

The Psychophysics of Speech Perception



NATO ASI Series

Advanced Science Institutes Series

A Series presenting the results of activities sponsored by the NATO Science Committee, which aims at the dissemination of advanced scientific and technological knowledge, with a view to strengthening links between scientific communities.

The Series is published by an international board of publishers in conjunction with the NATO Scientific Affairs Division

A	Life Sciences	Plenum Publishing Corporation
B	Physics	London and New York
C	Mathematical and Physical Sciences	D. Reidel Publishing Company Dordrecht, Boston, Lancaster and Tokyo
D	Behavioural and Social Sciences	Martinus Nijhoff Publishers Boston, Dordrecht and Lancaster
E	Applied Sciences	
F	Computer and Systems Sciences	Springer-Verlag Berlin, Heidelberg and New York
G	Ecological Sciences	London, Paris and Tokyo
H	Cell Biology	



The Psychophysics of Speech Perception

edited by:

M.E.H. Schouten

Institute of Phonetics
University of Utrecht
Utrecht
The Netherlands

1987 **Martinus Nijhoff Publishers**
Dordrecht / Boston / Lancaster
Published in cooperation with NATO Scientific Affairs Division

Proceedings of the NATO Advanced Research Workshop on "The Psychophysics of Speech Perception", Utrecht, The Netherlands, June 30-July 4, 1986

Library of Congress Cataloging in Publication Data

NATO Advanced Research Workshop on "The Psychophysics of Speech Perception" (1986 : Utrecht, Netherlands)
The psychophysics of speech perception.

(NATO ASO series. Series D, Behavioural and social sciences ; no. 39)

"Published in cooperation with NATO Scientific Affairs Division."

"Proceedings of the NATO Advanced Research Workshop on "The Psychophysics of Speech Perception", Utrecht, the Netherlands, June 30-July 4, 1986"--T.p. verso.

1. Speech perception--Congresses. 2. Psychoacoustics--Congresses. I. Schouten, Marten Egbertus Hendrik, 1946-. II. North Atlantic Treaty Organization. Scientific Affairs Division. III. Title. IV. Series. [DNLM: 1. Psychoacustics--congresses. 2. Speech Perception--physiology--congresses. WV 272 N2795p 1986]

BF463.S64N38 1987 153.6 87-12319

ISBN-13:978-94-010-8123-8

e-ISBN-13:978-94-009-3629-4

DOI:10.1007/978-94-009-3629-4

Distributors for the United States and Canada: Kluwer Academic Publishers, P.O. Box 358, Accord-Station, Hingham, MA 02018-0358, USA

Distributors for the UK and Ireland: Kluwer Academic Publishers, MTP Press Ltd, Falcon House, Queen Square, Lancaster LA1 1RN, UK

Distributors for all other countries: Kluwer Academic Publishers Group, Distribution Center, P.O. Box 322, 3300 AH Dordrecht, The Netherlands

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publishers,
Martinus Nijhoff Publishers, P.O. Box 163, 3300 AD Dordrecht, The Netherlands

Copyright © 1987 by Martinus Nijhoff Publishers, Dordrecht

Softcover reprint of the hardcover 1st edition 1987

TABLE OF CONTENTS

Preface by the editor	ix
Chapter 1 THE RELEVANCE OF PSYCHOPHYSICS FOR SPEECH PERCEPTION	1
B.H. Repp The role of psychophysics in understanding speech perception	3
N.A. Macmillan, L.D. Braida, and R.F. Goldberg Central and peripheral processes in the perception of speech and nonspeech sounds	28
D.W. Massaro Psychophysics versus specialized processes in speech perception: an alternative perspective	46
M.E.H. Schouten Speech perception and the role of long-term memory	66
B. Espinoza-Varas Levels of representation of phonemes and bandwidth of spectral-temporal integration	80
M.E.H. Schouten General discussion of session 1	91
Chapter 2 SEPARATION OF ACOUSTIC EVENTS	93
A.S. Bregman The meaning of duplex perception: sounds as transparent objects	95
C.J. Darwin and R.B. Gardner Perceptual separation of speech from concurrent sounds	112
M. Weintraub Sound separation and auditory perceptual organization	125
T. Houtgast On the significance of spectral synchrony for signal detection	135
Q. Summerfield and P. Assmann Auditory enhancement in speech perception	140
B.C.J. Moore General discussion of session 2	151

Chapter 3	153
DYNAMIC ASPECTS	
D.B. Pisoni and P.A. Luce	155
Trading relations, acoustic cue integration, and context effects in speech perception	
P. Howell and S. Rosen	173
Perceptual integration of rise time and silence in affricate/fricative and pluck/bow continua	
V.J. van Heuven	181
Reversal of the rise-time cue in the affricate/fricative contrast: an experiment on the silence of sound	
R.E. Pastore	188
Possible acoustic bases for the perception of voicing contrasts	
S. Rosen and P. Howell	199
Is there a natural sensitivity at 20ms in relative tone-onset-time continua? A reanalysis of Hirsh's (1959) data	
R.L. Diehl	210
Auditory constraints on speech perception	
D.G. Jamieson	220
Studies of possible psychoacoustic factors underlying speech perception	
L.C.W. Pols and M.E.H. Schouten	231
Perception of tone, band, and formant sweeps	
C. Sorin	241
Psychophysical representation of stop consonant and temporal masking in speech	
F. Lacerda	250
Effects of stimulus dynamics on frequency discrimination	
A. Bladon	258
Extending the search for a psychophysical basis for dynamic phonetic patterns	
D.B. Pisoni	264
General discussion of session 3	
Chapter 4	269
TIMBRE (PERIPHERAL CONSTRAINTS AND CENTRAL PROCESSES IN THE PERCEPTION OF COMPLEX SIGNALS).	
E. Terhardt	271
Psychophysics of audio signal processing and the role of pitch in speech.	

J.L. Schwartz and P. Escudier	284
Does the human auditory system include large scale spectral integration?	
H. Traunmüller	293
Some aspects of the sound of speech sounds	
B. Espinoza-Varas	306
Involvement of the critical band in identification, perceived distance, and discrimination of vowels	
D.M. Green and L.R. Bernstein	314
Profile analysis and speech perception	
L.C.W. Pols	328
General discussion of session 4	
 Chapter 5	331
PHYSIOLOGICAL CORRELATES OF SPEECH PERCEPTION	
B. Delgutte	333
Peripheral auditory processing of speech information: implications from a physiological study of intensity discrimination	
E.D. Young	354
Organization of the cochlear nucleus for information processing	
A.R. Palmer, I.M. Winter, R.B. Gardner and C.J. Darwin	371
Changes in the phonemic quality and neural representation of a vowel by alteration of the relative phase of harmonics near F1	
H. Traunmüller	377
Phase vowels	
J.W. Horst, E. Javel and G.R. Farley	385
Nonlinear responses in the auditory nerve to vowel-related complex stimuli	
G.F. Smoorenburg	393
Discussion of physiological correlates of speech perception	
G.F. Smoorenburg	400
General discussion of session 5	
 Chapter 6	403
PRIMARY SPEECH PERCEPTS	
J. Mehler and J. Segui	405
English and French speech processing: some psycholinguistic investigations	
R.E. Remez	419
Units of organization and analysis in the perception of speech	

P.W. Jusczyk Implications from infant speech studies on the unit of perception	433
A. Cohen General discussion of session 6	444
Chapter 7 PSYCHOPHYSICS AND SPEECH PERCEPTION IN THE HEARING- IMPAIRED	447
B.C.J. Moore and B.R. Glasberg Relationship between psychophysical abilities and speech perception for subjects with unilateral and bilateral cochlear hearing impairments	449
J.M. Festen Speech-reception threshold in a fluctuating background sound and its possible relation to temporal auditory resolution	461
A. Bosman and G.F. Smoorenburg Differences in listening strategies between normal and hearing-impaired listeners	467
Robert D. Celmer and Gordon R. Bienvenue Critical bands in the perception of speech signals by normal and sensorineural hearing loss listeners	473
Stuart Rosen Phase and the hearing-impaired	481

PREFACE

The following is a passage from our application for NATO-sponsorship:

"In the main, the participants in this workshop on the Psychophysics of Speech Perception come from two areas of research:

- one area is that of speech perception research, in which the perception of speech sounds is investigated;
- the other area is that of psychoacoustics, or auditory psychophysics, in which the perception of simple non-speech sounds, such as pure tones or noise bursts, is investigated, in order to determine the properties of the hearing mechanism.

Although there is widespread agreement among both speech researchers and auditory psychophysicists that there should be a great deal of co-operation between them, the two areas have, generally speaking, remained separate, each with its own research questions, paradigms, and above all, traditions. Psychoacousticians have, so far, continued to investigate the peripheral hearing organ by means of simple sounds, regarding the preoccupations of speech researchers as too many near-empty theories in need of a more solid factual base. Speech perception researchers, on the other hand, have continued to investigate the way human listeners classify vowels and consonants, claiming that psychoacoustics is not concerned with normal, everyday, human perception.

The two areas are inclined, then, to see each other as "dead ends", which could do with an infusion of ideas from the other side of the divide. Although such attitudes cannot be conducive to fruitful co-operation, developments in recent years have resulted in a more favourable climate:

- psychoacousticians are beginning to see increasingly that it is necessary to take the step from the question of what the limits of hearing are under laboratory conditions, to the question of how the hearing mechanism functions in everyday perception, especially speech perception;
- speech perception researchers are realising more and more that they have run up against a number of basic problems which cannot be resolved until more is known about how speech sounds are processed at the auditory periphery."

When I met Dr. Mario Di Lullo, head of NATO's Advanced-Research-Workshop programme, for the first and only time (sadly, he died just before the workshop began), his reaction to this passage was one of amazement that such a meeting had not already taken place a long time ago. I found it very hard to explain why this was so, but I could show him that many researchers felt the same way about it, although not many would formulate their opinions as starkly as I had done in the passage above.

The desire to bring about a meeting of researchers in psychoacoustics and in speech perception was not restricted to a few people in the Netherlands; in fact, it was shared by many people in many NATO countries. The degree to which it was shared was continually underestimated: we did not expect so many highly favourable reactions to our first circular; we did not expect so many people to submit papers; least of all did we expect all those people to turn up and actually present all those papers.

As a result, the workshop could have collapsed under its own weight; that it did not was due to that very same enthusiasm on the part of the participants, who were determined to turn it into a success, despite the long working days, without air conditioning, during a heat wave.

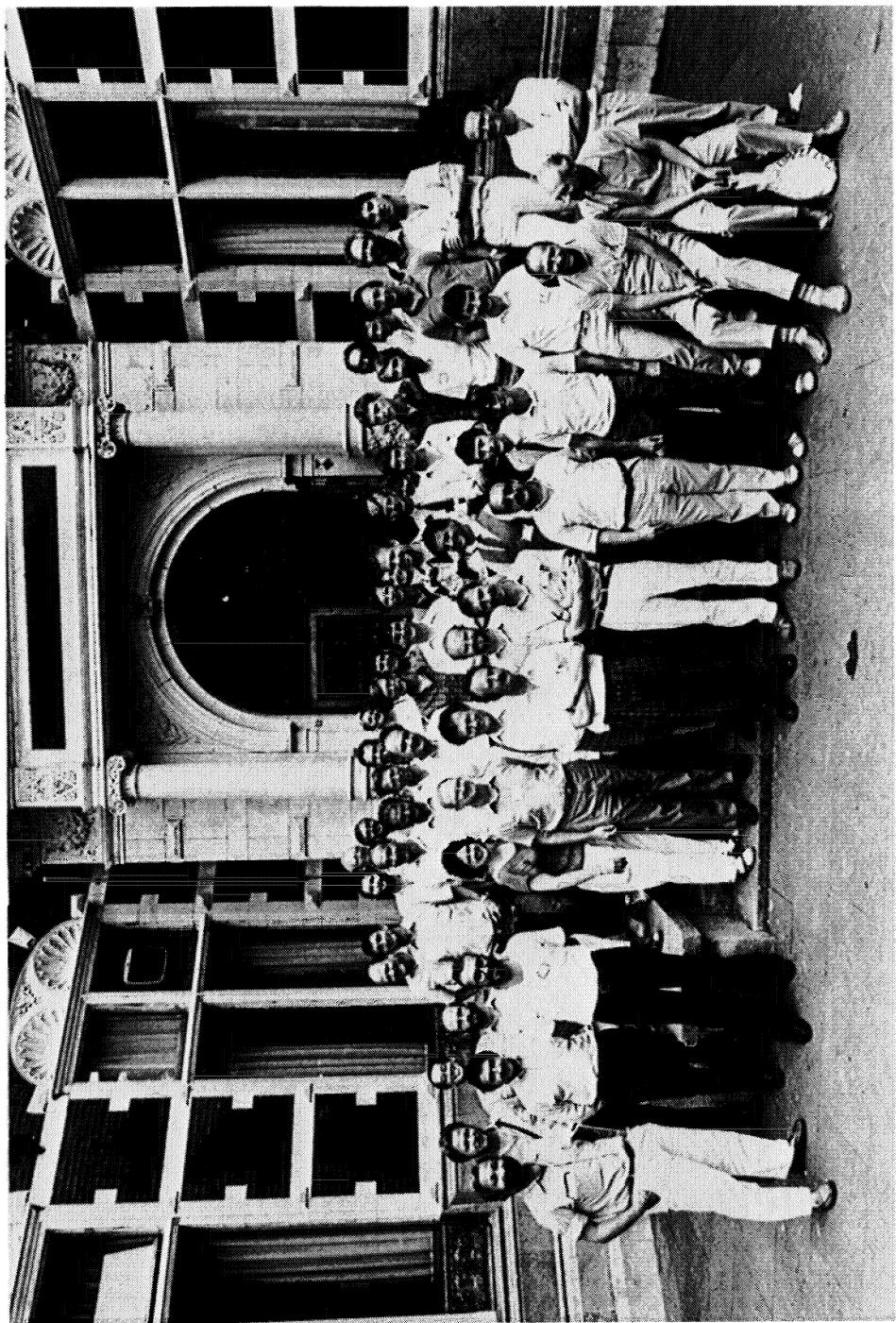
What have we learned from the workshop? A straight answer to this question will have to wait a few years, until the time has come for us to ask the participants whether a second workshop is desirable. Such a second workshop should have considerably fewer papers, most of which should report on research that was at least influenced by the present workshop. If such a meeting turns out to be impossible to organise, the conclusion will have to be that little of value has been learned from the present workshop, apart from a few relatively trivial conclusions that can even now be drawn:

1. it has given us a fairly comprehensive overview of the work that is currently going on near the boundary between speech perception and psychoacoustics;
2. it has taught us that a psychophysics of speech perception should be made up of a psychoacoustics (and a physiology) of complex, dynamically varying signals and a memory component.

This second conclusion involves a redefinition of the term "psychophysics", which, unlike "psychoacoustics", refers to more than just the periphery of hearing. This distinction is not a new one; researchers like N. Macmillan have been using it for years. It has not (yet) gained general currency: during the workshop the suggestion was made to change its name to "Psychophysics and speech perception", presumably on the assumption that "psychophysics" is more or less synonymous with "psychoacoustics". This suggestion has not been followed.

As has already been mentioned, the success of the workshop was due to the commitment of all participants. A lot of people have contributed, organisationally or otherwise, to the workshop and to this book. Those I'd especially like to mention include the chairmen, who turned the various sessions into relaxed but stimulating occasions, and Hellen König, whose hard work, friendliness, and charm were praised by all participants.

Bert Schouten



Participants of the NATO Advanced Research Workshop on
"The Psychophysics of Speech Perception"
Utrecht, The Netherlands, June 30 - July 4, 1986

Chapter 1

THE RELEVANCE OF PSYCHOPHYSICS FOR SPEECH PERCEPTION

THE ROLE OF PSYCHOPHYSICS IN UNDERSTANDING SPEECH PERCEPTION*

Bruno H. Repp

Haskins Laboratories,
270 Crown Street, New Haven, CT 06511, USA

0. INTRODUCTION

The purpose of this workshop is to discuss the psychophysics of speech perception. The program includes a variety of topics that presumably fall under this heading and that demonstrate that the psychophysics of speech perception is alive and well. Yet it is not really obvious what the psychophysics of speech perception is, what its goals and limitations are, and whether it is indeed a circumscribed area of investigation. It seems useful, therefore, to pose these basic questions explicitly and to include them in our discussions along with the many specific issues addressed by our research. The purpose of my paper is to stimulate such discussion by presenting a particular, possibly controversial, view of speech perception, psychophysics, and the relation between the two.

My presentation has five parts. First, I will attempt to define the psychophysics of speech perception and to discuss some of its assumptions and limitations. Then, turning to the second half of my title, I will consider briefly what it might mean to "understand" speech perception. Next, I will sketch a general view of phonetic perception and follow this with a discussion of what I believe to be the major research questions from that perspective. Finally, I will suggest a relatively novel application of psychophysics in the research enterprise I have envisioned.

1. WHAT IS THE PSYCHOPHYSICS OF SPEECH PERCEPTION?

I am starting with the assumption that there is indeed a psychophysics of speech perception--a particular area of scientific inquiry that the title of this workshop is intended to refer to. If so, what distinguishes the psychophysics of speech perception from the investigation of speech perception in general?

Psychophysics, as traditionally defined, is the science of describing the relationship between objective (physical) and subjective (psychological) dimensions. In a typical experiment, physical characteristics of a series of stimuli are measured or manipulated, and the subjects' judgments are obtained on an explicit or derived

*Preparation of this paper was supported by NICHD Grant HD-01994 to Haskins Laboratories. I am grateful to Al Bregman, Bob Crowder, Jim Flege, Ignatius Mattingly, Robert Remez, and Michael Studdert-Kennedy for helpful comments on an earlier draft.

numerical scale. The resulting stimulus-response relationship is often described in the form of a function, such as Weber's law or Stevens' exponential curves. However, there are many other ways of describing stimulus-response relationships, and it would be unwise to exclude any particular descriptions from the domain of psychophysics. Since virtually all speech perception research involves eliciting subjects' responses to stimuli that have been manipulated in some way, it seems to me that, at first blush, the psychophysics of speech perception is the only kind of research on speech perception that exists, especially if we exclude psycholinguistic topics such as word recognition and sentence comprehension, which concern the perception of meaning.

Is the title of this workshop then a tautology? Perhaps not. In fact, the term "psychophysics" is not commonly applied to all of the research on speech perception. Therefore, it has certain connotations that derive from the kinds of experiments it is explicitly associated with. That is, even though the boundaries of psychophysics are not clearly defined and may include a large variety of topics and methods, those researchers who consider themselves psychophysicists represent certain typical theoretical attitudes and preferences. Thus, psychophysics may be considered a particular approach to the study of speech perception that, without necessarily being programmatic, characterizes a fair amount of work in the field. I presume that, in choosing the title for this workshop, the organizers wished to highlight this approach, which I will now attempt to characterize.

1.1. FOCUS ON THE AUDITORY MODALITY

One attitude I associate with a psychophysical approach to speech perception is a preoccupation with psychoacoustics. Indeed, all presentations at this workshop are concerned with aspects of auditory speech perception. This is not to say that research on speech perception via the visual and tactile senses is not often psychophysical in character; in fact, much of it is, and several participants in this workshop have made important contributions to it. Nevertheless, this research has often been the province of specialists outside the mainstream of speech perception research. One consequence of this is that many speech perception researchers place special emphasis on auditory processes and thereby miss the more general insights to be gained from a multimodal approach.

Tactile speech perception, to be sure, is uncommon and requires special transduction devices; moreover, it is not clear whether tactile information feeds directly into the speech perception system the way auditory and visual information does (except for the Tadoma method, where articulation is felt directly). Visual speech perception, by contrast, is extremely common, especially in conjunction with listening. The extent to which auditory and visual information is integrated was strikingly demonstrated by McGurk and MacDonald (1976), who presented conflicting information in the two modalities and found that visual information may override the auditory information without the perceiver's awareness. In such instances, subjects believed they heard what in fact they saw. More often, the conscious percept represents a compromise between the inputs from the two modalities (Massaro & Cohen, 1983a; Summerfield, in press). It appears, therefore, that speech information from the two sensory modalities converges upon a common mental representation. As Summerfield (1979) and others have argued,

the information seems to be represented internally in a common metric that is amodal in nature.

If this kind of argument is accepted, it follows that not too much weight should be attached to descriptions of speech information that are tied to one modality. Rather, the basis for speech perception must be sought in information that is modality-independent and can be described in a common vocabulary. Such a vocabulary is provided by articulatory kinematics and/or by the dynamic parameters that underlie articulatory processes. To be sure, articulations taking place in the back of the vocal tract are transmitted exclusively by acoustic means, whereas movements of lips and jaw are prominent in the optic signal. This partial dissociation should not detract us from the fact, however, that in each case the information is about articulatory position and motion or, more abstractly, about the changing area function of the speaker's oral cavity.

Alternatively, it might be assumed that cues from different modalities are integrated in the process of categorical decision making, without recourse to a common metric (Massaro & Cohen, 1983a; Summerfield, in press). However, the question then arises: What motivates the integration in the first place? If the internal representations of stimuli are modality-specific, they can be related only through some form of association, either innate or acquired. In Massaro and Cohen's model, the associations reside in attribute lists which constitute phonetic category prototypes. Although this model seems to account well for audiovisual syllable perception, it seems less able to handle the intersensory integration of continuous dimensions such as speaking rate (Green & Miller, 1985) or prelinguistic infants' ability to recognize auditory-visual or visual-proprioceptive correspondences (Kuhl & Meltzoff, 1982; Meltzoff & Moore, 1985). A description of the stimulus information in articulatory terms eliminates the need to hypothesize independent mental representations of modality-specific correlates of articulation (see Yates, 1985), and it emphasizes the fact that the relation between visual and auditory manifestations of speech is nonarbitrary and possibly innately specified.

While it is generally taken for granted that we see the moving articulators when we look at them, not abstract optic patterns, there has been some reluctance in the field to accept the analogous proposition (Gibson, 1966; Neisser, 1976; Studdert-Kennedy, 1985) that, when we listen to speech, we hear the moving articulators and not the auditory patterns that constitute the proximal stimulus. Instead, researchers have been intensely preoccupied with acoustic variables such as formant transitions, delayed voicing onset, rise time, and so forth, as if the corresponding auditory percepts were the primary objects of speech perception. Whether they are is open to question, however (see, e.g., Liberman & Mattingly, 1985). Their prominent role in speech research may in large part be due to traditional techniques of acoustic analysis and synthesis, rather than to any compelling theoretical considerations. Many issues in the psychoacoustics of speech perception might never have been considered, had methods of articulatory analysis and synthesis preceded spectrographic and formant-based methods. As it is, we need to ponder whether these psychoacoustic issues are really pertinent to speech perception, or whether they merely have been forced upon us by the instruments we

have had available. In other words, if we had only articulatory synthesizers as well as devices that extract area functions from the acoustic (and/or optic) signal, what would be the theoretical status of phenomena such as backward masking, adaptation, contrast, spectral integration, etc., in speech perception research? How much would we lose if we talked only about articulation and not about acoustics at all?

1.2 FOCUS ON METHODOLOGY

A second tendency that may reasonably be associated with a psychophysical approach is a focus on methodology. Certainly in classical psychophysics the methods by which stimulus-response mappings are obtained have been of overriding concern. There are many examples of a similar concern in speech perception research. Many experiments have compared performance in different discrimination paradigms, such as AX, 4IAX, ABX, fixed versus roving standard, etc. (e.g., Pisoni & Lazarus, 1974; MacKain et al., 1981; Rosner, 1984; Macmillan et al., this volume), and even in the many studies using only a single method its choice has usually been a matter of concern. Other studies have compared different identification tasks, such as binary classification, numerical rating scales, absolute identification, and perceptual distance scaling (e.g., Ganong & Zatorre, 1980; Massaro & Cohen, 1983b; Vinegrad, 1972). In fact, it may be argued that most of categorical perception research, as well as much research on selective adaptation, contrast, auditory memory, etc., has been exercises in methodology. To be sure, the variations in methods have usually served to test some reasonable models or hypotheses, and I do not mean to imply that this research has been worthless. Nevertheless, the questions asked in such experiments often are somewhat removed from the original phenomena that stimulated the research; in other words, they have become methodological variations on a common theme, and sometimes variations themselves have become the themes for further variations.

Take categorical perception. The category boundary effect (Wood, 1976)--the well-known finding that discrimination performance is higher across a phonetic category boundary than within categories--is important because it tells us that the acoustic structure of speech is not very transparent to the typical listener, who habitually focuses only on linguistically significant information. Numerous studies have shown that the strength of the effect varies with methodological factors such as discrimination paradigm, interstimulus interval, training, instructions, language experience, types of stimuli, etc. (see review by Repp, 1984). The large majority of these studies has been concerned with subjects' ability to discriminate small acoustic differences among speech stimuli. This ability, not surprisingly, can be enhanced by training, reduction of stimulus uncertainty, short interstimulus intervals, etc. The studies that have shown this are prime examples of the psychophysics of speech perception, and they include many an elegant piece of experimentation. However, the important aspect of categorical perception that seems directly relevant to speech communication is not subjects' apparent inability to discriminate linguistically irrelevant differences along certain stimulus continua but rather their attention to linguistically distinctive information in the speech signal. To be sure, statements have been made in the literature (Studdert-Kennedy et al., 1970;

Liberman & Mattingly, 1985) to the effect that human listeners simply cannot perceive certain auditory properties of speech sounds, and this has, of course, been grist for the psychophysical mill. Apart from dismissing such extreme claims, however, little has been learned from all these studies about speech perception beyond the truism that perception within categories is not categorical. Rather, they have revealed some things about auditory discrimination and the methodological variables affecting it. Equivalent information could have been obtained by using nonspeech stimuli, and indeed one of the aims of psychophysical methodology (though this is rarely acknowledged) is to enable listeners to perceive speech as if it were a collection of arbitrary sounds. This leads me to another, related bias I associate with the psychophysics of speech perception.

1.3 FOCUS ON THE SOUNDS OF SPEECH

One possible definition of the psychophysics of speech perception is that it is the study of the perception of the sounds of speech. Unfortunately, the term "speech sounds" has often been used indiscriminately to denote both linguistically significant categories and acoustic components of the speech signal (and/or the auditory impressions associated with them). A clear distinction needs to be made between the auditory/acoustic and linguistic/articulatory domains, however (cf. Repp, 1981); the term "speech sounds" is appropriate for the former, whereas "phonetic categories" (or "phonemes") is appropriate for the latter. With this distinction in mind, my claim is that psychophysics is concerned, for the most part, with speech sound perception rather than with phoneme perception. It seems likely, however, that, except in very special circumstances, the sounds of speech as such do not play an important role in speech communication (see also Linell, 1982; Liberman & Mattingly, 1985; Traunmüller, this volume). Rather, I presume it is the more abstract, articulatory information that is used by listeners to decode the linguistic message. In fact, the only context in which the auditory qualities of speech segments may have a communicative function is in poetry, where an (unconscious) apprehension of the segmental sound pattern may enhance connotative and aesthetic qualities of the text (Fónagy, 1961; Hrushovski, 1980). Paradoxically, it seems that, so far, poetry has not attracted the attention of psychophysicists. (See, however, Marks, 1978.)

Why should one be interested in perceptual qualities that do not serve any important function in speech communication? There could be many valid reasons, such as questions about the auditory processing of complex sounds, the consequences of hearing impairment, skills of analytic perception, etc.--all topics worthy of scientific investigation. Nevertheless, these topics may be largely irrelevant to the perception of phonetic structure, and their study may therefore not contribute to our understanding of speech perception. To the hard-core psychophysicist, speech is primarily an acoustic signal of unusual complexity, which presents a challenge to the auditory system and to the experimenter's ingenuity. However, since this acoustic complexity is precisely what the speech perception system is equipped to handle, the speech signal actually has a very simple structure when viewed from the inside, as it were. For the speech perceiver, and for the speech researcher, perceptual complexity is defined by different criteria, such as the relative familiarity of a language, dialect, or

foreign accent, the rate of speech, or the fidelity of the acoustic signal. In other words, perceptual complexity is defined not absolutely but in terms of deviations from expectancies. In the case of synthetic or degraded speech, an acoustically simpler signal may pose a perceptual problem.

1.4 FOCUS ON THE NAIVE LISTENER

The bias that I have just portrayed--that psychophysics tends to be concerned with linguistically irrelevant aspects of speech--may seem to apply only to a small portion of speech research. After all, most speech perception experiments do require subjects to respond with phonemic categories (strictly speaking, with alphabetic symbols) to the speech sounds they hear, and not with numerical ratings or other kinds of nonphonetic responses. However, it is often assumed, if only implicitly, that the phonemic or orthographic symbols employed by listeners are simply convenient labels for auditory experiences. Hand in hand with that assumption goes the much-discussed hypothesis that phonetic categories, and particularly the boundaries between them, reflect constraints imposed by the mammalian auditory system (see, e.g., Kuhl, 1981; Liberman & Mattingly, 1985). This hypothesis dovetails with another bias of psychophysical research.

Classical psychophysics is rarely concerned with subjects' experience prior to an experimental session, except for task-specific training received under controlled conditions. Essentially, psychophysics is about basic processes of perceptual translation, most often from a continuous physical dimension to a continuous psychological dimension. If categories are to be employed as responses in a psychophysical task, they are usually defined within the limited context of the experimental situation, often exemplified by the extremes of a stimulus dimension. The boundaries between such categories are either arbitrary--e.g., they may just bisect a stimulus continuum and hence depend on its range--or, if they are not (as is more often the case with speech) they are assumed to coincide with a psychoacoustic discontinuity that gave rise to the categories in the first place. Although subjects obviously have much experience with the categories of speech outside the laboratory, this experience is often considered irrelevant because the psychoacoustic basis for the category division is assumed to be present in the stimuli. (If the stimuli are synthetic and unfamiliar-sounding, so much the better.) At best, language experience may have taught subjects to attend to one particular discontinuity and to ignore another, hence certain cross-language differences in boundary location.

These assumptions are perfectly appropriate within the framework of psychophysics. Indeed, in the quest for an elegant description of the perceptual translation from the objective to the subjective realm, any intrusion of pre-experimental knowledge is undesirable. Imagine an experiment involving the perceived similarity of various round shapes, in which subjects judge two shapes as more similar than the others because both happen to look like the same familiar object (e.g., an apple). This would be an undesirable artefact (Titchener, 1909, called it the "object error") that might distort the true psychophysical function underlying the similarity judgments. This function is assumed to be universal and independent of prior experience.

There is considerable evidence, however, that many, perhaps all, phonetic distinctions rest on linguistic, not psychoacoustic criteria (see Rosen & Howell, *in press*; Repp & Liberman, *in press*). These criteria are acquired--or, if innate, are modified--through experience with spoken language. Rather than referring to particular auditory experiences, phonetic category labels--once certain orthographic and linguistic conventions are stripped off--denote specific articulatory maneuvers whose auditory correlates, though systematic, are largely irrelevant. This is most strikingly demonstrated by the finding that phonetic structure can be perceived in auditorily anomalous stimuli composed of time-varying sinusoids that imitate formant movements and thus retain information about the changing shape of the vocal tract (Remez et al., 1981; Remez, *this volume*). The articulatory patterns characteristic of a language presumably have evolved according to articulatory and linguistic constraints (Lindblom, 1983; Ohala, 1983), and it seems unlikely that auditory limitations have played a significant role, except in the very general sense that phonetic contrasts that are difficult to discriminate tend to be avoided or, if they occur, may lead to language change (Ohala, 1981; Bladon, *this volume*). I will argue below that listeners refer to their knowledge of language-specific articulatory norms when listening to speech. This reference is external to the experimental situation and inside the listener. Rather than emerging from acoustic properties of the stimulus or the stimulus ensemble, the phonetic structure imposed by the talker and recovered by the listener represents a learned conventional pattern constrained by universal articulatory possibilities.

Since it is the linguistic structure that is important in speech communication, and not the auditory properties of speech components, it is natural that human listeners focus their attention on the former and not on the latter. This attention to a discrete representation of speech influences subjects' judgments in a variety of psychophysical tasks designed to assess the psychological transformation of acoustic stimulus dimensions. For example, it is probably responsible for the category boundary effect in categorical perception experiments, as hypothesized long ago by proponents of the so-called dual-process model (e.g., Fujisaki and Kawashima, 1969, 1970; Pisoni, 1973; Samuel, 1977). However, some researchers committed to psychophysical approaches (e.g., Macmillan et al., 1977) have taken these perceptual nonlinearities to be inherent in the auditory stimulus representation. Although auditory nonlinearities do seem to occur along certain acoustic dimensions of speech, they may be unrelated to the discontinuities imposed by the mental organization of the listener (see, e.g., Watson et al., 1985; Howell & Rosen, 1983; Rosen & Howell, *in press*; Schouten, *this volume*). The same may be said about so-called phonetic trading relations and context effects (see review by Repp, 1982) which, for the most part, reflect not psychoacoustic interactions among signal components but the listener's imposition of multidimensional criteria in the process of phonetic categorization (Derr & Massaro, 1980; Massaro, *this volume*; Repp, 1983; however, see also Diehl, *this volume*).

By these arguments, speech is a particularly unwieldy object for psychophysical and psychoacoustic experimentation. If questions of auditory perception are to be addressed, why not use simpler stimuli?

If questions of speech communication are to be addressed, why use a psychophysical approach? As Massaro (this volume) aptly points out, a large part of modern speech perception research consists of either (a) applying reductionistic models to laboratory phenomena in a search for the auditory mechanisms that accomplish phonetic categorization, or (b) appealing to "special" mechanisms that do the job. Both enterprises have been sterile--the first in that it has not revealed any relevant mechanisms, and the second in that it has postponed or even relinquished the search for them. One problem with both approaches is that they represent models of speech perception according to which linguistically distinctive information somehow must emerge from the stimulus alone, without recourse to long-term mental representations of linguistic knowledge. One notable exception has been the work of Massaro and his collaborators who have consistently pursued the idea that speech perception proceeds by reference to internal category "prototypes" (see Massaro & Cohen, 1980a; Massaro, this volume). Their model, and similar ideas in the literature, lead the way toward a relational (or systemic) theory of speech perception, to be sketched further below.

2. UNDERSTANDING SPEECH PERCEPTION

The goal of speech perception researchers is to understand (or explain) speech perception--that much is obvious. However, what does this really mean? What is speech perception, and what does understanding (explaining) it entail? Probing these questions too deeply leads to profound epistemological issues. I offer only a few comments for discussion.

2.1 TWO DEFINITIONS OF PERCEPTION

The term "perception" is being used in different ways by different researchers, as has been pointed out by Chistovich (1971) and Shepard (1984), among others. An example of one usage is provided by Massaro's recent writings on categorical perception (Hary & Massaro, 1982; Massaro & Cohen, 1983b; Massaro, in press, and this volume). He argues that "categorical results do not imply categorical perception": The perception of speech continua is revealed to be continuous if only the right methods are employed. According to Hary & Massaro (1982), "a central issue in auditory information processing is whether certain auditory continua are perceived categorically rather than continuously" (p. 409). That is, it must be one or the other: Perception is entirely a function of the input. Perception is thus equated with sensory transduction--an immutable process that is insensitive to attention and experience. Of course, this is exactly what psychophysics is concerned with. The goal of speech perception research, in this view, is to find out what speech perception really is like, once all constraints imposed by attentional and experiential factors have been removed. The classification by reference to prototypes, which plays such a prominent part in Massaro's model, apparently is a post-perceptual process in his definition.

This view needs to be contrasted with a definition of perception that includes categorization and attentional filtering. According to this (my preferred) view, perception is what occurs when the transduced stimulus meets the mental structures (the "model of the world") laid down by past experience and possibly by genetic

transmission (Hayek, 1952; Shepard, 1980, 1984; Yates, 1985). The result of perception is the outcome of that encounter, not the input to it. According to Fodor (1983, p. 40), "what perception must do is so to represent the world as to make it accessible to thought" through processes of transduction and inference. Categorical perception, and the apparent invariance of the categorical percept, represent the outcome of the inferential process. To find behavioral evidence of the (largely) continuous, transduced information that feeds into this process, a listener's perceptual strategy must be altered through instructions and training, or some measure of decision uncertainty (e.g., reaction time) must be obtained. Since there are a variety of mental structures a stimulus may relate to, there are often alternative ways of perceiving the same input, depending on the perceiver's experience (i.e., form of the mental representations) and attention (i.e., selection from among them). Thus, in this view, categorical results do imply categorical perception, and noncategorical results imply noncategorical perception.

Speech perception thus can mean different things depending on the situation and the subject's strategies. In addition, it has a double meaning from another perspective, depending on whether "speech" is taken to refer to the stimulus or the percept. Psychophysical research can be snugly accommodated under the stimulus-based definition that speech perception is whatever occurs when speech signals are presented to a listener. I favor a percept-based definition--that speech perception occurs when a stimulus is perceived as speech, i.e., when the listener interprets the stimulus in relation to the linguistic system. By that definition, many psychophysical experiments deal not with speech perception but with the perception of speechlike auditory stimuli. This distinction is not intended as a value judgment (indeed, psychophysical research generally surpasses speech perception research in rigor and methodological sophistication), but as a separation of largely independent domains of inquiry.

2.2 TWO DEFINITIONS OF UNDERSTANDING

What does it mean to understand (or explain) speech perception? According to one view, it involves building or programming a machine to recognize speech. For example, Chistovich (1980) presented this approach as the one taken by the Leningrad group. This pragmatic goal of "teachability" deserves our respect (for a critique, see Studdert-Kennedy, 1985). Even though the operations of the machine may not resemble those of the human brain, a speech recognition algorithm approximating human capabilities would represent a useful model of speech perception and thus increase our understanding of the process. Unfortunately, it seems that psychophysics has little to contribute to this enterprise. Psychoacoustic and physiological research has uncovered transformations in the auditory system that could be simulated by a speech processing system. However, incorporating auditory transforms into the machine representation of speech apparently does not improve speech recognition scores (Blomberg et al., 1986). This is perhaps not surprising. Machine representations need to capture the relationships between stimulus properties and precompiled knowledge structures (Shepard, 1980), and relational properties are likely to be largely invariant under transformations. Moreover, transformations of the input cannot result in an information gain, let alone in the magical

emergence of properties that cannot also be computed by a central algorithm, so the most detailed coding of the speech signal is likely to be the most useful one for machines. Unless the goal is to build an analog of a complex biological system (and we are far from that stage), insights derived from psychophysical and psychophysiological research are likely to be of little use to computers. The essential problem to be solved in speech recognition research, I presume, is not that of stimulus coding but that of phonetic knowledge representation and utilization.

The alternative approach to scientific explanation is a purely theoretical one. Scientists and other human beings, of course, can perceive speech and need not (cannot) be taught explicitly, so the teachability criterion does not apply. This approach to explanation, therefore, is fundamentally different from that provided by the automatic speech recognition research. Theory construction, in psychology at least, is a cognitive act subject to individual preferences, sociological factors, and philosophical considerations (see Toulmin, 1972). One person's explanation may be another's tautology.

A variety of scientific philosophies are evident in the speech perception field, and their coexistence for a number of years suggests that they represent, in large part, individual preferences and not theories subject to empirical disconfirmation. What is worse, they do not agree on what really needs to be explained about speech perception. Rather than discussing the current theories or endorsing any of them, I am going to present a personal view below, at the danger of adding to the general confusion. My own ideas are neither fully worked out nor entirely original. (See, for example, Bregman, 1977; Elman & McClelland, 1984, 1986; Hayek, 1952; Liberman & Mattingly, 1985; Massaro & Oden, 1980a; Shepard, 1980, 1984; Yates, 1985.) Whatever their merit, however, they may serve as a useful basis for discussion at this workshop. After presenting my view, I will discuss what seem to be the major research questions from this perspective and what role psychophysics might play in this enterprise.

3. SPEECH PERCEPTION AS A RELATIONAL PROCESS

Phonetic perception--i.e., the perception of the phonological structure of speech without regard to its semantic content--has often been considered a purely input-driven process, to be contrasted with the largely knowledge-driven processes of language understanding (e.g., Marslen-Wilson & Welsh, 1978; Studdert-Kennedy, 1982). That is, it is often assumed that phonological structure is in the speech signal (e.g., Gibson, 1966; Fowler, 1984; Stevens & Blumstein, 1981) or emerges from it via specialized neural processes (Liberman & Mattingly, 1985). The present proposal contrasts with these views in that it assumes that speech perception requires two complementary ingredients: the input signal and the perceiver's internal representation of the speech domain. In other words, I am assuming that phonological structure emerges, especially in its language-specific details, from the relation between a stimulus and a "phonetic lexicon" in the perceiver's head which (in mature individuals) provides an exhaustive knowledge base representing all the characteristics associated with the structural units of a language.

In this view, it is not the stimulus as such (or its auditory transform) that is perceived, but rather its relationship to the phonetic knowledge base; perception thus is a relational process, a two-valued function. Its output is also two-valued: The relation of the input to the pre-existing internal structures yields (potential) awareness of the structure that provides the best fit, plus some measure of goodness of fit which may be experienced as degree of confidence or uncertainty.

How is the phonetic knowledge represented in the brain? One possible conceptualization is in terms of "prototypes" (schemata, norms, ideals, logogens, basic categories) abstracted from language experience (cf. Massaro & Oden, 1980; Flege, 1986; Yates, 1985). The mechanisms enabling this abstraction during language acquisition are unknown and may either reside in a specialized "module" (Fodor, 1983; Liberman & Mattingly, 1985) or represent general neural design principles (e.g., Grossberg, 1986). Language-specific phonetic categories are assumed to "crystallize" around central tendencies extracted from the variable input under the guidance of linguistic distinctiveness criteria. How this occurs is one of the great unsolved questions in speech research.

Just like the stimulus itself, the contents of the listener's knowledge base can be described in acoustic (optic), auditory (visual), or articulatory terms; that is, the lexicon is assumed to contain information about typical articulatory motions and their acoustic and optic concomitants, as well as possibly about their underlying dynamic parameters. The articulatory information is primary in so far as it also serves to control speech production and silent (imagined) speech, because it relates more directly to linguistic and orthographic symbols, and because it unites the different sensory modalities (as pointed out earlier). Whatever metaphor is used to describe the knowledge base--and we cannot expect to capture in words the state of a complex neural network--the important consequence of having it is that a perceiver is able, at each moment in time, to evaluate the information in the speech signal as to whether it fits the language norms. Deviations from these expectations may be perceived as unnaturalness, foreign accent, or individual speaker characteristics; or they may pass unnoticed.

Speech that is pronounced clearly, free of noise, and typical of the language is perceived "directly": The appropriate prototypes "resonate" to the input (Shepard, 1984). Ambiguous or degraded speech is represented in terms of its relative similarities to the most relevant prototypes. Whenever a decision is required, one prototype is selected that provides the best fit to the input (cf. Massaro & Oden, 1980). Explicit linguistic category decisions, however, are basically a response phenomenon governed by (laboratory) task requirements. Whether or not overt categorical decisions are made, the structural linguistic information is always present, being implicit in the prototypes and their relations to each other (cf. Lindblom et al., 1983). The size of the "perceptual units," and with it the size of the prototypes, is variable, being a joint function of cognitive accessibility and real-time requirements (cf. Warren's, 1981, LAME model). Thus, even though explicit recognition of individual phonemes is likely to be a function of literacy and linguistic awareness (cf. Mattingly, 1972; Morais et al., 1979), phonemic structure is nevertheless implicit in the prototype inventory: For example, /b/ is perceived when all prototypes

transcribable as /b.../ are "active," i.e., resemble the input (cf. Elman & McClelland, 1984, 1986).

Properties of the speech signal become linguistic information only by virtue of their relation to the listener's knowledge base. One could imagine that the stimulus is represented in terms of a "similarity vector" (Chistovich, 1985) containing relative deviations from prototypes in some perceptual metric. This form of coding may be viewed as an effective way of information reduction, though it is by no means clear that the brain needs such a reduction the way we need it when thinking about the system's operation. That is, a similarity vector is better thought of as a set of potentials or relationships, not of physically instantiated quantities.

In my view, the "special" nature of speech, which has received so much emphasis in the past (e.g., Liberman, 1982), resides primarily in the fact that speech is a unique system of articulatory and acoustic events. In contrast to adherents of the modularity hypothesis (Fodor, 1983; Liberman & Mattingly, 1985) I suspect that the mechanisms of speech perception are general--i.e., that they can be conceptualized in terms of domain-independent models, such as adaptive systems theory (Grossberg, 1986), interactive activation theory (Elman & McClelland, 1984), or information integration theory (Massaro & Oden, 1980). In other words, I believe that the specialness of speech lies in those properties that define it as a unique phenomenon (i.e., its production mechanism, its peculiar acoustic properties, its linguistic structure and function) but not in the way the input makes contact with mental representations in the course of perception. That is, as long as we can only rely on models of the perceptual mechanism, it is likely that significant similarities will obtain across different domains, even though the physiological substrates may be quite different. This is a consequence of the relatively limited options we have for constructing models of perception and decision making.

To go one step further: If speech is special but speech perception is not, it follows that there is a lot to be learned about speech, but relatively little about speech perception. This conclusion, for what it is worth, suggests a "vertical" research strategy (giving a twist to Fodor's, 1983, arguments): The way to learn more about the speech system is to investigate its many special characteristics. This is a multidisciplinary venture, a task for the specialist called "speech researcher." By contrast, study of speech perception as such is open to a "horizontal" approach by psychologists interested in perception in general. However, there is comparatively little to be learned about that process. While there are lots of interesting facts to be uncovered about speech, the "mechanisms" of perception are a figment of the scientist's imagination (as is the mechanistic analogy itself). It is quite likely that, once we know enough about speech and have characterized the perceiver's knowledge in a suitably economic form, we also will have explained speech perception in its essential aspects.

4. A PROGRAM FOR SPEECH PERCEPTION RESEARCH

From the perspective I have adopted, there are four major questions for research on speech perception: What is the phonetic knowledge? How is it used? How is it acquired? How can it be modified?

4.1 DESCRIPTION OF THE KNOWLEDGE BASE

Before we can ask any questions about speech perception, we need to know what speech is, so we can account for the perceiver's expectations. This seemingly obvious requirement is often neglected by psychologists who plunge into speech perception experiments without considering the relevance of acoustic, articulatory, and linguistic phonetics. Even so useful a tool as Massaro's "fuzzy logical model" of information integration (Massaro & Oden, 1980a, 1980b) yields parameters characterizing phonetic prototypes whose relation to the normative properties of English utterances often remains unclear. It is often assumed that these properties will emerge from studies involving the classification of acoustically impoverished stimuli (see also Samuel, 1982). This is unlikely, however, because perceivers have detailed expectations about the full complement of acoustic properties, including those held constant in a given experiment, and they will often shift their criteria for stimulus classification along some critical dimension to compensate for the constancy or absence of others. While demonstration of this fact may be a worthwhile goal of some experiments, a more important point is that the perceivers' expectations can be assessed directly and independently (at least to a first approximation) by collecting facts about the acoustic and articulatory norms of their language, which constitute their knowledge base. Ever since Chomsky's (1965, 1968) seminal publications, the study of syntax, semantics, and phonology has been considered part of cognitive science, leading to a description of the language user's knowledge. I would like to add (normative) phonetics: The study of articulatory norms, too, yields a description of the average listener-speaker's "competence" (cf. Tatham, 1980).

I am thus proposing that the study of acoustic and articulatory phonetics be part and parcel of speech perception research. Incidentally, psychologists, with their thorough understanding of measurement and sampling problems, are especially well equipped to conduct phonetic and articulatory research, which too often has taken a case study approach in the past. Representative measurements are also important for automatic speech recognition research (Klatt, 1986). They would not make experimental determinations of prototypical perceptual parameters superfluous but rather provide a basis for their interpretation: The normative characteristics of a language are what a perceiver ought to have internalized. If deviations from the norm and/or individual differences emerge from such a comparison, the search for their causes should be an interesting and important undertaking.

In what form phonetic knowledge is represented in the brain is a question that cannot be answered conclusively by psychologists, who may choose from a number of alternative conceptualizations. As Shepard (1980, p. 181) has aptly stated, "there are many possible levels of description, and although they may appear very different in character, the various levels all pertain to the same underlying system. In this respect, the internal representation is no different from the external object." Choosing one particular level of description is basically a matter of preference and, perhaps, parsimony.

4.2 PERCEPTUAL WEIGHTS AND DISTANCES

One empirical question that psychologists may usefully address, however, is how phonetic knowledge is applied. Since a clear, unambiguous stimulus poses no challenge to the perceptual system and therefore cannot reveal its workings (cf. Shepard, 1984), the principal question is how phonetic ambiguities created by realistic signal degradation or by deliberate signal manipulation are resolved (explicitly) by the perceiver in the absence of lexical, syntactic, or other higher-order constraints. In such a situation, the perceiver must make a decision based on the perceptual distances of the input from the possible phonetic alternatives (prototypes) stored in his or her permanent knowledge base. The decision rule may be assumed to be straightforward: Select the prototype that matches the input most closely. However, what determines the degree of the match? What makes an ambiguous utterance more similar to one prototype than another? In other words, what is the phonetic distance metric, what are the dimensions of the perceptual space in which it operates, and what are the perceptual weights of these dimensions?

There are opportunities for the useful application of psychophysical methods here, since the distance metric may be, in part, a function of auditory parameters (see, e.g., Bladon & Lindblom, 1981). However, the relative importance of different acoustic dimensions for a given phonetic contrast cannot be predicted from psychophysical data alone, since it depends heavily on the nature and magnitude of the differences among the relevant prototypes, in combination with their auditory salience. Traditional psychophysics is concerned with perceptual similarities and differences between stimuli, whereas the present application requires a multidimensional psychophysics dealing with the similarity of stimuli to mental representations. The many confusion studies in the literature (beginning with Miller and Nicely, 1955) would seem to be about this issue, but the data have always been analyzed in terms of stimulus-stimulus, not stimulus-prototype similarities (which they indeed represent), and it is possible that important information has been missed. Research such as Massaro's modelling of information integration in phoneme identification (e.g., Derr & Massaro, 1980; Massaro & Oden, 1980a, 1980b) is an exemplary effort from the present viewpoint, despite certain limitations. Massaro has found again and again that stimulus attributes are evaluated in an independent and multiplicative (or log-additive) fashion in phonetic classification, and this has obvious implications for the nature of a phonetic distance metric. Many experiments on the perceptual integration and relative power of acoustic cues (e.g., Abramson & Lisker, 1985; Bailey & Summerfield, 1980; Lisker et al., 1977; Repp, 1982) also contribute relevant information. Experiments that avoid the fractionation of acoustic signals into "cues" and search for a phonetic distance metric based on more global spectral properties (Klatt, 1982, 1986) are promising but still at a very early stage.

Even though perceptual distances may reflect certain facts about auditory processing, these influences on phonetic perception are probably limited. The principal reason is that the mental structures that determine speech categorization have been built up from past experience with speech that underwent essentially the same auditory transformations as the current input is undergoing. That is, all transformations occurring during stimulus transduction are necessarily

represented in the central knowledge base. Therefore, it makes relatively little difference whether we think of the input as sequences of raw spectra and of the mental categories as prototypical spectral sequences (e.g., Klatt, 1979), or whether we consider both in terms of some auditory transform or collection of discrete cues. It is the relation between the two that matters, and that relation is likely to remain topologically invariant under transformations. Only nonlinear transformations will have some influence on phonetic distances (Klatt, 1986).

4.3 PERCEPTUAL DEVELOPMENT

In addition to asking how phonetic knowledge is utilized, we must ask how and when it is acquired. Much developmental and comparative research in the past has focused on auditory discrimination abilities, and the approach has been quite psychophysical in character. The "categorical" effects that have been observed in infants and animals may not reflect phonetic perception but certain psychoacoustic discontinuities on speech continua (Jusczyk, 1985, 1986), although this suggestion becomes doubtful in view of findings (Sachs & Grant, 1976; Soli, 1983; Watson et al., 1985) that the category boundary effect can be trained away in adults. Alternatively, category boundary effects in infants may reflect an innate predisposition for perceiving a universal articulatory inventory (Werker et al., 1981). The interpretation of these data is uncertain at present. Speech perception research in older children (e.g., Elliot et al., 1981; Tallal & Stark, 1981) also has often focused on their auditory abilities, not specifically on their criteria for phonetic identification and on the nature of their phonetic knowledge. Only more recently, following the lead of researchers such as Kuhl (1979) and Werker et. al. (1981) has phonetic categorization in infancy been studied more carefully. A finding of special significance is the discovery (Werker & Tees, 1984) that infants' ability to perceive phonetic contrasts foreign to their parents' language declines precipitously before 1 year of age. This stage seems to mark the beginnings of a language-specific phonetic lexicon. It is an important research endeavor to trace the accumulation and refinement of phonetic knowledge through different stages of development, and much work remains to be done (see Jusczyk, this volume).

4.4 PERCEPTUAL LEARNING

Another question of great theoretical and practical importance is how the phonetic knowledge, once it is established in the mature adult, can be augmented and modified. This concerns the process of second language learning and also, to some extent, the skills acquired by professional phoneticians (and even by subjects in a laboratory task, although their skills may be rather temporary). Furthermore, there is the very interesting question of bilingualism--the separation and interaction of two different, fully established phonetic knowledge bases. Until recently, little rigorous research had been carried out in this predominantly education-oriented area. Research is burgeoning, however, and is yielding interesting results (see Flege, in press).

Another, related question is to what extent reduced or distorted auditory input over longer time periods affects the internal representation of phonetic knowledge. For example, it has been

reported recently that otitis media in childhood (Welsh et al., 1983) or monaural hearing deprivation in adulthood (Silman et al., 1984) may result in reduced speech perception capabilities. Certainly, the congenitally hearing-impaired must have a very different representation of their limited phonetic experiences, and hearing impairments acquired later in life may distort the knowledge base as well. It has often been observed that the speech perception of the hearing-impaired is not completely predictable from assessments of auditory capacity (e.g., Tyler et al., 1982). One reason for this may be that there are distortions, not only in the auditory processing of speech (to which they are commonly attributed), but in the mental representations that hearing-impaired listeners refer to in phonetic classification. Such distortions are especially likely to result when hearing deteriorates progressively at a rate that exceeds the rate at which mental prototypes can be modified: A listener then expects to hear things that the auditory system cannot deliver. On the other hand, if the prototypes are degraded from many years of impoverished auditory experience, then there is little hope of improving speech perception by "improving" the acoustic signal, at least not without extensive training to rebuild the prototypes (cf. Sidwell & Summerfield, 1985).

5. MAKING PSYCHOPHYSICS MORE RELEVANT TO SPEECH RESEARCH

One characteristic of the psychophysical approach is that it is domain-independent. The psychophysical methods applied in the study of speech perception are essentially the same as those applied in research on auditory, visual, or tactile perception of nonspeech stimuli. Indeed, the generality across different stimulus domains and modalities of Weber's law or the law of temporal summation has been an important discovery. Such laws are in accord with behaviorist and information-processing orientations in psychology, which assume that perception and cognition are governed by general-purpose, domain-independent processes. The description of such processes is an important part of psychological research.

By focusing on domain-independent laws of perception, however, psychophysics essentially ignores those features that are specific to speech and whose investigation is critical to an understanding of speech perception as distinct from perception in general. Of course, there are many aspects that speech shares with nonspeech sounds and even with stimuli in other modalities. Research on the perception of those, however, leads only to an understanding of sound perception, temporal change perception, timbre perception, even categorization--in short, of all the things that speech perception has in common with nonspeech perception. What is missing is the main ingredient: the content. To understand speech perception fully, research needs to focus on the unique properties of speech, which include the facts that it is articulated (and hence peculiarly structured), capable of being imitated by a perceiver, and perceived as segmentally structured for purposes of linguistic communication. I see at least one way in which the sophisticated methods of psychophysics could be adapted to these special features and thus be made more relevant to speech research.

Psychoacoustic approaches to speech perception deal with both stimulus and response at some remove from the mechanism that is

directly responsible for most (if not all) special properties of speech: the vocal tract. A more speech-relevant psychophysics might examine the articulatory source of the acoustic signal in relation to what is probably the most direct evidence that perception has occurred--the perceiver's vocal reproduction of what has been heard (or seen). I am thus proposing an articulatory psychophysics based on the realization that speech is constituted of motor events (cf. Liberman & Mattingly, 1985). Its goal would be to describe the lawful relationships between a talker's articulations and a listener's perception or imitation of them.

A first step in this enterprise would be to look at the speech signal not in terms of its acoustic properties, but in terms of the articulatory information that it conveys. This is done most easily by generating the stimuli using an articulatory synthesizer or an actual human talker, perhaps in conjunction with analytic methods for extracting the vocal tract area function from the acoustic signal (e.g., Atal et al., 1978; Ladefoged et al., 1978; Schroeder & Strube, 1979). Articulatory synthesis studies in the literature (e.g., Lindblom and Sundberg, 1971; Rubin et al., 1981; Abramson et al., 1981; Kasuya et al., 1982) illustrate this approach. A second step would be to examine subjects' articulatory (rather than just written) response to speech stimuli. Studies of vocal imitation (e.g., Chistovich et al., 1966; Kent, 1973; Repp & Williams, 1985) have commonly analyzed stimulus-response relationships in terms of acoustic parameters and thus fall somewhat short of the stated goal. In the wide field of speech production research, there are few studies that have required subjects to listen to speech stimuli and reproduce them; almost always the task has been to read words or nonsense materials, and measurements have focused on normative productions characteristic of a language, not on talkers' imitative or articulatory skills. The final step towards a true articulatory psychophysics would be to measure subjects' articulatory response to articulatorily defined stimuli, generated either by an articulatory synthesizer or by a human model whose articulators are likewise monitored. An important (though necessarily crude) example of this still rare approach is the work of Meltzoff and Moore (see 1985) on facial imitation in infancy. More detailed studies of adult subjects should benefit from the development of more economic descriptions of articulation and its underlying control parameters (Brownman & Goldstein, 1985; Kelso et al., 1985).

Such studies would assess how articulatory dimensions such as jaw height, lip rounding, mouth opening, or velar elevation--or perhaps more global articulatory parameters such as the vocal tract area function--are apprehended by a listener/speaker, and how they are translated and rescaled to fit his or her own articulatory dimensions. Rather than relating physical stimulus parameters to some subjective auditory scale that is irrelevant to speech communication, the psychophysical function would relate equivalent articulatory measures in the model speaker and the imitator. Such functions would relate more directly to questions of speech acquisition and phonetic language learning than any measure of auditory perception. Even though articulatory psychophysics is likely to encounter various influences of linguistic categories on the subject's articulatory response, reflecting aspects of motor control that have become established through habit and practice, at least it would bypass the stage of overt categorical decisions (cf. Chistovich et al., 1966) which characterizes so many laboratory tasks. It may be possible to overcome these articulatory

habits through training, and such training may not only yield better estimates of articulatory information transfer but also potential practical benefits for second-language learners and speech pathologists (more so than training in auditory discrimination). An ancillary, hitherto little-investigated topic is that of articulatory awareness--a talker's ability to consciously observe and manipulate his or her articulators.

6. SUMMARY

Before summing up, one qualification is in order concerning the role of psychophysics in understanding speech perception. I have argued that this role is limited, and undoubtedly many will disagree with this opinion. In addition, however, I have followed the custom of the mainstream speech perception literature (and my own proclivities) by considering speech perception to be synonymous with the perception of phonetic structure. There are many other aspects of speech, however, such as intonation, stress, speaking rate, effort, rhythm, emotion, voice quality, speaker characteristics, room reverberation, and separation from other environmental sounds. All these aspects are worthy of detailed investigation, and although speech-specific knowledge also plays a role in their perception (e.g., Tuller & Fowler, 1980; Darwin, 1984; Ainsworth & Lindsay, 1985), auditory psychophysics probably has a more important contribution to make to research on these topics. The perception of subtle gradations becomes especially important in the registration of paralinguistic information. Thus, in yet another sense, the relevance of psychophysics to speech perception depends on how broadly or narrowly the field of speech perception research is defined.

In this paper I have tried to do five things. First, I have attempted to characterize the psychophysics of speech perception in terms of certain biases: heavy emphasis on the auditory modality; preoccupation with methodology; treatment of speech as a collection of sounds; neglect of the perceiver's knowledge and expectations. This characterization may well seem a caricature to those who espouse a broad definition of psychophysics. However, even though only a small part of speech perception research may fit my description, it represents an extreme (a prototype of psychophysical orthodoxy, as it were) that, though only rarely instantiated in its pure form, nevertheless exerts a certain "pull" on research in the field.

Second, I have tried to ask what it means to understand speech perception. Far from giving a satisfactory answer to this difficult question, I have made two points: Perception can be defined narrowly as a rigid process of transduction, or more broadly as a flexible process of relating the input to a knowledge base; I favor the second definition. As to understanding, it can mean producing some tangible evidence, such as a good recognition algorithm, or it can remain largely a matter of personal indulgence. My sympathies are with the former approach, but my own research has been very much within the latter.

Third, I have characterized speech perception as the application of detailed phonetic knowledge. I have argued that the mechanisms of speech perception may be quite general, but that the system as a whole is unique, thus stating a modified (possibly trivial) version of

the modularity hypothesis (Fodor, 1983; Liberman & Mattingly, 1985). This has led me further to suggest that speech perception, when considered divorced from the whole system, is a relatively shallow topic for investigation, and that a better understanding of speech perception will result indirectly from studying the whole "speech chain" (Denes & Pinson, 1963).

Fourth, I have discussed four major research questions that follow from the view taken here: Description of the phonetic knowledge; rules of its application; time course of its acquisition; and its modifiability in adulthood. The first and third of these topics are considered central to speech research. Many traditional core questions of speech perception, together with opportunities for the application for psychophysical methods, are contained in the second topic and thus are assigned a secondary role. Special emphasis is placed on articulatory and acoustic phonetics as a means for gaining insight into the language user's perceptual knowledge.

Finally, I have proposed the possibility of an articulatory psychophysics as a way of increasing the relevance of psychophysical methods to speech research.

In sum, I have painted a somewhat pessimistic picture of speech perception research, and in particular of the contribution of psychophysical approaches. This should not be taken as an assault on auditory psychophysics as such; on the contrary, the investigation of auditory function is an important area in which much excellent work is being done, as illustrated by many contributions to this workshop. What is at issue is the relevance of this general approach to the study of speech perception. If my paper stimulates discussion of this fundamental question, it will have served its purpose.

REFERENCES

1. Abramson, A., and Lisker, L. (1985). Relative power of cues: F0 versus voice timing. In V.A. Fromkin (Ed.), Phonetic linguistics. Essays in honor of Peter Ladefoged. New York: Academic. Pp. 25-33.
2. Abramson, A.S., Nye, P.W., Henderson, J.B., and Marshall, C.W. (1981). Vowel height and the perception of consonantal nasality. Journal of the Acoustical Society of America, 70, 329-339.
3. Ainsworth, W.A., and Lindsay, D. (1986). Perception of pitch movement on tonic syllables in British English. Journal of the Acoustical Society of America, 79, 472-480.
4. Atal, B.S., Chang, J.J., Mathews, M.V., and Tukey, J.W. (1978). Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. Journal of the Acoustical Society of America, 63, 1535-1555.
5. Bailey, P.J., and Summerfield, Q. (1980). Information in speech: Observations on the perception of [s]-stop clusters. Journal of Experimental Psychology: Human Perception and Performance, 6, 536- 563.
6. Bladon, R.A.W., and Lindblom, B.E.F. (1981). Modeling the judgment of vowel quality. Journal of the Acoustical Society of America, 69, 1414-1422.
7. Blomberg, M., Carlson, R., Elenius, K. and Granström, B. (1986). Auditory models as front ends in speech recognition systems. In

- J.S. Perkell and D.H. Klatt (Eds.), Invariance and variability in speech processes. Hillsdale, N.J.: Erlbaum. Pp. 108-114.
8. Blumstein, S.E., and Stevens, K.N. (1979). Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. Journal of the Acoustical Society of America, 66, 1001-1017.
9. Browman, C.P., and Goldstein, L.M. (1985). Dynamic modeling of phonetic structure. In V.A. Fromkin (Ed.), Phonetic linguistics. Essays in honor of Peter Ladefoged. New York: Academic. Pp. 35-53.
10. Bregman, A.S. (1977). Perception and behavior as compositions of ideals. Cognitive Psychology, 9, 250-292.
11. Chistovich, L.A. (1971). Problems of speech perception. In L.L. Hammerich, R. Jakobson, and E. Zwirner, (Eds.), Form & substance. Copenhagen: Akademisk Forlag. Pp. 83-93.
12. Chistovich, L.A. (1980). Auditory processing of speech. Language and Speech, 23, 67-75.
13. Chistovich, L.A. (1985). Central auditory processing of peripheral vowel spectra. Journal of the Acoustical Society of America, 77, 789- 805.
14. Chistovich, L.A., Fant, G., de Serpa-Leitao, A., and Tjernlund, P. (1966). Mimicking and perception of synthetic vowels. Quarterly Progress and Status Report (Royal Technical University, Speech Transmission Laboratory, Stockholm), 2, 1-18.
15. Chomsky, N. (1965). Aspects of the theory of syntax. Cambridge, MA: MIT Press.
16. Chomsky, N. (1968). Language and mind. New York: Harcourt, Brace & World.
17. Darwin, C.J. (1984). Perceiving vowels in the presence of another sound: Constraints on formant perception. Journal of the Acoustical Society of America, 76, 1636-1647.
18. Denes, P.B., and Pinson, E.N. (1963). The speech chain. Bell Telephone Laboratories.
19. Derr, M.A., and Massaro, D.W. (1980). The contribution of vowel duration, F0 contour, and frication duration as cues to the /juz/-/jus/ distinction. Perception & Psychophysics, 27, 51-59.
20. Elliot, L.L., Longinotti, C., Clifton, L.-A., and Meyer, D. (1981). Detection and identification thresholds for consonant-vowel syllables. Perception & Psychophysics, 30, 411-416.
21. Elman, J.L., and McClelland, J.L. (1984). Speech perception as a cognitive process: The interactive activation model. In N.J. Lass (Ed.), Speech and Language: Advances in research and practice. Vol. 10. New York: Academic. Pp. 337-374.
22. Elman, J.L., and McClelland, J.L. (1986). Exploiting lawful variability in the speech wave. In J.S. Perkell and D.H. Klatt (Eds.), Invariance and variability in speech processes. Hillsdale, NJ: Erlbaum. Pp. 360-380.
23. Flege, J.E. (in press). The production and perception of foreign language speech sounds. In H. Winitz (Ed.), Human communication and its disorders, Vol. 1. Norwood, NJ: Ablex.
24. Fodor, J. (1983). The modularity of mind. Cambridge, MA: MIT Press.
25. Fónagy, I. (1961). Communication in poetry. Word, 17, 194-218.
26. Fowler, C.A. (1984). Segmentation of coarticulated speech in perception. Perception & Psychophysics, 36, 359-368.
27. Fujisaki, H., and Kawashima, T. (1969). On the modes and mechanisms of speech perception. Annual Report of the

- Engineering Research Institute (Faculty of Engineering, University of Tokyo), 28, 67-73.
28. Fujisaki, H., and Kawashima, T. (1970). Some experiments on speech perception and a model for the perceptual mechanism. Annual Report of the Engineering Research Institute (Faculty of Engineering, University of Tokyo), 29, 207-214.
29. Ganong, W.F., III, and Zatorre, R.J. (1980). Measuring phoneme boundaries four ways. Journal of the Acoustical Society of America, 68, 431-439.
30. Gibson, J.J. (1966). The senses considered as perceptual systems. Boston: Houghton Mifflin.
31. Green, K.P., and Miller, J.L. (1985). On the role of visual rate information in phonetic perception. Perception & Psychophysics, 38, 269-276.
32. Grossberg, S. (1986). The adaptive self-organization of serial order in behavior: Speech, language, and motor control. In E.C. Schwab and H.C. Nusbaum (Eds.), Pattern recognition by humans and machines, Vol. I. New York: Academic.
33. Hary, J.M., and Massaro, D.W. (1982). Categorical results do not imply categorical perception. Perception & Psychophysics, 32, 409-418.
34. Hayek, F.A. (1952). The sensory order. Chicago: University of Chicago Press.
35. Howell, P., and Rosen, S. (1983). Production and perception of rise time in the voiceless affricate/fricative distinction. Journal of the Acoustical Society of America, 73, 976-984.
36. Hrushovski, B. (1980). The meaning of sound patterns in poetry. Poetics Today, 2, 39-56.
37. Jusczyk, P.W. (1985). On characterizing the development of speech perception. In J. Mehler and R. Fox (Eds.), Neonate cognition: Beyond the blooming buzzing confusion. Hillsdale, NJ: Erlbaum. Pp. 199-229.
38. Jusczyk, P.W. (1986). Toward a model of the development of speech perception. In J.S. Perkell and D.H. Klatt (Eds.), Invariance and variability in speech processes. Hillsdale, NJ: Erlbaum. Pp. 1-18.
39. Kasuya, H., Takeuchi, S., Sato, S., and Kido, K. (1982). Articulatory parameters for the perception of bilabials. Phonetica, 39, 61-70.
40. Kelso, J.A.S., Vatikiotis-Bateson, E., Saltzman, E.L., and Kay, B. (1985). A qualitative dynamic analysis of reiterant speech production: Phase portraits, kinematics, and dynamic modeling. Journal of the Acoustical Society of America, 77, 266-280.
41. Kent, R.D. (1973). The imitation of synthetic vowels and some implications for speech memory. Phonetica, 28, 1-25.
42. Klatt, D.H. (1979). Speech perception: a model of acoustic-phonetic analysis and lexical access. Journal of Phonetics, 7, 276-312.
43. Klatt, D.H. (1982). Prediction of perceived phonetic distance from critical-band spectra: A first step. Proceedings of the IEEE International Conference on Acoustic, Speech, and Signal Processing, Paris, France. Pp. 1278-1281.
44. Klatt, D.H. (1986). Problem of variability in speech recognition and in models of speech perception. In J.S. Perkell and D.H. Klatt (Eds.), Invariance and variability in speech processes. Hillsdale, NJ: Erlbaum. Pp. 300-319.

45. Kuhl, P.K. (1979). Speech perception in early infancy: Perceptual constancy for spectrally dissimilar vowel categories. Journal of the Acoustical Society of America, 66, 1668-1679.
46. Kuhl, P.K. (1981). Discrimination of speech by nonhuman animals: Basic auditory sensitivities conducive to the perception of speech-sound categories. Journal of the Acoustical Society of America, 70, 340-349.
47. Kuhl, P.K. (1985). Categorization of speech by infants. In J. Mehler and R. Fox (Eds.), Neonate cognition: Beyond the blooming buzzing confusion. Hillsdale, NJ: Erlbaum. Pp. 231-262.
48. Kuhl, P.K., and Meltzoff, A.N. (1982). The bimodal perception of speech in infancy. Science, 218, 1138-1141.
49. Ladefoged, P., Harshman, R., Goldstein, L., and Rice, L. (1978). Generating vocal tract shapes from formant frequencies. Journal of the Acoustical Society of America, 64, 1027-1035.
50. Liberman, A.M. (1982). On finding that speech is special. American Psychologist, 37, 148-167.
51. Liberman, A.M., and Mattingly, I.G. (1985). The motor theory of speech perception revised. Cognition, 21, 1-36.
52. Lindblom, B. (1983). Economy of speech gestures. In P.F. MacNeilage (Ed.), The production of speech. New York: Springer-Verlag. Pp. 207-246.
53. Lindblom, B., MacNeilage, P., and Studdert-Kennedy, M. (1983). Self-organizing processes and the explanation of phonological universals. In B. Butterworth, B. Comrie, and D. Dahl (Eds.), Explanations of linguistic universals. The Hague: Mouton. Pp. 181-203.
54. Lindblom, B.E.F., and Sundberg, J.E.F. (1971). Acoustical consequences of lip, tongue, jaw, and larynx movement. Journal of the Acoustical Society of America, 50, 1166-1179.
55. Linell, P. (1982). The concept of phonological form and the activities of speech production and speech perception. Journal of Phonetics, 10, 37-72.
56. Lisker, L., Liberman, A.M., Erickson, D.M., Dechovitz, D., and Mandler, R. (1977). On pushing the voice-onset-time (VOT) boundary about. Language and Speech, 20, 209-216.
57. MacKain, K.S., Best, C.T. and Strange, W. (1981). Categorical perception of English /r/ and /l/ by Japanese bilinguals. Applied Psycholinguistics, 2, 369-390.
58. Macmillan, N.A., Kaplan, H.L., and Creelman, C.D. (1977). The psychophysics of categorical perception. Psychological Review, 84, 452-471.
59. Marks, L.E. (1978). The unity of the senses. New York: Academic.
60. Marslen-Wilson, W.D., and Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. Cognitive Psychology, 10, 29-63.
61. Massaro, D.W. (in press). Categorical partition: A fuzzy logical model of categorization behavior. In S.N. Harnad (Ed.), Categorical Perception. New York: Cambridge University Press.
62. Massaro, D.W., and Cohen, M.M. (1983a). Evaluation and integration of visual and auditory information in speech perception. Journal of Experimental Psychology: Human Perception and Performance, 9, 753-771.
63. Massaro, D.W., and Cohen, M.M. (1983b). Categorical or continuous speech perception: a new test. Speech Communication, 2, 15-35.

64. Massaro, D.W., and Oden, G.C. (1980a). Speech perception: A framework for research and theory. In N.J. Lass (Ed.), Speech and Language: Advances in research and practice. Vol.3. New York: Academic. Pp. 129-165.
65. Massaro, D.W., and Oden, G.C. (1980b). Evaluation and integration of acoustic features in speech perception. Journal of the Acoustical Society of America, 67, 996-1013.
66. Mattingly, I.G. (1972). Reading, the linguistic process, and linguistic awareness. In J.F. Kavanagh and I.G. Mattingly (Eds.), Language by ear and eye: The relationships between speech and reading. Cambridge, MA: MIT Press. Pp. 133-148.
67. McGurk, H., and MacDonald, J. (1976). Hearing lips and seeing voices. Nature, 264, 746-748.
68. Meltzoff, A.N., and Moore, M.K. (1985). Cognitive foundations and social functions of imitation and intermodal representation in infancy. In J. Mehler and R. Fox (Eds.), Neonate cognition: Beyond the blooming buzzing confusion. Hillsdale, NJ: Erlb Pp. 139-156.
69. Miller, G.A., and Nicely, P. (1955). An analysis of perceptual confusions among some English consonants. Journal of the Acoustical Society of America, 27, 338-352.
70. Morais, J., Cary, L., Alegria, J., and Bertelson, P. (1979). Does awareness of speech as a sequence of phones arise spontaneously? Cognition, 7, 323-331.
71. Neisser, U. (1976). Cognition and reality: Principles and implications of cognitive psychology. San Francisco: Freeman.
72. Ohala, J.J. (1981). The listener as a source of sound change. In C.S. Marek, R.A. Hendrick, and M.F. Miller (Eds.), Papers from the parasession in language and behaviour. Chicago: Chicago Linguistic Society. Pp. 178-203.
73. Ohala, J.J. (1983). The origin of sound patterns in vocal tract constraints. In P.F. MacNeilage (Ed.), The production of speech. New York: Springer-Verlag. Pp. 189-216.
74. Pisoni, D.B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. Perception & Psychophysics, 13, 253-260.
75. Pisoni, D.B., and Lazarus, J.H. (1974). Categorical and noncategorical modes of speech perception along the voicing continuum. Journal of the Acoustical Society of America, 55, 328-333.
76. Remez, R.E., Rubin, P.E., Pisoni, D.B., and Carrell, T.D. (1981). Speech perception without traditional speech cues. Science, 212, 947-950.
77. Repp, B.H. (1981). On levels of description in speech research. Journal of the Acoustical Society of America, 69, 1462-1464.
78. Repp, B.H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. Psychological Bulletin, 92, 81-110.
79. Repp, B.H. (1983). Trading relations among acoustic cues in speech perception are largely a result of phonetic categorization. Speech Communication, 2, 341-362.
80. Repp, B.H. (1984). Categorical perception: Issues, methods, findings. In N.J. Lass (Ed.), Speech and language: Advances in Research and practice. Vol.10. New York: Academic. Pp. 243-335.
81. Repp, B.H., and Liberman, A.M. (in press). Phonetic category boundaries are flexible. In S.N. Harnad (Ed.), Categorical Perception. New York: Cambridge University Press.

82. Repp, B.H., and Williams, D.R. (1985). Categorical trends in vowel imitation: Preliminary observations from a replication experiment. *Speech Communication*, 4, 105-120.
83. Rosen, S., and Howell, P. (in press). Auditory, articulatory, and learning explanations of categorical perception in speech. In S.N. Harnard (Ed.), Categorical Perception. New York: Cambridge University Press.
84. Rosner, B.S. (1984). Perception of voice-onset-time continua: A signal detection analysis. *Journal of the Acoustical Society of America*, 75, 1231-1242.
85. Rubin, P., Baer, T., and Mermelstein, P. (1981). An articulatory synthesizer for perceptual research. *Journal of the Acoustical Society of America*, 70, 321-328.
86. Sachs, R.M., and Grant, K.W. (1976). Stimulus correlates in the perception of voice onset time (VOT): II. Discrimination of speech with high and low stimulus uncertainty. *Journal of the Acoustical Society of America*, 60, (Suppl. No. 1), S91. (Abstract)
87. Samuel, A.G. (1977). The effect of discrimination training on speech perception: Noncategorical perception. *Perception & Psychophysics*, 22, 321-330.
88. Samuel, A.G. (1982). Phonetic prototypes. *Perception & Psychophysics*, 31, 307-314.
89. Schroeder, M.R., and Strube, H.W. (1979). Acoustic measurements of articulator motions. *Phonetica*, 36, 302-313.
90. Shepard, R.N. (1980). Psychophysical complementarity. In M. Kubovy and J.R. Pomerantz (Eds.), Perceptual organization. Hillsdale, NJ: Erlbaum. Pp. 279-341.
91. Shepard, R.N. (1984). Ecological constraints on internal representation: Resonant kinematics of perceiving, imagining, thinking, and dreaming. *Psychological Review*, 91, 417-447.
92. Sidwell, A., and Summerfield, Q. (1985). The effect of enhanced spectral contrast on the internal representation of vowel-shaped noise. *Journal of the Acoustical Society of America*, 78, 495-506.
93. Silman, S., Gelfand, S.A., and Silverman, K.E.A. (1984). Late-onset auditory deprivation: Effects of monaural versus binaural hearing aids. *Journal of the Acoustical Society of America*, 76, 1357-1362.
94. Soli, S.D. (1983). The role of spectral cues in discrimination of voice onset time differences. *Journal of the Acoustical Society of America*, 73, 2150-2165.
95. Stevens, K.N., and Blumstein, S.E. (1978). Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America*, 64, 1358-1368.
96. Stevens, K.N., and Blumstein, S.E. (1981). The search for invariant acoustic correlates of phonetic features. In P.D. Eimas and J.L. Miller (Eds.), Perspectives in the study of speech. Hillsdale, NJ: Erlbaum. Pp. 1-38.
97. Studdert-Kennedy, M. (1982). On the dissociation of auditory and phonetic perception. In R. Carlson and B. Granström (Eds.), The representation of speech in the peripheral auditory system. Amsterdam: Elsevier. Pp. 9-26.
98. Studdert-Kennedy, M. (1985). Perceiving phonetic events. In W.H. Warren and R.E. Shaw (Eds.), Persistence and change: Proceedings of the first international conference on event perception. Hillsdale, NJ: Erlbaum.
99. Studdert-Kennedy, M., Liberman, A.M., Harris, K.S., and Cooper, F.S. (1970). Motor theory of speech perception: A reply to Lane's critical review. *Psychological Review*, 77, 234-249.

100. Summerfield, Q. (1979). Use of visual information for phonetic perception. *Phonetica*, 36, 314-331.
101. Summerfield, Q. (in press). Preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd and R. Campbell (Eds.), Hearing by eye. Hillsdale, NJ: Erlbaum.
102. Tallal, P., and Stark, R. (1981). Speech acoustic-cue discrimination abilities of normally developing and language-impaired children. *Journal of the Acoustical Society of America*, 69, 568-574.
103. Tatham, M.A.A. (1980). Phonology and phonetics as part of the language encoding/decoding system. In N.J. Lass (Ed.), Speech and language: Advances in research and practice. Vol.3. New York: Academic. Pp. 35-73.
104. Titchener, E.B. (1909). Lectures on the elementary psychology of the thought-process.
105. Toulmin, S. (1972). Human understanding: The collective use and evolution of concepts. Princeton: Princeton University Press.
106. Tuller, B., and Fowler, C.A. (1980). Some articulatory correlates of perceptual isochrony. *Perception & Psychophysics*, 27, 277-283.
107. Tyler, R.S., Summerfield, Q., Wood, E.J. and Fernandes, M.A. (1982). Psychoacoustic and phonetic temporal processing in normal and hearing-impaired listeners. *Journal of the Acoustical Society of America*, 72, 740-752.
108. Vinegrad, M.D. (1972). A direct magnitude scaling method to investigate categorical versus continuous modes of speech perception. *Language and Speech*, 15, 114-121.
109. Warren, R.M. (1981). Chairman's comments. In T. Myers, J. Laver, and J. Anderson (Eds.), The cognitive representation of speech. Amsterdam: North Holland. Pp. 34-37.
110. Watson, C.S., Kewley-Port, D., and Foyle, D.C. (1985). Temporal acuity for speech and nonspeech sounds: The role of stimulus uncertainty. *Journal of the Acoustical Society of America*, 77, (Suppl. No. 1), S27. (Abstract)
111. Welsh, L.W., Welsh, J.J., and Healy, M.P. (1983). Effect of sound deprivation on central hearing. *Laryngoscope*, 93, 1569-1575.
112. Werker, J.F., and Tees, R.C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, 7, 49-63.
113. Werker, J.F., Gilbert, J.H.V., Humphrey, K., and Tees, R.C. (1981). Developmental aspects of cross-language speech perception. *Child Development*, 52, 349-355.
114. Wood, C.C. (1976). Discrimination, response bias, and phoneme categories in discrimination of voice onset time. *Journal of the Acoustical Society of America*, 60, 1381-1389.
115. Yates, J. (1985). The content of awareness is a model of the world. *Psychological Review* 92, 249-284.

CENTRAL AND PERIPHERAL PROCESSES IN THE PERCEPTION OF SPEECH AND NONSPEECH SOUNDS*

Neil A. Macmillan**, Louis D. Braida, and Rina F. Goldberg***

Research Laboratory of Electronics, MIT

In this chapter, we apply a psychophysical theory to several speech and nonspeech continua. Our approach meets many of Repp's (this volume) criteria for psychophysics: it is auditory-based, methodological, and concerned with speech sounds. We deviate from Repp's characterization, however, in that our theory is not psychoacoustic, that is, is not limited to relations between stimulus characteristics and sensations.

The theory generalizes the Durlach and Braida (1969; Braida and Durlach, 1986) intensity perception model; the domain of the theory is resolution experiments, in which observers discriminate or classify stimuli. Resolution depends on both peripheral factors, which are task-independent and presumed to be related to neural transduction in the auditory periphery; and central factors, which are task dependent, and can be separated into those affecting sensitivity and those affecting response bias. The primary central process with which the theory is concerned is memory.

Two distinct memory processes play a role in comparing sounds. Context-coding compares sounds to a general context of sounds; this process is highly sensitive to the context width of sounds which must be compared, but insensitive to the time between successive sounds. Trace-maintenance compares one sound with the image of a previous sound: this process is highly sensitive to the passage of time, but not to context.

These distinctions are not unfamiliar to students of speech perception. The context-coding/trace-maintenance distinction is related (but not identical) to that between phonetic and "auditory" coding. In particular, the theory is similar in many respects to the dual-processing theory of Fujisaki and Kawashima (1969, 1970) and Pisoni (1975). Experimental tests to distinguish these theories are possible (see Macmillan, in press; Cowan and Morse, 1986), but they

*This work was supported by grants from NSF and the PSC-CUNY Award Program to the first author, and a grant from NSF to the second author. We are grateful to Nelson Cowan, Ken Grant, Burt Rosner, and Rosalie Uchanski for helpful comments on previous drafts.

**Present address: Department of Psychology, Brooklyn College, Brooklyn, NY, USA 11210.

***Now at Bell Laboratories, Holmdel, NJ.

are not our focus here. Rather, we take advantage of the scope of the Durlach-Braida theory to discuss a fairly wide range of issues arising from "phoneme-resolution" experiments, in which stimuli differ primarily in one phoneme, e.g., between /ba/ and /pa/ or between /i/ and /I/. Ades (1977) was the first to describe speech perception data of this sort in terms of the Durlach-Braida framework.

We first discuss experiments in which, according to the theory, peripheral sensitivity is being measured. Here, the theory offers a method by which the existence of natural auditory sensitivities can be tested. Second, we present a specific model for context coding that provides an account of some "categorical" phenomena, aspects of perceptual learning, and some contrast effects. Third, we describe the effects of interstimulus interval (ISI) on sensitivity as trace-maintenance attenuated by context coding.

Throughout, we use the theory to compare speech data with intensity data, seeking to characterize necessary quantitative changes (in the values of parameters) and qualitative ones. We focus on heavily-studied speech continua: consonant-vowel syllables in which place of articulation or voicing is varied, and steady-state English vowels varying from /i/ to /I/ (and sometimes /e/).

PERIPHERAL PROCESSES AND BASIC SENSITIVITY

By basic sensitivity, we mean the best possible resolution performance. Since performance is limited by both unavoidable sensory factors and task-dependent memory effects, best performance occurs when memory limitations are minimal. In general, listeners use both context and trace coding. Memory requirements can be minimized by using a small stimulus range and a short ISI, so that listeners can employ a combination of two highly efficient memory modes, and memory variance will be very small.

A paradigm in which memory variance is virtually eliminated is fixed discrimination, in which the only two stimuli ever presented in a block of trials are those being discriminated. Designs in which two stimuli are presented per trial are best: the trace mode is not used in one-interval paradigms, while in three-interval paradigms (like ABX and oddity) the time between stimuli necessarily becomes large and interference effects may occur. The two most popular two-interval paradigms are two-interval forced-choice (2IFC) and variable-standard same-different (AX). Models exist for separating sensitivity and bias in each of these paradigms (Green and Swets, 1974; Macmillan, Kaplan and Creelman, 1977). Fixed discrimination paradigms are not often used with speech stimuli; roving discrimination paradigms, in which pairs of sounds are drawn from a wider range, are far more common. We shall see, however, that roving experiments involve greater memory limitations than fixed experiments.

Intensity

Data on intensity perception illustrating the difference between roving and fixed discrimination are shown in Figure 1(a) (data from Rabinowitz, Lim, Braida and Durlach, 1976; Berliner, Durlach and Braida, 1977). The upper curve summarizes data for the fixed paradigm, the lower curve from the roving paradigm. For a given decibel

difference between intensities, sensory limitations, revealed by the fixed data, decrease gradually as overall intensity increases, in accordance with the famous "near-miss" to Weber's Law (McGill and Goldberg, 1968). Performance in roving discrimination is substantially worse, and appears not to be as simply related to intensity. An interpretation of the roving discrimination curve (and an explanation of panel (b)) will be provided in a later section.

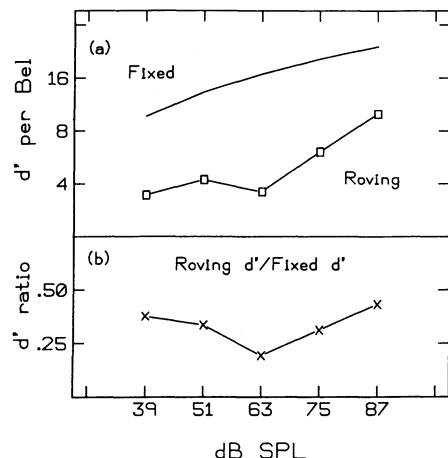


FIGURE 1, (a) Fixed and roving discrimination of pure tone intensity. (b) Roving discrimination relative to the fixed discrimination baseline. Data from Rabinowitz, et al. (1976) and Berliner, et al. (1977).

Vowels

Figure 2 compares roving and fixed discrimination for steady-state synthetic vowels. In panel (a) (based on Goldberg, 1986), these range from /i/ to /I/. Sensitivity in fixed discrimination is higher than in roving, and the two curves also differ significantly in shape: the gentle peak observed in roving discrimination is not evident in fixed discrimination. Panel (b) (from Goldberg, et al., 1985) displays data for the wider vowel range /i/ to /I/ to /ɛ/, and compares fixed discrimination with identification, in which, theoretically, the context-coding mode is used exclusively. Again, fixed discrimination is better, and lacks the peaks observed in the identification task.

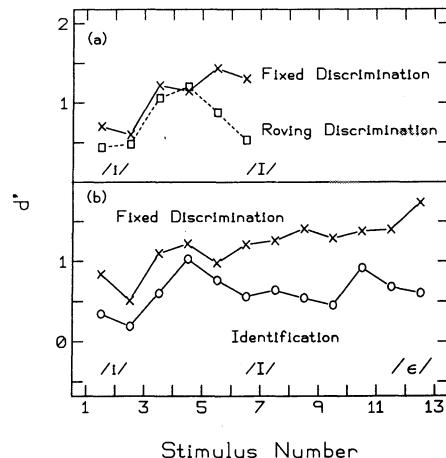


FIGURE 2, Fixed and roving discrimination of vowels from /i/ to /I/ (Goldberg, 1986). (b) Fixed discrimination and identification of vowels /i-I-ɛ/ (Goldberg, et al., 1985)

Consonants

Goldberg (1986) measured fixed discrimination on a bilabial voice-onset time (VOT) continuum. Figure 3(a) shows that performance was better than in roving discrimination or identification, but followed the same qualitative pattern: sensitivity peaked in mid-range.

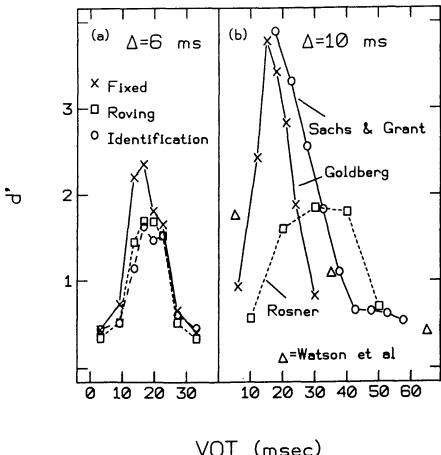


FIGURE 3, (a) Fixed and roving discrimination and identification of bilabial consonants differing in VOT (Goldberg, 1986). (b) Fixed discrimination of three VOT continua: bilabial (Goldberg, 1986), alveolar (Rosner, 1984), and velar (Sachs and Grant, 1976). All d' values have been estimated for 6-ms stimulus differences in panel (a), 10-ms differences in panel (b).

Three other experiments have measured fixed discrimination for VOT continua: Watson, Kewley-Port, Foyle (1985) studied a bilabial continuum, Rosner (1984) an alveolar one, and Sachs and Grant (1976) used velar stimuli. A comparison of the data from all studies is shown in Figure 3(b), for a constant 10-ms difference in VOT. The data of Watson, et al. and of Sachs and Grant are consistent with Goldberg's: neither found a discrimination peak, because neither collected data near 15 ms. Except for Rosner's data, the pattern of results thus reflects high basic sensitivity for VOTs in the range 10-20 ms, independent of place of articulation. Rosner's listeners had less training than the others, and may have had a wider-than-optimal effective range; evidence that memory limitations were unusually large in Rosner's experiment will be presented later. Clearly this interpretation of the data requires further test.

Basic sensitivity peaks on a place-of-articulation continuum can be seen in Figure 4 (from Stephens, 1986). The stop-consonant continua of voicing and place have been those most reliably reported

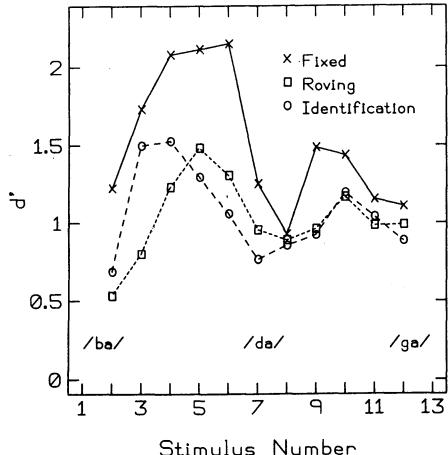


FIGURE 4, Fixed and roving discrimination and identification of consonants differing in place of articulation (Stephens, 1986).

to yield sensitivity peaks, and it appears that these peaks result from regions of high auditory sensitivity, not memory processing.

Fixed discrimination data are an essential baseline against which the effects of these cognitive processes can be discerned. Psychoacousticians often refer results to a different baseline, Weber's Law. Data which satisfy Weber's Law are often seen as needing no further explanation, whereas those that do not may generate complex theorizing (as has been the case for the near-miss). When speech discrimination functions have been compared to Weber's law (see, e.g., Miller, Wier, Pastore, Kelly and Dooling, 1976; Pastore, in press), only those parts of the function that fail to fit Weber's Law have become the grist for theory. Weber's law plays no such special role in our theory; that it provides a good description of intensity perception in several modalities does not imply that it is hidden in the data on VOT, noise-buzz, or vowels. In our theory, all fixed discrimination data require psychoacoustic explanation, whether they resemble Weber's Law, display a peak, or are monotonic.

Fixed discrimination data are the appropriate input for psychoacoustic theories, or those based on peripheral physiology. That is, the data just presented would lead us to expect, as a working hypothesis, to observe peaks in consonant, but not vowel sensitivity at the level, say, of the auditory nerve. There is thus a need for more psychophysical data of the fixed type. Failure to appreciate the fixed/roving distinction can lead theorists to offer peripheral explanations for central events.

CONTEXT CODING

Context variance and stimulus range

In identification experiments, a single sound is presented on each trial, and the observer classifies it by making a comparison with a stable context of sounds heard in the past. Our ability to identify sounds is poorer than our ability to discriminate them, to a degree that depends on the stimulus range. In intensity experiments, identification of intensity is only slightly worse than discrimination when the range is small (Pynn, Braida and Durlach, 1972), but declines by a factor of about four for a large (60 dB) range. We call this phenomenon the range effect (1). Sensitivity in identification is limited by both sensory variance and range-dependent context variance, which add to limit performance. Fixed discrimination and identification performance are sufficient to calculate the relative context variance, which is the context variance measured in units of the sensory variance. (Quantitative predictions of the model are given in the Appendix. See equations (A1) to (A3).)

Context variance can be estimated in this way only if d' values can be obtained from identification, that is, if observers make errors. In many speech experiments, the possible responses correspond to phonetic categories, of which there are usually only two or three, and many stimuli are responded to with unanimity. Much more information can be gathered if more responses are permitted, either by giving listeners multiple response alternatives directly, or by combining the phonetic response with a rating scale.

Range effect. Goldberg, Macmillan, and Braida (1985) measured identification of vowels /i-I-ɛ/, and also of vowels varying over half that range; results are shown in Figure 5. Identification d' has been normalized by dividing by d' in fixed discrimination, but the comparison between full- and half-range conditions is unaffected. There is clearly a range effect for this continuum: reducing the range improved total d' , the total sensitivity across the stimulus set, by an average of 30 percent. The relative context variance is 3.0 for the full range, but averages only 0.94 for half-range identification. These are the values which would be obtained in intensity experiments with ranges of about 14 and 8 decibels.

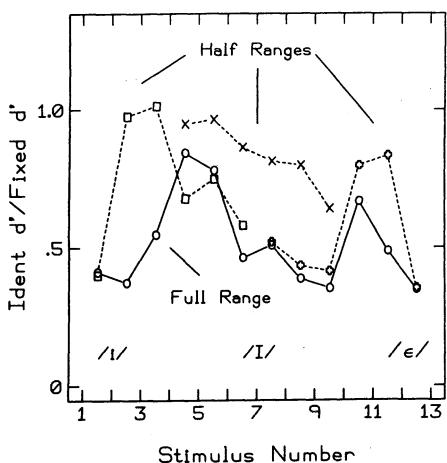


FIGURE 5, identification d' , relative to fixed discrimination for vowels /i-I-ɛ/ and three subranges (Goldberg, et al., 1985).

We are somewhat limited in studying the range effect with synthetic speech stimuli since, unlike tones differing in intensity, speech continua cannot be extended arbitrarily. On the other hand, natural speech is multidimensional. A suggestive hypothesis is that the context variance would be increased if speech stimuli were allowed to vary on dimensions other than the one most relevant to the observer's judgment. We (Macmillan and Braida, 1985) have explored this idea by comparing identification for /i-I-ɛ/ vowels which either did or did not vary in F0; multiplicity of tokens reduced total d' by about 20 percent. Since the token effect on fixed discrimination was far less, introducing F0 uncertainty may indeed be similar to increasing the stimulus range.

Comparing context variance across continua

When the range can be measured quantitatively, as in intensity experiments, equation (A2) in the Appendix provides a direct test of the theory (see Braida and Durlach, 1972), but for other continua range must be measured indirectly. One possible measure is total discrimination d' , the number of fixed discrimination jnds required to span the continuum. In intensity, total d' increases (nonlinearly) with the range, so this is a natural translation.

Does total discrimination d' predict total identification d' across stimuli? If so, then context memory is independent of stimulus continuum. This was the conjecture of Ades (1977), who attributed differences between consonant and vowel discrimination to a larger total d' for vowels. Figure 6 shows that this hypothesis is probably too simple.

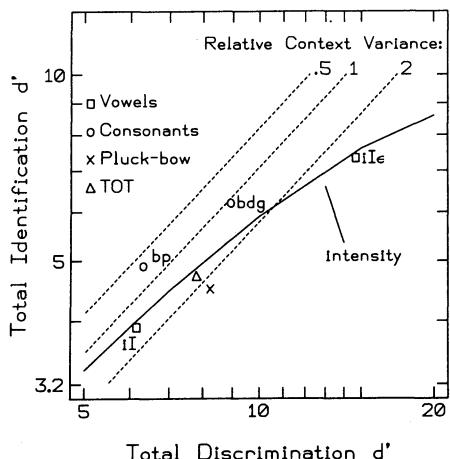


FIGURE 6. Total identification d' vs. total discrimination d' ; dashed lines indicate values of relative context variance. Seven continua are plotted. Data are from Goldberg, 1986 (/ba-pa/ and /i-I/); Goldberg, et al., 1985 (/i-I-e/); Stephens, 1986 (/ba-da-ga/); Macmillan, 1979 (pluck-bow); Macmillan, in preparation (tone-onset-time); and Braida and Durlach, 1986 (intensity). See text for details.

Shown in the figure is identification d' versus fixed discrimination d' for seven continua studied in our laboratory. The solid curve shows how identification of intensity increases with total discrimination d' . Lines of constant relative context variance in the figure provide a measure of the range effect: as total d' increases, so does relative context variance, so that sensitivity for any given stimulus difference decreases with range.

Figure 6 allows a number of useful comparisons among stimulus sets. First, continua do differ in total d' . The three-category vowel continuum /i-I-e/ has a larger range than its two-category subset /i-I/; the range of /ba-da-ga/ is greater than that of /ba-pa/.

Second, these range differences, measured on the intensity continuum, can account for memory differences in vowel resolution: data points for the two vowel continua lie very close to the intensity curve.

Third, range differences cannot reconcile the consonant data with the other continua. Total discrimination d' for /i-I/ is about the same as for /ba-pa/, but the consonants are much better identified (relative context variance is 0.65 vs 1.48). Both consonant continua are characterized by small context variance. Since small context variance means that identification is nearly equivalent to discrimination, this conclusion is similar in spirit to the assertion that consonants are categorically perceived, although, since our identification task permits responses other than phonetic category names, our conclusion is somewhat less extreme.

Fourth, two nonspeech continua developed in attempts to mimic consonantal ones, pluck-bows (Cutting and Rosner, 1974) and tone-onset time (Pisoni, 1977), have relative context variance of the order of vowels, not consonants.

Finally, whether the basis of comparison is total d' or relative context variance, all continua correspond to modest ranges of intensity. For a 54-dB intensity range, total discrimination d' is over 100 and relative context variance is about 50. The speech continuum studied which had the largest range (and largest relative context

variance), /i-I-ɛ/, is equivalent to an intensity continuum with about a 15-dB range

Perceptual anchors

We turn now from our examination of overall or average memory sensitivity to a consideration of the variation of sensitivity within the stimulus range, starting once again with the intensity continuum. Recall that fixed and roving intensity discrimination have different trends, as shown in Figure 1(a). Since fixed discrimination is a sensory task, we can uncover the pattern of memory processing by plotting the ratio of roving d' to fixed d' across the range. The result, for intensity, is shown in Figure 1(b): memory efficiency is substantially better at the edges, near the strongest and weakest stimuli.

This sensitivity edge effect suggests that comparative judgments are being made by employing the edges of the stimulus range as perceptual anchors and measuring the distance of a given intensity from these references with a precision that decreases as the distance from the reference increases (Braida, Durlach, Lim, Berliner, Rabinowitz and Purks, 1984). Intuitively, the extreme stimuli are memorable, i.e., are coded with small variance. The Perceptual Anchor Model of context coding formalizes this idea, and provides a good account of intensity identification data. The model predicts sensitivity peaks only for the larger ranges; for smaller ranges, anchors are still used, but all stimuli are near an anchor, so performance is uniformly elevated.

Sensitivity peaks for vowels and consonants. To discover whether there are perceptual anchors for other dimensions, we follow the strategy illustrated in Figure 1: roving (or identification) sensitivity is plotted relative to basic sensitivity, and local maxima serve to indicate perceptual anchors. Roving discrimination and identification should reveal the same pattern, since trace coding (the only process that distinguishes them) is equally effective at all points in the range.

The relevant data for the /i-I-ɛ/ vowel continuum have already been presented (see Figure 5), and show peaks between phonemic categories. The boundary stimuli appear to be remembered almost perfectly--identification d' is virtually as large as fixed discrimination d' --and the accuracy with which other stimuli are judged decreases with their distance from these anchors. We have already determined that the sensitivity peak in roving discrimination vowel data is a cognitive effect rather than a sensory one; we now conclude that this peak results from a decision process in which vowel tokens are compared with a well-remembered perceptual reference near the boundary.

The consonant data of Goldberg (1986) for VOT and Stephens (1986) for place can also be used to probe for anchors. The resulting picture (Figure 7) is different from that for vowels in one important respect: the peaks appear to be located near good exemplars, rather than between them. Context coding of consonants is with reference to well-remembered prototypes, rather than with reference to boundaries.

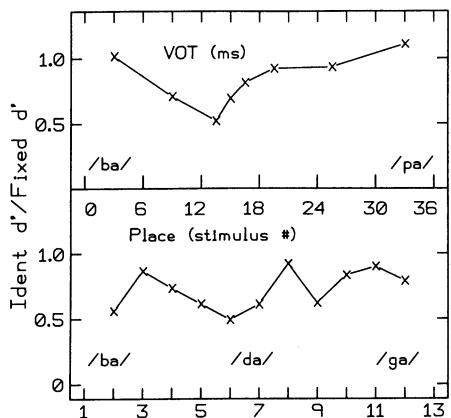


FIGURE 7, (a) Identification sensitivity relative to fixed discrimination for bilabial VOT (Goldberg, 1986).

(b) Identification sensitivity relative to fixed discrimination for place of articulation (Stephens, 1986).

Perceptual learning

Perceptual learning is important in resolution tasks, and the appropriate way to characterize the changes occurring during the course of an experiment has been a topic of some controversy in speech perception. Most writers (e.g., Carney, Widin and Viemeister, 1977) have stressed changes in discrimination sensitivity which occur with practice, and we suggested earlier that such improvements may underlie discrepancies among studies of VOT continua (see Figure 3). But improvements in discrimination are often accomplished by reducing experimental uncertainty, making the task more like fixed discrimination. We consider in this section a simple hypothesis with a different focus: perceptual learning in the tasks we have been discussing consists of changes in context coding. Two types of changes can occur: improvements in sensitivity, and shifts in response bias. Improvements could occur because the variance of anchors is reduced, or because mean anchor locations become more optimal. To evaluate this hypothesis, we first consider direct measurements of overall context variance during training, then changes in the pattern of sensitivity which might result from shifts in perceptual anchors.

Reducing context variance. We have examined changes in fixed discrimination and identification during the first four hours of listening to isolated vowels (Macmillan and Braida, 1985). Identification improved 17% while fixed discrimination improved only 5%; the relative context variance declined from about 0.7 to 0.3.

A similar analysis of Rosner's (1984) data leads to the conflicting conclusion that VOT depends on a central anchor (Macmillan, in press). An important difference between Rosner's experiment and those in our laboratory is that Rosner's subjects, probably because they were untrained, were inefficient in context memory: relative context variance was 8.1, compared for 0.65 and 1.06 for Goldberg and Stephens. We argue below that perceptual learning in these tasks consists largely of improved context coding, but it is an empirical matter whether anchor locations change with experience, as suggested by this set of results.

Inducing anchors. According to the theory, context coding consists of comparisons between sensations and perceptual anchors. Learning effects can be studied by providing listeners with unusual distributions of sounds, to see if new anchors can be created. This strategy (but not necessarily this rationale) has been applied to both intensity and speech continua.

Chase, Bugnacki, Braida, and Durlach (1983) compared intensity identification (using a 36-dB range) under three conditions: (1) all stimuli were presented with equal probability; (2) the middle stimulus was presented on one-third of all trials; and (3) the end stimuli were each presented on one-fifth of all trials. When trial-to-trial feedback was provided, condition (2) produced only an insignificant improvement in sensitivity near the frequently-presented stimulus. Conditions (2) and (3) did, however, produce strong bias changes: listeners adjusted their criteria so as to improve their overall identification score, i.e., increased their bias towards responding with one of the frequently-presented stimuli.

Sawusch, Nusbaum, and Schwab (1980) performed an analogous identification experiment using the /i-I/ continuum. They compared the equal-presentation condition with conditions in which either the extreme /i/ or the extreme /I/ was presented on 40 percent of the trials. Unlike Chase et al., they observed sensitivity changes: when /i/ was presented frequently, sensitivity improved at the /i/ end of the continuum; but no analogous change occurred in the frequent-/I/ condition. In both conditions (but especially for the frequent-/I/ case), there was a bias shift, but in the opposite direction of those found by Chase et al., (1983): observers moved their criteria toward the frequently-presented stimuli, thus decreasing their use of the corresponding responses. The result was consisted with Parducci's (1965) range-frequency theory, which Chase et al. were able to reject for their data.

The discrepancy between the intensity and vowel data can probably be attributed to instruction differences: unlike the listeners of Chase et al., those of Sawusch et al. were not informed of the presentation probabilities, and did not receive trial-by-trial feedback. Macmillan and Braida (1985) replicated the Sawusch et al. (1980) experiment, but provided their listeners with feedback and knowledge of presentation probabilities. The pattern of results was in agreement with the procedurally similar Chase et al (1983) intensity experiment: sensitivity changes were small and inconsistent, but listeners increased their use of responses near the one corresponding to the modal stimulus.

Whether frequent presentation of a stimulus should be expected to strengthen an anchor is an arguable question. Intensity anchors exist just outside the range of stimuli; the anchor on the /i-I/ continuum is interior, i.e., at a location where exemplars occur infrequently. We know of no attempt to induce anchors in midrange by using a bimodal stimulus distribution, but this may be a good analog to normal speech perception development.

Providing explicit standards. Berliner et al. (1977) showed that the presentation of an explicit standard improved performance in large-range intensity identification experiments only when the standard

did not coincide with an anchor. The theoretical effect of a standard, according to the anchor model, has been explored by Khazatsky (1985) for the intensity continuum. His model assumes that observers estimate the perceptual distance of a sound from both the standard and the nearest edge anchor. A standard can improve performance in midrange, provided that the variability of the internally-represented standard is not too large, and the locations of the edge anchors and the anchor induced by the standard are not highly correlated. Under these conditions, the standard acts roughly to enable the listener to bisect the effective stimulus range.

A number of investigators have measured identification of speech sounds which have been preceded by other sounds, but most have not tried to determine whether sensitivity changed. In many "selective adaptation" studies, primary interest has been in shifts in the identification function. In the original experiment of this type, Eimas and Corbit (1973) showed that repeated presentation of /ba/ shifted identification of /ba-pa/ stimuli towards /pa/. The use of repeated stimuli appears to be unnecessary--Diehl, Elman, and McCusker (1978) found that "adaptation" effects were equally strong when a single standard was used--so these experiments may be viewed as conventional identification with a standard. In a signal detection analysis, Elman (1979) concluded that adaptation effects were entirely due to response bias: no sensitivity effects of the sort found by Berliner et al. (1977) for intensity occurred.

By using a multi-response identification task and SDT analysis, we have examined sensitivity and bias effects separately, for both the /i/-/l/ vowel continuum and consonants differing in VOT. Goldberg (1986) found only small effects of standards on sensitivity. The primary effect of an endpoint standard, for both continua, was to shift criteria towards that standard, and thus increase responding in categories at the opposite end. Thus when the standard in vowel identification was /i/, stimulus 1, more responses fell in higher-numbered categories than when there was no standard; the opposite was true for /l/. The pattern of results was the same for the consonants. Why standards should locally improve sensitivity for intensity but not for our speech continua is unclear, but one testable hypothesis is that the range (and thus the context variance) is too small in speech. Berliner, et al. (1977) found their effects for a 54-dB range.

TRACE MAINTENANCE

The passage of time or the existence of interfering sounds or tasks limits our ability to compare sounds separated by an interval of time. These degradations suggest the operation of a trace-maintenance mechanism which provides a trace of the one sound for comparison with a subsequent sound. The variance of the trace increases linearly with time. If only the trace mode is used, the rate at which sensitivity decreases with increasing ISI depends only on the ratio between trace variance and sensation variance, which we call the relative trace variance. The trace-maintenance mechanism is unaffected by the range of stimuli, and thus exhibits neither a range effect nor a sensitivity edge effect.

When ISI is varied in two-interval intensity discrimination, sensitivity does indeed decline with ISI, but it is immediately clear that listeners are not relying entirely on the trace mode. In fixed discrimination, relative trace variance equals 1.0 at an ISI of about 3.5 s, implying that resolution degrades only by a factor of three in an interval of a minute. This low estimate arises because listeners can also use the context mode to compare sounds in roving discrimination. A better description of decay effects can be obtained under the assumption that roving discrimination uses both trace and context modes, so that apparent time effects depend on the context-mode parameters of range and anchor location. (See Appendix, equations (A4) and (A5).)

Trace decay and stimulus range

Intensity. Figure 8 shows that the apparent decay rate in roving discrimination of intensity increases with stimulus range. In small-range experiments, the context mode is especially useful, and trace decay effects are suppressed. With large ranges, the utility of the context-coding mechanism is reduced, and decay effects are much greater. The curves fit to the data assume a single decay parameter and optimal combination of trace-maintenance and context-coding by the listener. The relative trace variance equals 1.0 when T is only about 0.05 sec. Thus estimates of trace decay which ignore the use of the context-coding mechanism underestimate the trace decay rate by a factor of roughly 70.

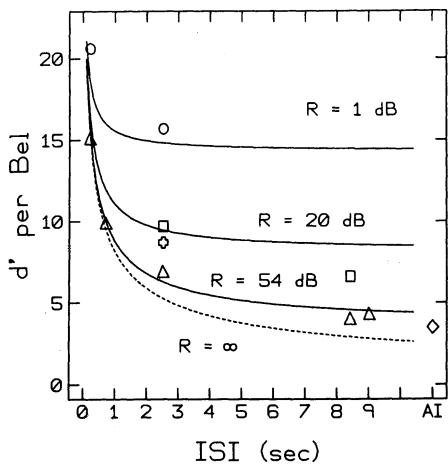


FIGURE 8. Roving intensity discrimination d' for different ISIs, with predictions of the model. Dashed curve is prediction for trace mechanism alone. Data from Berliner and Durlach (1973), figure from Braida and Durlach (1986).

second to as much as 30 seconds. Cowan (1984) has recently proposed that there are both short- and long-term auditory stores, the short

Speech. In speech discrimination experiments, the apparent rate of decline of sensitivity with ISI is faster for vowels than for consonants. Pisoni (1973) concluded from this result that different phonetic classes have different decay rates, but our model suggests that such decay rates should not be taken at face value, since they ignore the contribution of context-coding. Vowels have a greater relative context variance than consonants (see Figure 6), so if the same trace decay parameter applied to both continua, we would expect to observe a smaller rate of decline for consonants.

The variability of estimates of the duration of auditory memory from different types of experiments, with different stimuli, is substantial, ranging from a fraction of a

store holding unanalyzed input for perhaps one-quarter of a second, the long store holding partially-analyzed input for a longer period. Cowan finds evidence for short-term trace memory in tasks on auditory persistence, integration time, and backward and forward masking. Long-term trace memory is implicated in experiments on the "suffix effect" (e.g., Crowder, 1982), and selective attention to left versus right ear input (Treisman, 1964) or spatial location (Darwin, Turvey and Crowder, 1972). Our theory provides an alternative account. Cowan's short-term tasks offer a very limited range of stimuli and no distractions, making context coding a useful technique. In his long-term tasks, context-coding is degraded by (a) providing a wide range of potential stimuli, and (b) requiring the listener to focus attention on competing sources of information. A single estimate of decay rate (the longer one) may be enough to account for all the data, if the contribution of context coding to the short-term tasks is taken into account.

Trace decay and anchors

An alternative account of these data would postulate a trace decay rate that depends on the range, and (possibly for that reason) on stimulus continuum. This approach can be ruled out for intensity, however, because it fails to predict the resolution edge effects obtained in roving discrimination. Figure 9 shows that sensitivity declines less with ISI at the edges of the intensity range, where there are perceptual anchors, than elsewhere. The hypothesis that roving discrimination employs both context and trace coding explains the dependence of apparent trace decay on both range (Figure 8) and location within the range (Figure 9).

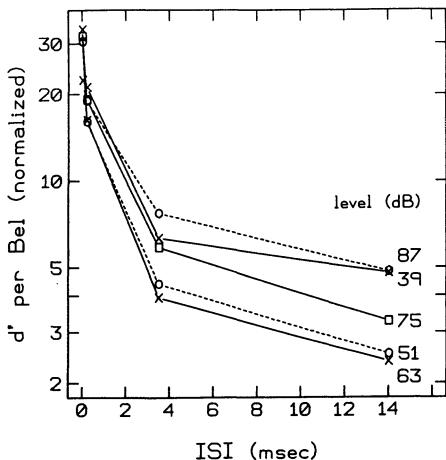


FIGURE 9, Roving discrimination of intensity (relative to fixed discrimination baseline) for different values of ISI. The parameter is the stimulus level, in a large-range experiment with tone duration 1.25 sec. Data replotted from Berliner, et al. (1977)

Differences in decay rates at different points on a stimulus dimension have also been reported for speech continua. Pisoni (1973) and, more recently, Cowan and Morse (1986) have studied the /i-I/ vowel dimension, and found that decay appears greater for stimuli near the midrange "boundary" than at the edges. (Cowan and Morse reached a different conclusion because they measured sensitivity differences, rather than the ratios implied by the use of d'). Pisoni's (1973) dual-process conclusion was that "phonetic memory" decays more slowly than "auditory memory". This effect is exactly analogous to the intensity effect shown in Figure 9. Recall that, according to our analysis, the /i-I/ continuum has a perceptual anchor in midrange; we therefore expect that apparent decay will be smaller there,

since efficient context-coding masks trace effects. Both the vowel and the intensity data support the conclusion that trace decay rates are constant, and seem variable only because listeners use the trace mode in combination with the context mode.

Contrast effects in the trace mode

As the memory trace decays, its variance increases; what about its mean? The theory says nothing about potential nonstationarities, but data from both intensity and speech suggest that decay is biased, that is, the mean does not stay the same.

Berliner et al. (1977) observed a bias edge effect in roving discrimination: listeners tended to hear the first of two high-intensity stimuli as softer, but to hear the first of two low-intensity stimuli as louder. A trace decay process the mean of which moves towards the middle of the range, away from the anchors, could account for this pattern. Such a process would not change sensitivity: on half the 2IFC trials, the biased decay harms performance, but on the other half it helps.

Some contrast effects observed in speech experiments can be described similarly. Repp, Healy, and Crowder (1979) performed a roving discrimination task with vowels, and an identification task using the same two-interval stimulus presentation. They reported strong contrast effects. Cowan and Morse (1986) have proposed a model for vowel discrimination in which the internal representations of steady-state vowels decay over time in the direction of /ə/, the neutral vowel. Thus an /ɪ/ standard decays away from other stimuli on the /i-I/ continuum, and the greater separation improves performance; but an /i/ standard decays towards the rest of the continuum, lowering performance. The model is consistent with the vowel data of Goldberg (1986) and Repp et al. (1979), and is similar to the bias edge effect. If these contrast effects are indeed characteristics of the trace mode, then it should be possible to shrink them by reducing the stimulus range, because listeners are less likely to be in the trace mode when the range is small.

SUMMARY

A major advantage of the present framework is the large set of questions about sensory and memory effects which may be posed within it. The present, preliminary application of the model has, however, suggested some substantive answers. The following short list of such conclusions illustrates the variety of problems for which the model can be useful tool:

- (1) On at least two consonant continua, there are regions of high basic sensitivity in midrange; on at least one vowel continuum, there is none.
- (2) Vowel resolution declines with increasing stimulus range to the same degree that intensity resolution does. Performance also declines with increasing variation on dimensions orthogonal to judgment.
- (3) Memory for vowels is about as good as memory for intensity, when the range in jnd's is equated, but memory for consonants is better.

(4) In memory-limited tasks, vowel stimuli are best remembered near a category boundary, whereas consonant stimuli are best remembered near good phonetic exemplars.

(5) Altering the stimulus distribution or providing explicit standards for comparison primarily affects response bias, not sensitivity.

(6) Estimates of memory decay inferred by varying ISI are range-dependent.

NOTE

1. In Parducci's (1965) range-frequency model, a shift in response selection as a function of stimulus subset is called a "range effect". Rosen (1979) has reported such an effect in consonant identification: a stimulus in the middle of a ten-step /ba-da/ continuum was more likely to be called "da" in the context of stimuli 1-5 than when the possible stimuli were 4-8. By range effect, we mean the relation between context width and sensitivity, as described by equation (A2); the Parducci-Rosen usage applies to changes in the relation between the average stimulus value and responding, which may be entirely due to response bias.

REFERENCES

1. Ades, A.E. (1977). Vowels, consonants, speech and nonspeech. Psychological Review, 84, 524-530.
2. Berliner, J.E. and Durlach, N.I. (1973). Intensity perception. IV. Resolution in roving-level discrimination. Journal of the Acoustical Society of America, 53, 1270-1287.
3. Berliner, J.E., Durlach, N.I., and Braida, L.D. (1977). Intensity perception. VII. Further data on roving level discrimination and the resolution and bias edge effects. Journal of the Acoustical Society of America, 61, 1577-1585.
4. Braida, L.D. and Durlach, N.I. (1986). Peripheral and central factors in intensity perception. In: G.M. Edelman, W.E. Gall, and W.M. Cohen (Eds.), Functions of the auditory system. New York: Wiley.
5. Braida, L.D., Durlach, N.I., Lim, J.S., Berliner, J.E., Rabinowitz, W.M., and Purks, S.R. (1984). Intensity perception. XIII. Perceptual anchor model of context coding. Journal of the Acoustical Society of America, 76, 722-731.
6. Carney, A.E., Widin, G.P., and Viemeister, N.F. (1977). Noncategorical perception of stop consonants differing in VOT. Journal of the Acoustical Society of America, 62, 961-970.
7. Chase, S., Bugnacki, P., Braida, L.D., and Durlach, N.I. (1983). Intensity perception. XII. Effect of presentation probability on absolute identification. Journal of the Acoustical Society of America, 73, 279-284.
8. Cowan, N. (1984). On short and long-term auditory stores. Psychological Bulletin, 96, 341-370.
9. Cowan, N. and Morse, P.A. (1986). The use of auditory and phonetic memory in vowel discrimination. Journal of the Acoustical Society of America, 79, 500-507.
10. Crowder, R.G. (1982). A common basis for auditory sensory

- storage in perception and immediate memory. Perception & Psychophysics, 31, 477-483.
11. Cutting, J.E. and Rosner, B.S. (1974). Categories and boundaries in speech and music. Perception & Psychophysics, 16, 564-570.
 12. Darwin, C.J., Turvey, M.T., and Crowder, R.G. (1972). An auditory analogue of the Sperling partial report procedure: Evidence for brief auditory storage. Cognitive Psychology, 3, 255-267.
 13. Diehl, R.L., Elman, J.L., and McCusker, S.B. (1978). Contrast effects in stop consonant identification. Journal of Experimental Psychology: Human Perception and Performance, 4, 599-609.
 14. Durlach, N.I. and Braida, L.D. (1969). Intensity perception. I. Preliminary theory of intensity resolution. Journal of the Acoustical Society of America, 46, 322-383.
 15. Eimas, P.D. and Corbit, J.D. (1973). Selective adaptation of linguistic feature detectors. Cognitive Psychology, 4, 99-109.
 16. Elman, J.L. (1979). Perceptual origins of the phoneme boundary effect and selective adaptation to speech: A signal detection theory analysis. Journal of the Acoustical Society of America, 65, 190-207.
 17. Fujisaki, H. and Kawashima, T. (1969). On the modes and mechanisms of speech perception. Annual report of the Engineering Research Institute, Faculty of Engineering, University of Tokyo, 28, 67-73.
 18. Fujisaki, H. and Kawashima, T. (1970). Some experiments on speech perception and a model for the perceptual mechanism. Annual report of the Engineering Research Institute, Faculty of Engineering, Univiversity of Tokyo, 29, 207-214.
 19. Goldberg, R.F. (1986). Perceptual anchors in vowel and consonant continua. M.S.E.E. thesis, M.I.T.
 20. Goldberg, R.F., Macmillan, N.A., and Braida, L.D. (1985). A perceptual-anchor interpretation of categorical phenomena on a vowel continuum. Journal of the Acoustical Society of America, 77, Suppl. 1, S7. (Abstract).
 21. Green, D.M. and Swets, J.A. (1974). Signal detection theory and psychophysics. Huntington, N.Y.: Krieger.
 22. Khazatsky, V. (1985). Anchor model in identification with a standard. Unpublished manuscript.
 23. Macmillan, N.A. (1979). Categorical perception of musical sounds: The psychophysics of plucks and bows. Bulletin of the Psychonomic Society, 11, 241 (Abstract).
 24. Macmillan, N.A. (in press). Beyond the categorical/continuous distinction: A psychophysical approach to processing modes. In: S. Harnad (Ed.), Categorical Perception. New York: Cambridge University Press.
 25. Macmillan, N.A. (in preparation). A psychophysical study of tone onset time resolution.
 26. Macmillan, N.A. and Braida, L.D. (1985). Toward a psychophysics of the speech mode. Bulletin of the Psychonomic Society, 23, 278 (Abstract).
 27. Macmillan, N.A., Kaplan, H.L., and Creelman, C.D. (1977). The psychophysics of categorical perception. Psychological Review, 84, 452-471.
 28. McGill, W.M. and Goldberg, J.P. (1968). A study of the near-miss involving Weber's law and pure-tone intensity discrimination. Perception & Psychophysics, 4, 105-109.
 29. Miller, J.D., Wier, C.C., Pastore, R.E., Kelly, W.J., and Dooling, R.J. (1976). Discrimination and labeling of noise-buzz sequences

- with varying noise-lead times: An example of categorical perception. Journal of the Acoustical Society of America, 60, 410-417.
30. Parducci, A. (1965). Category judgment: A range-frequency model. Psychological Review, 72, 407-418.
31. Pastore, R.E. (in press). Categorical perception: Some psychophysical models. In: S. Harnad (Ed.), Categorical Perception. New York: Cambridge University Press.
32. Pisoni, D.B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. Perception & Psychophysics, 13, 253-260.
33. Pisoni, D.B. (1975). Auditory short-term memory and vowel perception. Memory & Cognition, 3, 7-18.
34. Pisoni, D.B. (1977). Identification and discrimination of the relative onset time of two component tones: Implications for voicing perception in stops. Journal of the Acoustical Society of America, 61, 1352-1361.
35. Pynn, C.T., Braida, L.D., and Durlach, N.I. (1972). Intensity perception. III. Resolution in small range identification experiments. Journal of the Acoustical Society of America, 51, 559-566.
36. Rabinowitz, W.M., Lim, J.S., Braida, L.D., and Durlach, N.I. (1976). Intensity perception. VI. Summary of deviations from Weber's law for 1 kHz tones. Journal of the Acoustical Society of America, 59, 1506-1509.
37. Repp, B.H., Healy, A.F., and Crowder, R.G. (1979). Categories and context in the perception of isolated steady-state vowels. Journal of Experimental Psychology: Human Perception and Performance, 5, 129-145.
38. Rosen, S.M. (1979). Range and frequency effects in consonant categorization. Journal of Phonetics, 7, 393-402.
39. Rosner, B.S. (1984). Perception of voice-onset time: A signal-detection analysis. Journal of the Acoustical Society of America, 75, 1231-1242.
40. Sachs, R.M. and Grant, K.W. (1976). Stimulus correlates in the perception of voice onset time (VOT): II. Discrimination of speech with high and low stimulus uncertainty. Journal of the Acoustical Society of America, 60, S91 (Abstract).
41. Sawusch, J.R., Nusbaum, H.C., and Schwab, E.C. (1980). Contextual effects in vowel perception. II: Evidence for two processing mechanisms. Perception & Psychophysics, 27, 421-434.
42. Stephens, L.M. (1986). A study of the interpretation of sounds in a synthetic /ba/-/da/-/ga/ continuum. B.S.E.E. Thesis, M.I.T.
43. Watson, C.S., Kewley-Port, D. and Foyle, D.C. (1985). Temporal acuity for speech and nonspeech sounds: The role of stimulus uncertainty. Journal of the Acoustical Society of America, 77, Suppl. 1. S27. (Abstract).

Appendix: Expressions for sensitivity

Sensitivity d' in various paradigms depends on the sensory variance β^2 , range-dependent context variance $(GR)^2$, and time-dependent trace variance AT . The difference between the two distribution means is denoted a ; A and G are constants.

Fixed discrimination:

$$d'_{fixed} = \frac{a}{\beta}. \quad (\text{A1})$$

Identification:

$$d'_{ident} = \frac{a}{[\beta^2 + (GR)^2]^{1/2}}. \quad (\text{A2})$$

Identification and fixed discrimination data can be used to calculate the relative context variance:

$$\frac{(GR)^2}{\beta^2} = \left\{ \frac{d'_{fixed}}{d'_{ident}} \right\}^2 - 1 \quad (\text{A3})$$

Roving discrimination, trace mode only:

$$d'_{roving} = \frac{a}{[\beta^2 + AT]^{1/2}}. \quad (\text{A4})$$

Roving discrimination, trace and context modes:

$$d'_{roving} = \frac{a}{\left\{ \beta^2 + \frac{1}{\frac{1}{(GR)^2} + \frac{1}{AT}} \right\}^{1/2}}. \quad (\text{A5})$$

PSYCHOPHYSICS VERSUS SPECIALIZED PROCESSES IN SPEECH
PERCEPTION:
AN ALTERNATIVE PERSPECTIVE*

Dominic W. Massaro

Program in Experimental Psychology, University of California,
Santa Cruz , Santa Cruz, CA 95064, USA

The title of this conference describes one of the two major contrasting frameworks for speech perception research during the last three decades. This point of view is that speech perception can be understood by the principles of auditory psychophysics. Speech involves complex auditory signals and the processing and perception of speech can be understood by the rules of processing complex auditory signals. Research representative of the paradigm has been contributed by Cutting and Rosner (1974), Kuhl and Miller (1975, 1978), Pastore, Ahroon, Baffuto, Friedman, Puleo, and Fink (1977), and Pisoni (1977). The other point of view, the antithesis of the first, is that speech perception represents the operation of a set of specialized processes unique to speech. This view began as the motor theory of speech perception (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967) and has evolved into an illustration of the modularity principle (Fodor, 1983; Liberman & Mattingly, 1985). Representative studies within this paradigm can be found in Best, Morrongiello, and Robson (1981), Eimas and Corbit (1973), Eimas and Miller (1980), and Repp (1982).

These two schools of thought have imposed a very narrow research environment, falling short of contributing to a basic understanding of speech perception. The empirical work and controversy has bounced back and forth resembling a tennis match in which the server demonstrates that speech perception is special and the opponent replies that it is not. Caught up in the controversy of psychophysics versus specialization, little has been accomplished with respect to the question of how speech is perceived. A resolution of the controversy is offered by a third perspective proven successful in other domains such as reading and categorization of natural objects (Massaro, 1984; Massaro, in press, d).

INFORMATION-INTEGRATION PERSPECTIVE

One salient aspect of pattern recognition involves the processing of multiple sources of information. Consider recognition of the word "performance" in the spoken sentence: "The actress was

*The writing of this paper and the research reported in the paper were supported, in part, by NINCDS Grant 20314 from the Public Health Service and Grant BNS-83-15192 from the National Science Foundation. Michael M. Cohen made important contributions to the research enterprises, Neil Appel helped with the references, and Ervin R. Hafter provided helpful feedback on an earlier version of this paper.

praised for her outstanding performance". Recognition of the critical word is achieved via a variety of bottom-up and top-down sources of information. Top-down sources include semantic and syntactic constraints and bottom-up sources include acoustic features and syllables making up the word. Phonological constraints also have been shown to contribute to perceptual recognition at the word level (Massaro & Cohen, 1983d). Integrating multiple sources of information appears to be central to pattern recognition, not just speech perception.

Historically, the present approach can be traced, in part, to Egon Brunswik's (1952, 1956) Probabilistic Functionalism. He proposed that there are many cues influencing perception but that these cues are equivocal and only probabilistically related to the objects of interest. Brunswik realized "the limited ecological validity or trustworthiness of cues ... To improve its (the organism's) bet, it must accumulate and combine cues" (Brunswik, 1955, pg. 207). Methodologically, Brunswik called for representative designs or experiments that are random samples from natural phenomena. We reject this method, however, in favor of experiments that manipulate the environment. Only by independently varying naturally correlated cues are we able to determine which cues are functional in perception. To this end, we employ factorial designs and functional-measurement techniques (Anderson, 1981, 1982) and test mathematical models of perceptual performance (Massaro & Cohen, 1983c).

According to the present framework, well-learned patterns are recognized in accordance with a general algorithm regardless of the modality or particular nature of the patterns (Massaro, 1979; Oden & Massaro, 1978). The model postulates three operations in perceptual (primary) recognition: feature evaluation, prototype matching, and pattern classification. Continuously-valued features are evaluated, integrated, and matched against prototype descriptions in memory, and an identification decision is made on the basis of the relative goodness of match of the stimulus information with the relevant prototype descriptions. The model is called a fuzzy logical model of perception (abbreviated FLMP).

It is necessary to distinguish between environmental properties that are potentially informative about some object or event and the properties actually used in perception and recognition of the object or event (Massaro, in press c). The former might be called data and the latter information. With respect to bottom-up properties in speech, I have referred to them as acoustic characteristics and acoustic features, respectively (Massaro, 1975). One primary task of research is to determine which acoustic characteristics function as acoustic features in speech perception. Furthermore, it is necessary to specify the algorithms or computations that resolve the wide variety of acoustic features. The description does not end here, however, because the features, once resolved, must be combined or integrated to achieve a unitary perception. The processing of top-down sources of information must be described in the same manner as bottom-up sources. An adequate theory must also specify how the perceiver integrates bottom-up and top-down sources in real time during speech perception. Finally, decision processes must be revealed given that speech perception, like other forms of pattern recognition, represents a selection of one of several viable candidates or alternatives. I will

first discuss briefly the research areas of categorical perception, normalization, duplex perception, the McGurk effect, and trading relations, and contrast the two modal approaches to our approach to the study of speech perception. This latter approach not only offers a productive framework for describing the processes involved in speech perception, it provides major constraints on potential contenders for a theory of speech perception.

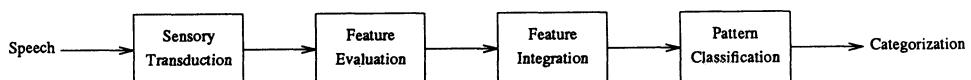


FIGURE 1, Schematic representation of Stages of Processing in Speech Perception.

Figure 1 gives a schematic representation of four stages of processing in categorization of speech. Sensory transduction transforms the physical stimulus into sensory data. The feature evaluation process assesses the sensory data with respect to the important dimensions of speech characterizing the speech segments in the language. The feature integration process integrates or combines the featural information from the different dimensions with respect to prototype representations in memory. The pattern classification process determines the relative goodness of match of the sensory information with the relevant prototypes in memory. The relative goodness of match is equal to the likelihood of identifying the speech event as an instance of the prototype category.

CATEGORICAL PERCEPTION

The speech-is-special school offered the phenomenon of categorical perception in its support (Liberman, et al., 1967). The contemporary field's repression of the concept of categorical perception makes transparent the sterility of this area of research. To this day I cannot understand why categorization behavior was (and continues to be) interpreted as evidence for categorical perception. At the risk of belaboring the obvious, I will illustrate very quickly how it is only natural that continuous perception should lead to sharp category boundaries along a stimulus continuum. Given a stimulus continuum from A to not-A that is perceived continuously, the goodness of A, abbreviated G(A), is an index of the degree to which the information represents the category A. The left panel of Figure 2 shows G(A) as a linear function of Variable A.

An optimal decision rule in a discrete judgment task would set the criterion value at .5 and classify the pattern as A for any value greater than this value. Otherwise, the pattern is classified as not-A. Given this decision rule, the probability of an A response, P(A) would take the step-function form shown in the right panel of Figure 2. That is, with a fixed criterion value and no variability, the decision operation changes the continuous linear function given by the perceptual operation into a step function. Although based on continuous perception, this function is identical to the idealized form

of categorical perception in a speech identification task (Studdert-Kennedy, Liberman, Harris, & Cooper 1970). It follows that a step function for identification is not evidence for categorical perception because it can occur given continuous information.

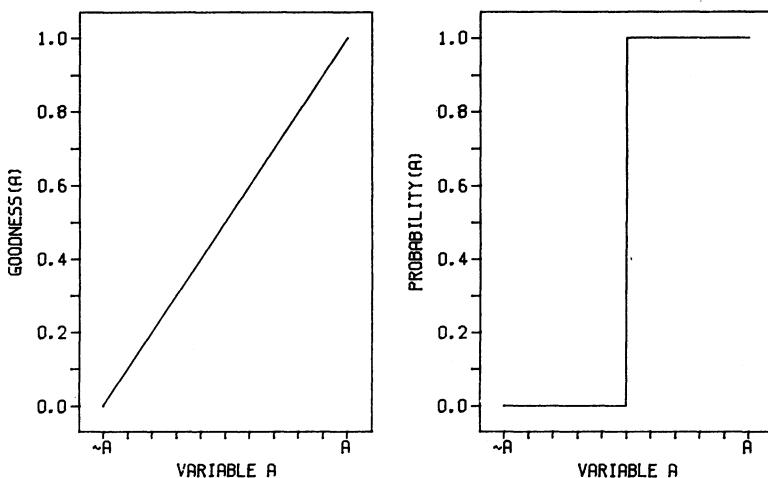


FIGURE 2, Left Panel. The degree to which a stimulus represents the category A, called Goodness(A) as a function of the level along a stimulus continuum between not-A and A. Right Panel. The probability of an A response, Probability(A), as a function of the stimulus continuum if the subject maintains a decision criterion at a particular value of Goodness(A) and responds A if and only if Goodness(A) exceeds the decision criterion.

If there is noise in the mapping from stimulus to identification, a given level of Variable A cannot be expected to produce the same identification judgment on each presentation. It is reasonable to assume that a given level of Variable A produces a normally distributed range of Goodness(A) values with a mean directly related to the level of Variable A and a variance equal across all levels of Variable A. If this is the case, noise will influence the identification judgment for the levels of Variable A near the criterion value more than the levels away from the criterion value. Figure 3 illustrates the expected outcome for identification if there is normally distributed noise with the same criterion value assumed in Figure 2.

If the noise is normal and has the same mean and variance across the continuum, a stimulus whose mean goodness, $G(A)$, is at the criterion value will produce random classifications. The value of $G(A)$ will be above the criterion on half of the trials and below the criterion on the other half. As the value of $G(A)$ moves away from the criterion value, the noise will have a diminishing effect on the identification judgments. Noise has a larger influence on identification in the middle of the range of $G(A)$ values than at the extremes since variability goes in both directions in the middle and only inward at the extremes.

This example shows that categorical decisions made on the basis of continuous information produce identification functions with sharp boundaries, previously taken to represent categorical perception. Strictly speaking, of course, categorical perception was considered present only if discrimination behavior did not exceed that predicted from categorization (Studdert-Kennedy, et al., 1970). However, one should not have been impressed with the failure of discrimination to exceed that predicted by categorization if the discrimination task resembled something more akin to categorization than discrimination (Fujisaki & Kawashima, 1970, Paap, 1975). Even in this period of enlightenment, however, we have authors such as Eimas (1985) using the concept of categorical perception to describe typical categorization behavior.

This analysis of categorical perception also makes explicit at what level in the processing system categorical perception must be demonstrated. Categorization behavior alone cannot be taken as evidence for categorical perception, for it is the mapping of stimulus information to sensory information (feature evaluation in our model) that is relevant, not simply that mapping of stimulus information to perceptual judgment. The issue formalized in Figure 1 is whether the mapping of Variable A to Goodness(A) is continuous or categorical. I don't understand Repp's (this volume) opinion that the present analysis of the problem precludes any contribution of experience and attention. In fact, the McClelland and Elman Trace Model places categorical perception at exactly the level of mapping of Variable A to Goodness(A) (see later discussion). If categorization implies categorical perception, we have abandoned any interest in the processes leading to perception and have joined the behavioristic camp of psychological inquiry.

The psychoacoustician's answer to categorical perception was not to reject the concept but to attempt to show that it also occurs for nonspeech stimuli. Thus speech cannot be considered special because both speech and nonspeech are perceived categorically. An attractive but incorrect solution was the idea of natural auditory sensitivities accounting for perceptual categories in speech. Little processing is needed if the speech categories fall on opposite sides of some perceptual discontinuity in the auditory system. For example, the most popular distinction is voice onset time (VOT), the time interval between the onset of the release of a stop consonant and the

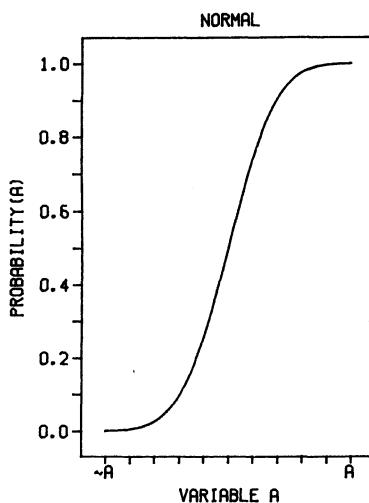


FIGURE 3, The probability(A) as a function of Variable A given the linear relationship between Goodness(A) and Variable A and the decision criterion represented in Figure 1, but with normally distributed noise added to the mapping of Variable A to Goodness(A).

onset of vocal cord vibration (see Rosen & Howell, this volume). The psychoacoustic explanation for VOT has been in terms of temporal order judgments (TOJ). In 1959, Hirsh found that listeners needed about a 17 ms onset difference to determine which of two auditory events (e.g., tones) occurred first. If the 17 ms difference is viewed as a threshold, then stimuli with onset asynchronies longer than this value would be heard in one way and stimuli with onset asynchronies shorter than this value would be heard in another way. Voiced stop consonants would have VOT values less than the threshold, whereas voiceless stops would have VOT values greater than the threshold.

There are several limitations with this proposal that make it an unlikely possibility for perception of voicing. First, listeners do not experience two auditory events one before the other, but instead qualitatively different types of percepts. (Massaro, 1972; Warren, 1974a, 1974b). Second, the onset difference at threshold depends on other factors such as the overall duration of the auditory events (Patterson & Green, 1970). Third, there is little evidence that the critical cue to voicing is temporal order as opposed to the acoustic events during the onset of the speech segment. In a neglected study, Winitz, La Riviere, and Herriman (1975) opposed VOT and the quality of the burst and aspiration in stop consonants in initial position. As an example, the burst of /du/ was isolated and separated from the periodic portion by an interval appropriate for the VOT of /tu/. In this situation, listeners tended to hear /du/ showing that VOT itself was not the critical cue to voicing. Fourth, Rosen and Howell (this volume) illustrate that Hirsh's (1959) results are inconsistent with a threshold or discontinuity in the perception of relative onset time. Other limitations to a TOJ threshold in speech exist and are mentioned in other sections of this commentary.

At best, there may be only a couple of speech contrasts even approaching categorical speech perception (Howell & Rosen, 1983). These few cases are better described as slight irregularities or discontinuities rather than by sharp thresholds in the auditory processing system (Hary & Massaro, 1982; Massaro & Oden, 1980b). By irregularity is meant that the mapping between some perceptual variable such as the discrimination of which stimulus came first and some physical variable such as relative onset time is irregular. Speech, like nonspeech, is unlikely to be perceived categorically, even in those few rare cases of irregularity across some continuum.

Some psychoacousticians had the inappropriate reaction to the notion of categorical perception. Macmillan, Kaplan, and Creelman (1977) redefined categorical perception within the theory of signal detection which, of course, assumes continuous information. Although categorical perception was defined as a match between categorization and discrimination performance, the underlying assumptions were those of continuous perception. Thus, a match between discrimination performance and categorization performance certainly could not mean categorical perception. More generally, evidence consistent with the traditional definition of categorical perception is not necessarily inconsistent with alternative models assuming continuous perception (Massaro, in press b). Using the traditional paradigm of the categorization/discrimination task, Hary and Massaro (1982) demonstrated that sounds that appear to be perceived categorically in one context appear to be perceived continuously in another. Massaro

and Cohen (1983a) showed that the distribution of rating responses to speech syllables was significantly better described by a continuous than by a categorical model of perception. None of us like to be reminded of the sterility of a particular research enterprise. If we can face up to mistakes in the past, however, we may be less susceptible to repeating them in the future.

The idea of categorical perception is not feasible for the perception of continuous speech. It is well known that the acoustic characteristics of the speech signal vary considerably with speaker, rate of speaking, and the surrounding segmental contextual. Categorical perception would prove to be too inflexible to deal with these context variations. As an example, Summerfield (1982) showed that the influence of voice-onset time (VOT) in the discrimination of voicing could not be categorically perceived. The contextual modifications resulting from the influence of this variable in speech are significantly different from what could be predicted from nonspeech. For example, the position of boundaries between phonetic categories on VOT continua depends on other spectral properties, such as the onset frequency of the first formant. Analogous spectral manipulations in the nonspeech analogs of the syllables do not result in a similar dependency.

It is also of considerable interest that the demise of categorical perception poses a serious problem for the Trace Model of speech perception (McClelland & Elman, 1986). Their model produces categorical-like behavior at the sensory (featural) level rather than at simply the decision stage. This occurs because of the nonindependence between the feature and phoneme levels in the model (which contrasts with the independence assumption of our fuzzy logical model). In the Trace Model, a stimulus pattern is presented and activation of the corresponding features sends more excitation to some phoneme units than others. Given the assumption of feedback from the phoneme to the feature level, the activation of a particular phoneme feeds down and activates the features corresponding to that phoneme (McClelland & Elman, 1986, p. 47). This effect of feedback produces enhanced sensitivity around a category boundary, exactly as predicted by categorical perception. Categorical perception is infrequent in speech if it occurs at all, and we have exposed an important weakness in a powerful and comprehensive model of speech perception.

NORMALIZATION

A second area of research has to do with normalization processes in speech perception. It is well known, for example, that a given speech segment in a sentence will be interpreted differently depending on the rate of speaking. Thus, for example, Miller and Liberman (1979) showed that the identification of /ba/ versus /wa/ as a function of transition duration also depended on the rate of speaking the sentence. Pisoni, Carrell, and Gans (1983) showed that similar context effects occur with nonspeech. However, the size of the normalization effect appeared to differ in the speech and nonspeech tasks. In one case, the context effects appeared to be significantly larger for the speech stimuli, and in another the context effects appeared to be significantly larger for the nonspeech stimuli. It is incumbent upon the psychoacoustician to show direct correspondences between the nonspeech and speech signals, not simply a rough

approximation. On the other hand, I am not optimistic about this research strategy given the limitations in comparing speech and nonspeech (see discussion in the section Trading Relations). Sine wave analogs of speech heard as speech or nonspeech might provide a better assessment of a psychoacoustic explanation of normalization (see Best et al., 1981).

DUPLEX PERCEPTION

Another arena of research controversy involves duplex perception. In this situation a single stimulus gives rise to two different perceptions: one is speech, and the other is nonspeech. If an isolated formant transition (the chirp) is presented to one ear, while the rest of the speech sound (the base) is presented to the other ear, subjects report hearing both speech and nonspeech sounds lateralized at different locations (Nusbaum, Schwab, & Sawusch, 1983; Rand, 1974). The speech percept must result from the fusion of the two inputs, whereas the nonspeech percept must result from the isolated formant transitions. In addition, different rules seem to describe the perceptual processes involved with the two percepts. The formant transition is perceived differently in the context of being fused with the rest of the speech sound relative to its perception as an isolated, nonspeech sound. In the speech-is-special camp, this result is interpreted as evidence for a specialized process in speech perception (Liberman, 1982; Repp, Milburn, & Ashkenas, 1983). In turn, the psychoacousticians take pains to illustrate that similar processes can occur with nonspeech. As an example, Pastore, Schmuckler, Rosenblum, and Szczesniak (1983) showed that duplex perception also occurs for musical stimuli, which then weakens the argument for specialized processes. The debate continues in this area; for recent papers see Repp (1984) and Nusbaum (1984). Nusbaum (1984) provides a reasonable interpretation of duplex perception. Following the idea of integrating multiple sources of information in speech perception, it is reasonable that both the base and the chirp contribute to perceptual recognition of the speech segment. This interpretation, if formalized within the fuzzy logical model of perception, also accounts for the finding that the contribution of the base should increase as the relevant cue given by the chirp becomes ambiguous.

McGURK EFFECT

An area of research that has captured much of my effort is speech perception by ear and by eye (McGurk & MacDonald, 1976). In this situation, watching a speaker articulate speech influences what the perceiver hears. This result obviously has no psychoacoustic explanation, and this provided much hope for the framework of speech as special. As summarized by one esteemed researcher, "Both motor (speech is special) theorists and Gibsonians went dancing through the streets of every major city in the Eastern U.S. the day in 1976 that McGurk and MacDonald's paper in Nature hit the newsstands." Once again, the problem seems to be that the speech-is-special camp did not consider that alternative explanations are also consistent with the McGurk effect (Massaro & Cohen, 1983b; Massaro, in press e). We might expect that bimodal perception resulting from sight and sound can occur in other situations such as watching and listening to a musician pluck or bow a string on a violin. The visual capture effect in which the sight of an object can attract localization of a sound

source might be used as the counterexample against a speech-is-special interpretation of the McGurk effect.

If not a psychoacoustic explanation of the McGurk effect, a more general physical explanation is still a remote possibility. Perhaps there is some inherent relationship between the mouth configurations and the sound configurations, independent of their representation of vocalization. Kuhl and Meltzoff (1982, 1984) found that five-month-old infants recognized cross-modal correspondences of the vowels /i/ and /a/. The infants viewed a film showing two side-by-side images of a talker articulating /i/ and the same talker articulating /a/, in synchrony, with one of the two vowel sounds. The infants looked longer at the face matching the sound than at the nonmatching face. To test the physical explanation, Kuhl and Meltzoff (1984) used pure tone analogs of /a/ and /i/. There was a complete reversal of the finding with speech; infants now looked longer at the articulation of /i/ when the pure tone analog of /a/ was played, and analogously for the pure tone analog of /i/ (Kuhl & Meltzoff, 1984). Subject to the limitation of a speech-nonspeech comparison, the experiment offers little hope for the adequacy of physics (analogous to the inadequacy of pure psychoacoustics) as an explanation of bimodal speech perception.

TRADING RELATIONS

The final area of research has to do with what are called trading relations in the perceptual categorization of speech. Multiple cues influence a perceptual discrimination, and these cues can be traded off for one another. As an example, voice-onset time can be traded off against the first formant frequency and transition at the onset of voicing in the identification of the voicing of initial stops and fricatives (Massaro & Cohen, 1976, 1977; Summerfield & Haggard, 1974). Trading relations have been interpreted by the speech-is-special camp as meaning that articulatory processes must intervene in the integration of these diverse cues. Once again, trading relations can be found in other domains in addition to speech. Oden (1981), for example, has shown that visual properties about cups and bowls are evaluated and integrated as predicted by the fuzzy logical model, the same model with a history of success in the domain of speech perception.

The term "trading relations" is incomplete and possibly misleading to describe the contribution of multiple sources of information in speech perception. It might seem reasonable to say that voice onset time trades off with the frequency of the first formant at the onset of voicing. But it seems unreasonable to use trading relations to describe the contributions of lexical information and voice onset time to the perception of voicing (Ganong, 1980; Massaro & Oden, 1980b). In the latter case, it is more obvious that there are multiple sources of information contributing to the perceptual interpretation of the message. The same is true in the former case, and it is necessary for a theory to describe how the sources of information are evaluated and integrated to give the tradeoff that is observed.

The historical study of this problem in speech perception may also be of interest. Early workers at Haskins Laboratories manipulated multiple cues to perceptual categorization but did not assess how the

cues were integrated (Hoffman, 1958). Stevens and Klatt (1974) varied both voice-onset time and the onset of the first formant but did not manipulate these in a complete factorial design (see also Sawusch & Pisoni, 1974). Massaro and Cohen (1976) manipulated two cues to voicing in a factorial design and tested mathematical models of their integration. Sometime afterwards, workers at Haskins Laboratories began using factorial designs and studying trading relations and arguing that these depict a special speech processer (Repp 1977, 1982).

Analogous to the nonspeech studies of categorical perception, we are now witnessing a flurry of experiments purportedly illustrating trading relations with nonspeech (Diehl, this volume). Several do not make direct comparisons between speech and nonspeech, precluding a direct comparison between the two. Unless a direct comparison is provided, we have no measure of whether the trading relations are the same in speech and nonspeech domains. For those studies involving a direct comparison, some kind of model is necessary to evaluate the similarities and differences between speech and nonspeech situations. Our fuzzy logical model of perception (Massaro, in press a) permits direct assessments of the information value of each property involved in the trading relation and the nature of the integration process generating the trading relation. Both of these questions are fundamental to assessing any psychoacoustical bases for trading relations.

Diehl (this volume) assessed the psychoacoustic basis for the tradeoff of vowel duration and closure duration of consonants in the perception of voicing of medial stop consonants. Square wave analogs were created by replacing the formants of the speech with square waves. Subjects judged the speech syllables as the voiced or voiceless alternatives and judged the nonspeech as having a short or long silent period (closure) in the middle of the sound. Figure 4 gives the average results for the speech and nonspeech continua. Although superficially the results appear to be comparable, a fine-grained analysis reveals large differences between the speech and nonspeech effects. This difference can be highlighted by fitting the results with two very different models: the FLMP and a weighted averaging model. These models make different predictions about the joint effect of two cues. The FLMP predicts that the contribution of one cue increases with increases in the ambiguity of the other cue, leading to a statistical interaction, as given in the top panel in Figure 4. The weighted averaging model, on the other hand, predicts additive effects and thus parallel curves similar to those shown in the bottom panel of Figure 4.

Quantitative descriptions of the results reinforce this graphical analysis. The FLMP gave a better description of the speech results than did the weighted averaging model, and the opposite outcome emerged for the nonspeech results. A fine-grained analysis appears to reveal important differences between speech and nonspeech analogs. Integrating multiple sources of information appears to be a psychological rather than a psychoacoustic phenomenon.

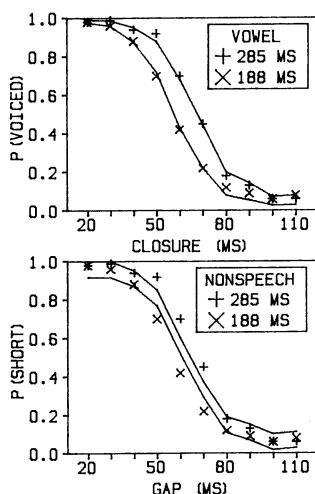


FIGURE 4, Top panel. Proportion of voicing judgements as a function of closure duration and vowel duration. The lines give the predictions of the fuzzy logical model of perception. Bottom panel. Proportion of short judgements as a function of silent gap and the duration of the nonspeech preceding the gap for square-wave analogs of the speech syllables. The lines give the predictions for a weighted averaging of the two properties of the nonspeech (results from Diehl et al., this volume).

The surprising finding of nearly equivalent voice-onset-boundaries as a function of place of articulation for chinchillas and humans could have been a result of the synthetic speech stimuli that were used. The stimuli of Lisker and Abramson (1970) allowed other properties of the stimulus to vary with changes in the speech variable of interest. In these stimuli, the duration of the first formant transition covaried with differences in the higher formants that were varied to cue place of articulation. This property could have been responsible for the differences in the VOT boundary as a function of place of articulation. Later versions of synthesized speech without this property do not replicate the large differences that were originally found (Miller, 1977, Massaro & Oden, 1980a, Oden & Massaro 1978). Hence, the original trading relation that was observed appears to have a psychoacoustic rather than a psychological explanation.

Additivity also appears for speech when the two properties function psychoacoustically as illustrated in van Heuven's (this volume) study. Vowel decay time and friction-rise time should have comparable acoustic effects in a vowel-consonant sequence. Both properties contribute to perceived silence between the vowel and consonant. The silent interval cues the difference between the affricate chop and the fricative shop, with longer intervals cueing the affricate. In sequences of say chop/shop, the effects of vowel decay time and friction rise time appeared to have additive effects on the categorization of the affricative-fricative. Increasing the decay time of the vowel and increasing the rise time of the consonant both increased the likelihood of an affricate categorization, supposedly by increasing the perceived silence between the vowel and consonant.

The psychoacoustical basis of trading relations has been assessed using nonhuman subjects, again repeating a strategy used in the categorical-perception controversy (Kuhl & Miller, 1975, 1978; Kuhl & Padden, 1982, 1983; Waters & Wilson, 1976).

Rejecting the psychoacoustical basis for trading relations by showing differences between speech and nonspeech is perhaps a losing battle. The reason, analogous to studies of categorical perception, is the persistent possibility that an inappropriate nonspeech stimulus was used. Workers at Haskins Laboratories have solved this problem to some extent by using sine-wave analogs of speech and testing subjects in both speech and nonspeech modes of perceiving these signals (Best, et al., 1981). They have succeeded in demonstrating large differences between the two modes of perceiving, but without a formal analysis we don't know enough about how the two modes differ. The Best, et al. (1981) results appear similar to those of Diehl in that nonspeech integration appears to be more additive relative to the multiplicative integration of speech. When the signals are heard as nonspeech, subjects might use only one of the two varying properties to categorize the sounds. When they use two properties, they appear to integrate the two nonspeech cues according to a weighted-averaging rule. When the signals are heard as speech, however, the two cues are integrated in such a way that the least ambiguous cue has the greatest impact on the decision.

Diehl (this volume) claims that the integration of closure duration and closure pulsing has an auditory basis, perhaps the glottal pulsing biasing the observer to hear the interval as short. However, the actual relationship between the amplitude of voicing during the closure and the closure interval for voiced stops differs for different languages (Lindau & Ladefoged, 1986). It seems that these two variables are independently variable in articulation and different languages combine them differently to convey voicing information. This arbitrariness between the two dimensions precludes any psychoacoustic explanation of the integration of the two cues. In our framework, the integration of the two cues will reflect how these cues are used by the speaker to inform the listener.

Psychophysical explanations should be able to illuminate the contrasts between short and long vowels. Two vowels that have highly similar formant frequencies tend to differ from one another in duration. If psychoacoustics has any relevance, the relationship between formant structure and duration should be systematic. The normally longer vowel should be heard as longer relative to the normally shorter vowel, even though presented at the same duration. For example, a vowel with a high F2 might be heard as longer than a vowel with a low F2. In this case, duration might acquire cue status, as suggested by Stevens, Keyser, and Kawasaki (1986). Consider the three pairs /i/-/ɪ/, /u/-/ʊ/, and /ʌ/-/e/ and their first two formants and average durations shown in Table 1.

For two of the pairs, F1 is lower in frequency for the longer member and higher for the longer vowel of the third pair. The frequency of F2 is higher for the longer member of two of the pairs and lower for the longer member of the third pair. Thus there is no systematic relationship between formant frequency and relative duration. Clearly it is unlikely that the integration of these two characteristics results from some low-level auditory interaction.

Table 1. Average values of first and second formants and the durations of six vowels (from Peterson & Barney, 1952, and Peterson & Lehiste, 1960)

	/i/	-	/ɪ/		/u/	-	/U/		/ʌ/	-	/ɛ/
F ₁ (Hz)	270		390		300		440		660		530
F ₂ (Hz)	2290		1990		870		1020		1720		1840
Duration (ms)	240		180		260		200		330		200

Having observed limitations with nonspeech and nonhuman subjects, I offer in their place developmental and cross-linguistic comparisons with respect to trading relations. In an early observation, Simon and Fourcin (1978) found differences in the contribution of F1 to perceived voicing of initial stop consonants as a function of development and language. Onset frequency of F1 is more informative for voicing in English relative to French and interestingly French children acquire this cue sometime later than their English cohorts. If this result is reliable, it goes well beyond what could be predicted by a psychoacoustic interpretation. If this psychoacoustic interpretation is broadened to include an important contribution of perceptual learning, however, it can postdict the results. While achieving this worthwhile goal it would also become much less like its parsimonious predecessor and more like the speech-is-special (Liberman 1982) and information-integration (Massaro & Oden, 1980b) viewpoints.

Cross-linguistic comparisons offer a direct assessment of the psychoacoustic basis of trading relations. If a trading relation between two properties of the speech signal exists for subjects of one language but not for subjects of another, there is little merit to a psychoacoustic explanation. The results would be compatible with the integration of the two properties in one case and not in the other. This outcome would supposedly occur when a given property is ecologically valid in one language but not in the other.

A recent cross-linguistic study provides a simple rejection of the psychoacoustical-basis hypothesis. The contrast of interest was postvocalic voicing as cued by the duration of the preceding vowel and the duration of the consonant (Denes, 1955). Like the two pronunciations of use, the words peas and peace differ in the voicing of the final consonant. Since Denes (1955), we have known that both vowel duration and the aperiodic fricative duration contribute to this distinction in English (Derr & Massaro, 1980; Massaro & Cohen, 1976, 1977; Raphael, 1972). Flege and Hillenbrand (1986) observed that the /s/-/z/ distinction is not learned in Swedish and Finnish, since neither language possesses a /z/ phoneme. If a psychoacoustic basis existed for the trading relation between vowel and consonant durations in English, then learning this new distinction should be an easy charge for Swedish and Finnish speakers acquiring English. However, this distinction is difficult to learn and the question is whether the cues used by English listeners are learned by these individuals acquiring English as a second language. Both experienced and inexperienced

listeners were tested on the peas-peace contrast. Five durations of the periodic vowel were factorially combined with five durations of the aperiodic frication and subjects were instructed in English to categorize the words as peas or peace.

The results revealed that the Swedish and Finnish listeners did not use fricative duration as a cue to the /s/-/z/ contrast in English, but based their identification decision on only vowel duration. Although this result generates a variety of interesting questions, for our purposes, it weakens a simple psychoacoustic explanation of the trading relation in English. There is nothing inherent in the auditory resolution of the English syllables that leads to the tradeoff between vowel duration and consonant duration. If there were, the identification results for the Swedish and Finnish subjects should have been identical to those for the English speakers. The trading relation exists because the English listener integrates these two cues in perceptual recognition, not because the auditory system naturally categorizes short vowel-long consonant syllables as one class and long vowel-short consonant syllables as another class. The evaluation of vowel duration and consonant duration appear to occur relatively independently of one another, as assumed in our fuzzy logical model (Derr & Massaro, 1980; Massaro & Cohen, 1977, 1983b). They are integrated by native English speakers because both are informative about the identity of voicing of the final consonant.

The observation that Swedish and Finnish speakers are not influenced by consonant duration also weakens the appeal of consonant/vowel (C/V) ratio as the cue to voicing (Port & Dalby, 1982). If C/V ratio were used, consonant duration would necessarily have an influence on perceptual categorization. Conceptualizing vowel duration and consonant duration as independent sources of information about voicing, however, describes the results parsimoniously (Massaro & Cohen, 1983b). Consonant duration does not acquire cue value for the Swedish and Finnish speakers because their language does not have a /z/ phoneme. Learning English as a second language does not seem to change this situation, possibly because these individuals continue to speak their native language. In this case, the cue value of various sources of information is not easily normalized to take into account the language currently being perceived.

Remaining questions are why the Swedish and Finnish listeners used vowel duration as a cue and whether they use consonant duration as a cue for other contrasts such as stop consonants in medial position. Vowel duration would be functional for stops in final position (Raphael, 1972) and this could have generalized to the new /z/-/s/ contrast in English.

We have parallel results comparing Chinese and English subjects on the perception of a vowel contrast [i]-[y] that exists in Chinese but not English (Massaro, Tseng, & Cohen, 1983). The English subjects had no experience with Chinese. The vowels differ not only in their formant pattern but also in loudness (Fant, 1973) in that [i] tends to be louder than [y]. Chinese listeners should know this (at a procedural not necessarily a declarative level), but English listeners should not. Massaro et al. (1983) utilized this logic in comparing Chinese and English speakers on the contribution of F0 pattern to identification of Chinese vowels. Given that loudness has no ecological validity in

distinguishing English vowels, the English subjects can be conceptualized as serving as a "chinchilla" control group. Thus, Chinese but not English should hear a louder vowel as more like [i] than like [y]. To test this hypothesis, five levels of formant structure between [i] and [y] were factorially combined with 3 amplitude levels, producing a total of 15 syllables. Both the Chinese and English subjects were simply instructed to identify each syllable as [i] or [y]. The results provide evidence for a psychological integration of formant structure and amplitude for Chinese listeners. It is not a psychoacoustic integration because the English subjects are not influenced by amplitude even though they use formant structure in the same manner as the Chinese subjects. Figure 5 shows that the Chinese reveal a significantly larger effect of amplitude when the formant structure is ambiguous, exactly as predicted by the fuzzy logical model.

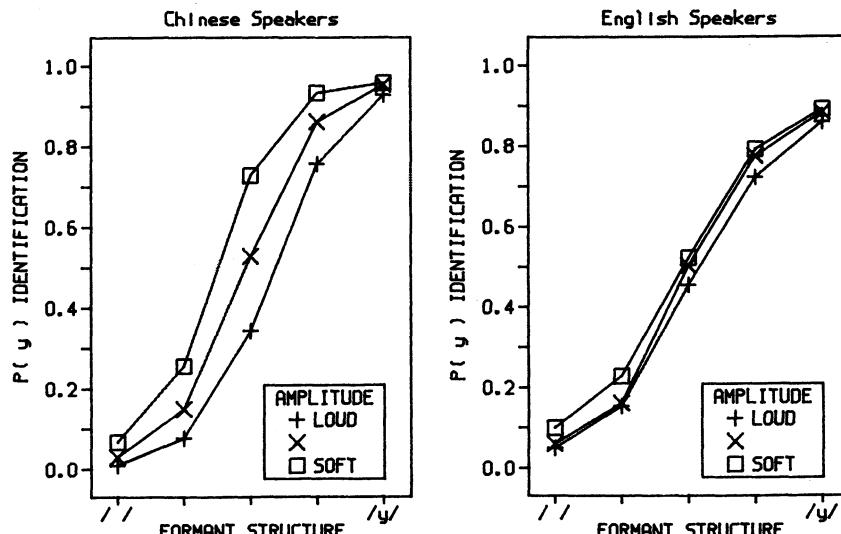


FIGURE 5, Proportion of /y/ identification for Chinese and English subjects as a function of the Formant Structure of the test syllable; the amplitude of the syllable is the curve parameter.

The English speakers show no hint of a similar effect, even though they discriminate the formant structure as well as the Chinese. The small effect of amplitude on identification for the English speakers was not statistically significant, nor, was the interaction of formant structure and amplitude.

In summary, the contribution of multiple sources of information appears to be a fundamental characteristic of speech perception. These sources of information differ for different languages and, therefore, the integration of the sources is not easily accounted for by psychoacoustic principles. The situation is better conceptualized as pattern recognition in which multiple sources of information are brought to bear on a decision. The FLMP provides a good quantitative description of the integration of the multiple sources of information across a variety of speech contrasts. The model not only describes the integration of acoustic sources but also their integration

with visible speech and with phonological, lexical, syntactic, and semantic sources of information (Glucksberg, Kreuz, & Rho, 1986; Massaro, in press c).

REFERENCES

1. Anderson, N.H. (1981). Foundations of information integration theory. New York: Academic.
2. Anderson, N.H. (1982). Methods of information integration theory. New York: Academic.
3. Best, C.T., Morrongiello, B. and Robson, R. (1981). Perceptual equivalence of acoustic cues in speech and nonspeech perception. Perception and Psychophysics, 29, 191-211.
4. Brunswik, E. (1952). The conceptual framework of psychology. Chicago: University of Chicago Press.
5. Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. Chicago: University of Chicago Press.
6. Brunswik, E. (1956). Perception and the Representative Design of Psychological Experiments. Berkeley, CA: University of California Press.
7. Cutting, J.E. and Rosner, B.S. (1974). Categories and boundaries in speech and music. Perception & Psychophysics, 16, 564-570.
8. Denes, P. (1955). Effects of duration on the perception of voicing. Journal of the Acoustical Society of America, 27, 761-764.
9. Derr, M.D. and Massaro, D.W. (1980). The contribution of vowel duration, F0 contour, and frication duration as cues to the /juz-/jus/ distinction. Perception and Psychophysics, 27, 51-59.
10. Eimas, P.D. (1985). The perception of speech in early infancy. Scientific American, 252, no. 1, 46-52.
11. Eimas, P.D. and Corbit, J.D. (1973). Selective adaptation of linguistic feature detectors. Cognitive Psychology, 4, 99-109.
12. Eimas, P.D. and Miller, J.L. (1980). Contextual effects in infant speech perception. Science, 209, 1140-1141.
13. Fant, G. (1973). Speech sounds and features. Cambridge, MA: MIT Press.
14. Flege, J.E. and Hillenbrand, J. (1986). Differential use of temporal cues to the /s/-/z/ contrast by native and non-native speakers of English. Journal of the Acoustical Society of America, 79, 508-517.
15. Fodor, J.A. (1983). Modularity of Mind. Cambridge, Mass.: Bradford Books.
16. Fujisaki, H. and Kawashima, T. (1970). Some experiments on speech perception and a model for the perceptual mechanism. Annual Report of the Engineering Research Institute. University of Tokyo, 29, 206-214.
17. Ganong, W.F. (1980). Phonetic categorization in auditory word perception. Journal of Experimental Psychology: Human Perception and Performance, 6, 110-125.
18. Glucksberg, S., Kreuz, R.J. and Rho, S.H. (1986). Context can constrain lexical access: Implications for models of language comprehension. Journal of Experimental Psychology: Learning, Memory, and Cognition, 12, 323-335.
19. Hary, J.M. and Massaro, D.W. (1982). Categorical results do not imply categorical perception. Perception and Psychophysics, 32, 409-418.

20. Hirsh, I.J. (1959). Auditory perception of temporal order. Journal of the Acoustical Society of America, 31, 759-767.
21. Hoffman, H.S. (1958). Studies of some cues in the perception of the voiced stop consonants. Journal of the Acoustical Society of America, 33, 1035-1041.
22. Howell, P. and Rosen, S. (1983). Natural auditory sensitivities as universal determiners of phonemic contrasts. Linguistics, 21, 205-235.
23. Kuhl, P.K. and Meltzoff, A.N. (1982). The bimodal perception of speech in infancy. Science, 218, 1138-1141.
24. Kuhl, P.K. and Meltzoff, A.N. (1984). Infants' recognition of cross-modal correspondences for speech: Is it based on physics or phonetics? Journal of the Acoustical Society of America, 76, Suppl. 1, S80(A).
25. Kuhl, P.K. and Miller, J.D. (1975). Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar-plosive consonants. Science 190, 69-72.
26. Kuhl, P.K. and Miller, J.D. (1978). Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. Journal of the Acoustical Society of America, 63, 905-917.
27. Kuhl, P.K. and Padden, D.M. (1982). Enhanced discrimination at the phonetic boundaries for the voicing feature in macaques. Perception and Psychophysics, 32, 542-550.
28. Kuhl, P.K. and Padden, D.M. (1983). Enhanced discriminability at the phonetic boundaries for the place feature in macaques. Journal of the Acoustical Society of America, 73, 1003-1010.
29. Liberman, A.M. (1982). On finding that speech is special. American Psychologist, 37, 148-167.
30. Liberman, A.M., Cooper, F.S., Shankweiler, D.P. and Studdert-Kennedy, M. (1967). Perception of the speech code. Psychological Review, 74, 431-461.
31. Liberman, A.M. and Mattingly, I.G. (1985). The motor theory of speech perception revised. Cognition, 21, 1-36.
32. Lindau, M. and Ladefoged, P. (1986). Variability of feature specifications. In J.S. Perkell and D.H. Klatt (Eds.), Invariance and Variability in Speech Processes. Hillsdale, NJ: Lawrence Erlbaum Associates, 464-478.
33. Lisker, L. and Abramson, A. (1970). The voicing dimension: Some experiments in comparative phonetics. Proceedings of the Sixth International Congress of Phonetic Sciences, Prague, 1967, (Academia, Prague), 563-567.
34. Macmillan, N.A., Kaplan, H.L., and Creelman, C.D. (1977). The psychophysics of categorical perception. Psychological Review, 84, 452-471.
35. Massaro, D.W. (1972). Preperceptual images, processing time and perceptual units in auditory perception. Psychological Review, 79, 124-145.
36. Massaro, D.W. (1979). Reading and listening. In P.A. Kolers, M. Wrolstad, and H. Bouma (Eds.), Processing of Visible Language 1, New York: Plenum. 331-354.
37. Massaro, D.W. (1984). Building and testing models of reading processes. In P.D. Pearson (Ed.), Handbook of Reading Research. New York: Longman. 111-146.
38. Massaro, D.W. (in press a). A fuzzy logical model of speech perception. In W.A. Lea (Ed.), Towards Robustness in Speech Recognition. Apple Valley, Minnesota: Speech Science Publications.

39. Massaro, D.W. (in press b.). Categorical partition: A fuzzy logical model of categorization behavior. In S. Harnad (Ed.), Categorical Perception. New York.
40. Massaro, D.W. (in press c.). Information-processing theory and strong inference: A paradigm for psychological inquiry. In H. Heuer and A.F. Sanders (Eds.), Perspectives on Perception and Action. Hillsdale, N.J.: Erlbaum.
41. Massaro, D.W. (in press d.). Integrating multiple sources of information in listening and reading. In D.A. Allport, D.G. MacKay, W. Prinz, and E. Scheerer (Eds.), Language Perception and Production: Shared Mechanisms in Listening, Speaking, Reading and Writing. Academic Press.
42. Massaro, D.W. (in press e.). Speech perception by ear and eye. In B. Dodd and R. Campbell (Eds.), Hearing by Eye: Experimental studies in the psychology of lipreading. Hillsdale, N.J.: Erlbaum.
43. Massaro, D.W. and Cohen, M.M. (1976). The contribution of fundamental frequency and voice onset time to the /zi/-/si/ distinction. Journal of the Acoustical Society of America, 60, 704-717.
44. Massaro, D.W. and Cohen, M.M. (1977). Voice onset time and fundamental frequency as cues to the /zi/-/si/ distinction. Perception and Psychophysics, 22 373-382.
45. Massaro, D.W. and Cohen, M.M. (1983a). Categorical or continuous speech perception: A new test. Speech Communication, 2, 15-35.
46. Massaro, D.W. and Cohen, M.M. (1983b). Consonant/vowel ratio: An improbable cue in speech. Perception and Psychophysics, 33, 501-505.
47. Massaro, D.W. and Cohen, M.M. (1983c). Evaluation and integration of visual and auditory information in speech perception. Journal of Experimental Psychology: Human Perception and Performance, 9, 753- 771.
48. Massaro, D.W. and Cohen, M.M. (1983d). Phonological context in speech perception. Perception and Psychophysics, 34, 338-348.
49. Massaro, D.W. and Oden, G.C. (1980a.). Evaluation and integration of acoustic features in speech perception. Journal of the Acoustical Society of America, 67, 996-1013.
50. Massaro, D.W. and Oden, G.C. (1980b.). Speech perception: A framework for research and theory. In N.J. Lass (Ed.), Speech and Language: Advances in Basic Research and Practice, 3, New York: Academic Press. 129-165.
51. Massaro, D.W., Tseng, C.Y., and Cohen, M.M. (1982). Vowel and lexical tone perception in Mandarin Chinese: Psycholinguistic and psychoacoustic contributions. Quantitative Linguistics, 19, 76-102.
52. McClelland, J.L. & Elman, J.L. (1986). The TRACE model of speech perception. Cognitive Psychology, 18, 1-86.
53. McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. Nature, 264, 746-748.
54. Miller, J.D. (1977). Nonindependence of feature processing in initial consonants. Journal of Speech and Hearing Research, 20, 510-518.
55. Miller, J.L. and Liberman, A.M. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. Perception and Psychophysics, 25, 457-465.
56. Nusbaum, H.C. (1984). Possible mechanisms of duplex perception: "chirp" identification versus dichotic fusion. Perception and Psychophysics, 35, 94-101.
57. Nusbaum, H.C., Schwab, E.C., and Sawusch, J.R. (1983). The role

- of "chirp" identification in duplex perception. Perception and Psychophysics, 33, 323-332.
58. Oden, G.C. (1981). Fuzzy propositional model of concept structure and use: A case study in object identification. In G.W. Lasker (Ed.), Applied Systems Research and Cybernetics. Elmsford, NY: Pergamon Press.
59. Oden, G.C. and Massaro, D.W. (1973). Integration of featural information in speech perception. Psychological Review 85, 172-191.
60. Paap, K.R. (1975). Theories of speech perception. In D.W. Massaro (Ed.), Understanding Language: An Information Processing Analysis of Speech Perception, Reading and Psycholinguistics. New York: Academic Press. 151-204.
61. Pastore, R.E., Ahroon, W.A., Baffuto, K.J., Friedman, C., Puleo, J.S., and Fink, E.A. (1977). Common-factor model of categorical perception. Journal of Experimental Psychology: Human Perception and Performance, 3, 686-696.
62. Pastore, R.E., Schmuckler, M.A., Rosenblum, L., and Szczesniak, R. (1983). Duplex perception with musical stimuli. Perception and Psychophysics, 33, 469-474.
63. Patterson, J.H. and Green, D.M. (1970). Discrimination of transient signals having identical energy. Journal of the Acoustical Society of America, 48, 894-905.
64. Peterson, J.H. and Barney, H.L. (1952). Control Methods used in a study of the vowels. Journal of the Acoustical Society of America, 24, 175-184.
65. Peterson, G.E. and Lehiste, I. (1960). Duration of syllable nuclei in English. Journal of the Acoustical Society of America, 32, 693-703.
66. Pisoni, D.B. (1977). Identification and discrimination of the relative onset time of two-component tones: Implications for voicing perception in stops. Journal of the Acoustical Society of America, 61, 1352-1361.
67. Pisoni, D.B., Carrell, T.D., and Gans, S.J. (1983). Perception of the duration of rapid spectrum changes: Evidence for context effects with speech and nonspeech signals. Perception and Psychophysics, 34, 314- 322.
68. Port, R.F. and Dalby, J. (1982). Consonant/vowel ratio as a cue for voicing in English. Perception and Psychophysics, 32, 141-152.
69. Rand, T.C. (1974). Dichotic release from masking for speech. Journal of the Acoustical Society of America, 55, 678-680.
70. Rapheal, L.J. (1972). Preceding vowel duration as a cue to the voicing of the voicing characteristic of word-final consonants in American English. Journal of the Acoustical Society of America, 51, 1296-1303.
71. Repp, B.H. (1977). Interdependence of voicing and place decisions. Haskins Labs, New Haven CT, September 1977 (unpublished).
72. Repp, B.H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. Psychological Bulletin, 92, 81-110.
73. Repp, B.H. (1984). Against a role of "chirp" identification in duplex perception. Perception and Psychophysics, 35, 89-93.
74. Repp, B.H., Milburn, C., and Ashkenas, J. (1983). Duplex perception: Confirmation of fusion. Perception and Psychophysics, 33, 333-337.
75. Sawusch, J.R. and Pisoni, D.B. (1974). On the identification of

- place and voicing features in synthetic stop consonants. Journal of Phonetics, 2, 181-194.
76. Simon, C. and Fourcin, A.J. (1978). Cross-language study of speech-pattern learning. Journal of the Acoustical Society of America, 63, 925-935.
77. Stevens, K.N., Keyser, S.J., and Kawasaki, H. (1986). Toward a phonetic and phonological theory of redundant features. In J.S. Perkell and D.H. Klatt (Eds.), Invariance and Variability in Speech Processes. Hillsdale, NJ: Lawrence Erlbaum Associates. 426-449.
78. Stevens, K.N. and Klatt, D.H. (1974). Role of formant transitions in the voiced-voiceless distinction for stops. Journal of the Acoustical Society of America, 55, 653-659.
79. Studdert-Kennedy, M., Liberman, A.M., Harris, K.S., and Cooper, F.S. (1970). The motor theory of speech perception: A reply to Lane's critical review. Psychological review, 77, 234-249.
80. Summerfield, A.Q. (1982). Differences between spectral dependencies in auditory and phonetic temporal processing: Relevance to the perception of voicing in initial stops. Journal of the Acoustical Society of America, 72, 51-61.
81. Summerfield, A.Q. and Haggard, M.P. (1974). Perceptual processing of multiple cues and contexts: Effects of following vowel on stop consonant voicing. Journal of Phonetics, 2, 279-295.
82. Warren, R.M. (1974a). Auditory pattern discrimination by untrained listeners. Perception and Psychophysics, 15, 495-500.
83. Warren, R.M. (1974b). Auditory temporal discrimination by trained listeners. Cognitive Psychology, 6, 237-256.
84. Waters, R.S. and Wilson, W.A. Jr. (1976). Speech perception by rhesus monkeys: The voicing distinction in synthesized labial and velar stop consonants. Perception and Psychophysics, 19, 285-289.
85. Winitz, H., La Riviere, C., and Herriman, E. (1975). Variations in VOT for English initial stops. Journal of Phonetics, 3, 41-52.

SPEECH PERCEPTION AND THE ROLE OF LONG-TERM MEMORY

M.E.H. Schouten

Institute of Phonetics, University of Utrecht, Trans 14,
3512 JK Utrecht, The Netherlands

INTRODUCTION

The history of speech perception since the fifties has really been the history of categorical perception. Ever since it was discovered (Liberman, Harris, Hoffman, and Griffith, 1957) that particular speech stimuli are not only identified but also discriminated as members of phoneme categories, most of the research in the field of speech perception has been aimed at explaining this discovery, or at explaining it away. The nature of the early categorical discrimination data was such that both approaches seemed about equally plausible, as discrimination was never really completely categorical. Usually, all that could be claimed was that stimuli drawn from different identification categories were discriminated better than stimuli drawn from the same identification category. However, the ideal categorical discrimination case, in which within-category discrimination is at chance level and between-category discrimination is perfect, was seldom if ever observed. Depending on personal preferences one therefore had two options: one either regarded the increased sensitivity at the boundary between the categories as being due to an extra cue in the form of a threshold (the psychoacoustic approach), or one regarded the better-than-chance within-category performance as being due to a residual sensitivity to purely acoustic differences (the phonetic approach).

Logically or empirically speaking, there was never a great deal to choose between the two approaches. The main disadvantage of the psychoacoustic position was that it fell down whenever no extra cue could be found. The phonetic position, on the other hand, could only explain the categorical phenomena it observed by invoking a black box such as articulatory knowledge (see Repp, this volume) and thus shifting the explanation problem to another time and place. Nevertheless, the phonetic position had one great advantage: it assumed that speech perception is determined to a great extent by the fact that phonemes are well-established cognitive entities, although it failed to provide even the beginning of a description of the perceptual process.

This last sentence contains a non-empirical value judgment, which will be the starting point of this paper. It is that a theory of speech perception, whether it explains or rejects categorical discrimination, simply has to assume that speech is represented cognitively by a limited set of categories which play a very important role in perception. The only evidence for this position is to be found in common sense; one either finds it attractive or one does not. This is not so bad, however, since it applies to most theories. In the

terminology used in this paper, a theory is a set of qualitative assumptions about the various stages of a process and the relations between these stages; a theory cannot be falsified, only liked or disliked. A model, on the other hand, is one of an infinite set of explicit and quantitatively specified versions of a theory; it can be falsified by showing that its predictions differ significantly from the outcome of an experiment. This distinction between theory and model is in agreement with the way the two terms are generally used.

In this paper, we will first have a look at the psychoacoustic position; after concluding that psychoacoustic thresholds are seldom the cause of categorical perception, we will then try to sketch out the factors that we think do cause categorical perception; finally we will review three theories to see whether and how they incorporate these factors.

The main question will be: does the theory contain a long-term memory component, and if so, how is this component brought to bear on the perception process? One important theory will be left out of this discussion, and that is the motor theory, which claims that phonemes are represented as articulatory patterns. One reason for this omission is that this theory already has a strong advocate in this workshop; a second reason is that I have very little sympathy with it, since it does not say anything at all about non-speech, assuming apparently that non-speech categories are represented in a completely different way.

1. CATEGORICAL PERCEPTION AND PSYCHOACOUSTIC THRESHOLDS

An example of what has just been called the psychoacoustic position is Schouten (1980), who claims that the categorical perception of speech and other sounds is based on auditory threshold phenomena. Thus the sharp division of a voice-onset-time (VOT) continuum separating voiced (b, d, g) and voiceless (p, t, k) plosive consonants is considered to be due to a threshold in temporal order detection, and the categorical perception of place-of-articulation of plosives (separating e.g. p, t, and k) is explained as the result of a threshold for frequency change permitting only three categories: up, down, and no change. Of course, if one of the parameters along which a stimulus continuum varies crosses a threshold at some point within the continuum, then categorical perception of the continuum will be inevitable; however, it must be regarded as extremely doubtful that such thresholds occur within the voice-onset-time or the place-of-articulation continua. For a comprehensive review of the evidence, the reader is referred to Repp (1984); here only an outline will be given.

The first explicit statement of the view that categorical perception may be due to threshold phenomena is the common-factor theory put forward by Pastore, Ahroon, Baffuto, Friedman, Puleo, and Fink (1977). They proposed that the categorical perception of voice-onset-time is determined by a psychoacoustic threshold for onset asynchronies; if one assumes that voice-onset time is a cue to the distinction between voiced and voiceless plosive consonants, then this leads to the conclusion that the perception of voicing is at least partly the result of a psychoacoustic threshold. Experiments by Pisoni (1977) seemed to confirm this idea: using two-tone stimuli, he obtained category boundaries at onset asynchronies that seemed to be not very

different from what one would expect with a voice-onset time continuum. However, the idea of an onset-asynchrony threshold is based on the interpretation of the 75% correct identification point reported by Hirsh (1959) as a psychophysical threshold; as Rosen and Howell (this volume) point out, however, Hirsh's data do not really permit such an interpretation. As a result there is very little support for the claim that the voicing distinction is based only or even mainly on a psychoacoustic threshold. Pisoni's (1977) categorical discrimination results could be due to the amount of training he used; more about this will be said below.

The idea that the categorical perception of place of articulation might also be due to a psychoacoustic threshold derives, by analogy, from the results of Cutting and Rosner (1974), who obtained categorical perception on a non-speech rise-time continuum. Schouten (1980) guessed that this indicated the existence of a transient threshold enabling listeners to distinguish three categories of sounds: those containing rapid changes in one direction, those containing rapid changes in the other direction, and those containing no rapid changes. In Cutting and Rosner's (1974) stimuli, the changes were amplitude changes; in order for place-of-articulation continua to produce similar patterns, there would have to be a similar threshold for rapid frequency changes. However, Rosen and Howell (1981) attempted to demolish the categorical perception of amplitude rise-times, and even though in his reply Cutting (1982) succeeded in rebuilding some of it, its psychoacoustic foundations had now proved to be very shaky. Cutting did manage to show, however, that a continuum with logarithmic increments was perceived in a categorical manner. Perception of a rise-time continuum with linear increments, however, followed Weber's law and was therefore not categorical¹.

All in all, it no longer seems tenable to hold on to an assumed analogy between psychophysical thresholds for amplitude and frequency transients to explain categorical perception of place of articulation. Moreover, some experiments of my own (Schouten, 1985), involving rising and falling sinewave tones, strongly suggest that there is no such thing as a threshold which enables one to discriminate among three different frequency transition types.

More importantly, however attractive it may seem, on the one hand, to have a simple explanation of a seemingly complicated process, on the other hand, a theory of speech perception which does not take account of any cognitive representation whatever is not really acceptable. In the next section we will try to sketch the outline of an alternative theory in relation to existing research paradigms. Much of it will be speculative; it will suggest connections which have yet to be proved.

2. CONDITIONS FOR CATEGORICAL PERCEPTION

Research involving non-speech auditory stimuli has generally concentrated on very simple sounds such as pure tones and noise bands. In those cases in which more complex sounds have been used, subjects have usually been extremely well-trained, and they have been told to listen analytically and to report the presence or absence of a particular component. Occasionally, more general questions have been asked: judging "roughness" is, for example, not an analytical procedure,

but still hardly refers to subjects' everyday experience. For a survey of psychoacoustic research paradigms, see Plomp (1976).

I would like to propose that these preoccupations of psychoacoustics have usually prevented categorical perception from manifesting itself, since the degree of categorical perception could be a function of at least two factors - a stimulus/task factor and a memory factor:

1. If categorical perception is to occur along an auditory stimulus continuum, the stimuli have to be so complex spectrally or paced in such a way that analytical listening to the parameter that is being varied is difficult. Categorical perception could thus be a function of stimulus complexity and task difficulty.
2. The stimuli have to be so well-known to the listener that he need not search his memory for them: they should be readily classifiable - automatic classification is a necessary requirement for categorical perception to occur. Categorical perception could thus be a function of amount of training.

In what follows these two conditions will be elaborated, although it must be stressed once again that the second condition rests more on its inherent plausibility than on factual evidence.

2.1. Analytical vs. non-analytical tasks

Categorical perception mainly occurs with spectrally complex stimuli, in situations in which listeners cannot hear or are not given the chance to hear the varying stimulus attributes as separate components of the stimulus. Loudness, pitch, and duration are rarely discriminated categorically. With simple stimuli one is more or less restricted to these physical attributes, whereas with complex stimuli one can ask listeners to give them a label which in itself may have nothing to do with stimulus attributes.

For categorical perception to be observed, a labelling task is needed, in which attention is focussed on the stimulus as a whole, and not on (one of) its components. This accords quite well with remarks made by Burns and Ward (1978) and Cutting (1982) on high and low stimulus uncertainty: if the listener knows what aspect of a stimulus to listen for and is given enough time, perception is unlikely to be categorical. Van Heuven and Van den Broecke (1979) and Van den Broecke and Van Heuven (1983), for example, found no categorical perception of rise times with any of their stimuli, probably because they used an adjustment task, giving subjects, who knew well what to listen for, sufficient time to come up with a satisfactory match. In high-uncertainty tasks, however, rise times are often perceived categorically, other things being equal. A similar remark is made by Repp (1983), who in summarising Pisoni's work (e.g. Pisoni, 1973) and his own work (Repp, Healey, and Crowder, 1979) says that categoricity of perception depends, to some extent, on how much use can be made of auditory memory in a task: use of auditory memory decreases categoricity. Subjects differ in how much use they will make of auditory memory and in how automatically they will perceive a stimulus as one indivisible percept: some listeners are more "analytical" than others (see Foard and Kemler Nelson, 1984). It is perhaps not surprising that such interindividual differences are

especially apparent when synthetic speech sounds are used as stimuli (Repp, 1981, and Best, Morrongiello, and Robson 1981).

There seem to be two listening modes, then: a "labelling mode" and a "non-labelling mode". Except where it is based on a perceptual threshold along the stimulus continuum (and there are very few examples of this), categorical perception could be just an experimental phenomenon, which normally only occurs in the former mode. Which of the two modes is used, or to what extent, depends on the stimulus and the task, but mainly on the task. This should be investigated in an experimental design in which stimulus complexity and stimulus uncertainty are included as independent variables.

2.2. Familiar and unfamiliar sounds

Categorical perception has been observed mainly with speech stimuli, and only occasionally with other sounds. Everyone is thoroughly familiar with speech sounds and spends a great deal of time on activities involving their immediate and automatic classification. It is extremely difficult to attend to the components of speech sounds - it takes a great deal of training or the use of degraded synthetic speech stimuli to make this possible.

The question is: does this only apply to speech sounds? There are thousands of highly familiar sounds in the world around us that also seem to be classified immediately and automatically. Just as speech sounds vary with the language or the dialect we speak, these other sounds vary with the environment we live in; common to most of us are such sounds as the striking of a match, a car door being slammed, a train running, breaking glass, a dripping tap, turning the pages of a newspaper, and so on. None of these sounds is likely to occur quite as often as any speech sound, but they are all categorised immediately, despite the wide within-category variation. If they were not, we would not be well-adapted for the world we live in: the ability to classify immediately any frequently occurring sounds - this includes speech sounds - is a biological necessity.

Unfortunately, up to now no one has ever attempted to demonstrate categorical perception with familiar non-speech sounds, apart from a few experiments employing synthetic speech-like non-speech continua such as "plucks" and "bows" (Cutting and Rosner, 1974; Rosen and Howell, 1981; Cutting, 1982) - which may or may not have been very familiar to the subjects taking part, or musical intervals (Burns and Ward, 1978), presented to musicians and non-musicians. What are needed are high-stimulus-uncertainty experiments testing for categorical perception on a continuum linking two highly familiar non-speech sounds; the main difficulty would be to choose the endpoints in such a way that the continuum is not entirely meaningless. An example of a meaningful continuum could be the same note being struck on a piano and on a harpsichord, provided that listeners can be found who are highly familiar with both sounds, yet do not listen to them analytically.

Another line of research could be to investigate categorical perception as a function of training, by means of experiments in which amount of training was varied systematically. To my knowledge, such experiments have never been performed.

Like the argument about labelling in the previous section, the above argument about familiarity does not apply to cases where categorical perception is caused by a perceptual threshold along the stimulus continuum.

3. THREE THEORIES FOR CATEGORICAL PERCEPTION

In the preceding section, categorical perception was said to be conditional on task factors and category familiarity. This brings us back to our starting point, which is that any theory of categorical perception should incorporate the notion of long-term or "permanent" memory, in which categories are stored; it also means that a model derived from such a theory should specify to what extent long-term memory is used as a function of task factors in the experiment (this includes spectral and temporal complexity of the stimuli). Below, three theories about the relationship between identification and discrimination will be briefly discussed; the first two are very close to Signal Detection Theory (SDT; Green and Swets, 1966) and do not take long-term memory into account; the third theory, however, does include long-term memory.

3.1. Durlach and Braida

The best-known theory of auditory perception is the Theory of Intensity Resolution (TIR) introduced by Durlach and Braida (1969). In this theory, two "modes of memory operation" are assumed: a sensory-trace mode, in which the sensory image of the stimulus decays rapidly, and a context-coding mode, in which the sensation caused by a stimulus is compared with the range of sensations used in the experiment. A decision model and an internal-noise model relate the two modes to each other and to the response. Over the years the theory has remained largely unchanged, but in a long series of papers by various researchers at M.I.T., the models have been continually tested and modified. The question of whether TIR would be capable of explaining categorical perception is equivalent to asking whether the theory makes the right assumptions about the use of memory: it is not inconceivable that its models could be modified in such a way that a good description of categorical data is obtained, especially since categorical perception is also determined to a great extent by experimental task factors. But even with such a modification there would still be no way of knowing whether the assumptions about memory use were right; these assumptions can, at the moment, only be discussed in terms of their plausibility. In the present paper we assume, along with e.g. Massaro and Oden (1980) and Zatorre (1983), that long-term memory plays a crucial role in categorical perception; our question therefore must be: will the context-coding mode of TIR do as a representation of long-term memory? The answer is that in its original form it did not: the context-coding mode was meant to last only as long as the experiment and to encompass no more than just the sensations that are present in the experiment. A recent modification to the theory, the perceptual anchor model of context coding, presented by Braida, Lim, Berliner, Durlach, Rabinowitz, and Purks (1984), seems to open the context-coding mode to the effects of long-term memory: a perceptual anchor could be either an experiment-bound or a permanent phenomenon; the results presented by Macmillan

(this volume) are promising but still inconclusive. What is clear, however, is that an experiment-bound context-coding mode does play a part in categorical perception: Rosen (1979) has shown that categorically perceived stimulus continua are subject to range-frequency effects. Moreover, there are also the numerous "selective adaptation" studies, starting with Eimas and Corbit (1973), in which seemingly stable category boundaries are shifted by a change in the experimental context. A theory of categorical perception should therefore also take short-term context-coding into account. Whether this agrees with TIR is debatable: TIR claims that short-term context-coding plays a negligible role in discrimination experiments, provided that the stimulus range is small enough. Is it small enough in the average speech discrimination experiment? The authors of the theory to be discussed in the next paragraph prefer to assume that it is, but they are not entirely sure.

A lot has to be left open here: it is not yet clear whether perceptual anchors will do to define long-term memory categories, and we seem to be using short-term context-coding to account for "adaption" phenomena in a manner that does not entirely agree with the theory.

3.2. Macmillan, Kaplan, and Creelman

The second theory is explicitly about categorical perception and also bases itself entirely on SDT. It is the one described by Macmillan, Kaplan, and Creelman (1977); it is, in fact, a straight adaptation of TIR to the various discrimination paradigms used in categorical perception experiments. In its basic form, SDT is, as its name implies, a model for the detection of the presence or the absence of a signal: it is a decision model. The assumption which turns this model into a possible theory of categorical perception is that every form of perception is ultimately reducible to signal detection; this is the assumption underlying Macmillan et al. (1977). It has to be stressed again that, no matter how successful a model derived from SDT may be at describing categorical data, one's acceptance of the theory depends on one's acceptance of its assumptions; the ideas expressed in the present paper about long-term memory seem to be far removed from traditional SDT. Nevertheless, let us consider briefly how successful the SDT model is in dealing with categorical data.

Macmillan et al. define categorical perception as "equivalence of identification and discrimination"; this equivalence (or the lack of it) is expressed in d' terms (i.e. in terms of the "perceptual distances" between the stimuli of a continuum). They reanalyse percent correct identification and discrimination data from Cutting and Rosner (1974), which Rosen and Howell (1981) have since shown to be flawed. We will therefore not consider them here, but turn instead to Rosner (1984), who, just like Macmillan et al., performed a reanalysis in d' terms of other researchers' data in order to defend SDT as a theory of categorical perception. None of these reanalysed studies met the criteria for an SDT-type analysis; these criteria are that all stimuli should be presented a large number of times and that d' measures should be calculated separately for each subject. Rosner's own experiments, reported in the same paper, did meet SDT criteria, however. In his own experiment I, Rosner tested the hypothesis that

any difference between identification and discrimination of a speech continuum (in terms of d') is due to an under-differentiated set of identification categories; he therefore used an absolute identification paradigm (a separate category for each stimulus type) and a fixed-level ABX discrimination task on an eight-stimulus voice-onset-time continuum; the results, however, showed the usual effect of discrimination being better overall than identification, and seemed to support a dual-coding theory such as the one to be discussed below as our third theory of categorical perception. Rosner suspected, however, that the difference between identification and discrimination in this experiment could be due to the relative ease of the fixed-level discrimination task as compared to the absolute identification task; he therefore conducted a second experiment, in which the fixed-level discrimination task was replaced by a roving-level one, but the results were essentially the same, indicating that task differences could not account for the difference between identification and discrimination results. Although Rosner's experiments had been designed for an SDT-type analysis (apart from pooling the individual results before the analysis), the results seemed to indicate that categorical perception could not be explained in terms of d' -equivalence between identification and discrimination. Rosner then proceeded, as Macmillan et al. had done, to reanalyse a number of other categorical-perception studies which had not been designed for an SDT-type analysis, and all of which used two-category identification. He now found that between-category discrimination was hardly better than identification and concluded that this deviation from his own results was due to the fact that in these other studies judgemental stability was not affected by too large a response set: the response set was just right. Within-category discrimination in these reanalysed studies, however, was much higher than identification; this is once again blamed on the response set for identification being too small.

The most striking aspect of this series of arguments is that Rosner prefers an unsuitable reanalysis of old data to his own perfectly suitable new data, mainly because he prefers one theory to another one. Most researchers would have changed the parameters of the model to fit their data; Rosner rejects his own data. From the point of view of the theory there is no difference either way: its acceptance depends on its appeal. If any conclusion is to be drawn from Rosner's paper, however, it would seem to be just the opposite: SDT does not provide a good description of categorical perception. The data indicate that a dual-coding theory, which incorporates a long-term memory component, would be a more suitable candidate.

3.3. Fujisaki and Kawashima

The third theory of categorical perception is the Dual-Process Theory (DPT), proposed by Fujisaki and Kawashima (1971). The assumptions behind this theory are:

- (1) "in a discrimination procedure, ... identification of individual stimuli always precedes discrimination judgment, which requires the reception of all ... stimuli in a trial";
- (2) "... a properly selected monotone division (of a narrow stimulus range) produces an interval scale on the perceptual continuum";
- (3) "Only when the results of categorical judgment are useless for forced discrimination, subjects have recourse to comparative

judgment of the ... stimuli on the perceptual continuum of timbre, retrieved from the short-term memory for timbre."

It is hazardous to make a direct comparison between DPT and TIR. The main difficulty lies in the status of the level or mode defined by assumptions (2) and (3). On the one hand, the stimulus timbre stored in this "short-term memory for timbre" is said to be an "analog and continuous form of information", which would seem to place it very close to a sensory trace. On the other hand, this short-term memory can retain information about three stimuli (ABX discrimination procedure) for at least a few seconds, which seems to bring it closer to context-coding, if one assumes that a sensory trace is equivalent to Massaro's (e.g. 1974) precategorical storage, which has a life-span of about 300 ms. Such an assumption would relegate the trace mode to a level below the short-term memory for timbre of DPT, perhaps to a level called "the process of mapping" by Fujisaki and Kawashima, where Macmillan et al. (this volume) would probably place what they call "sensory variance", but certainly not "trace variance".

Although it is unclear, then, how exactly the two theories are related and where, if anywhere, Massaro's 300 ms store fits in, assumption (1) of DPT explicitly introduces a long-term-memory mode, which is absent from the other theories (although it may recently have crept in in the form of perceptual anchors), but which is indispensable if one wants to account for the perception of well-known (speech) sounds. It should be remarked at this point that the processes mentioned here are quite different from the three factors (auditory, phonetic, and phonemic) proposed by Werker and Logan (1985): their phonetic level appears to be situated between auditory context coding and phonemic memory. Apart from these considerations, DPT does what we have required of a categorical perception theory: it states that categorisation is automatic and precedes "lower-level" types of discrimination. Any model derived from this theory should specify how the degree of familiarity of a category and the degree of complexity of a stimulus influence the use of long-term memory, and how task factors and other experimental context affect short-term (context-coding) memory operation and its relation to long-term memory. In its present form, the theory only states that the use of short-term memory is inversely related to the use that can be made of the long-term-memory mode, and that is not enough.

Macmillan et al. (1977) criticised DPT as unparsimonious: they compared it to a weight-comparison theory in which two weights are not compared directly, but in which each weight is first compared separately with a standard weight. This objection misses the point since it ignores the fact that categorisation of incoming sounds is virtually automatic and inescapable, whereas analysis of sounds is, by comparison, highly unusual.

Rosner (1984) claims that his findings "... undercut a major argument for dual coding theory. These results and the quantitative failure of that theory now form serious grounds for its abandonment". In the terminology used in the present paper, a theory cannot fail quantitatively: only a model can. If DPT had been tested and refined as extensively as TIR, it is quite conceivable that its position as a theory of categorical perception would have been comparable to the

position of TIR in relation to intensity perception. Unfortunately, however, nothing of the kind has happened, so that DPT is still only judged on the strength of the first quantitative models derived from it in 1971.

Implicit support for dual processing comes from an experiment by Spiegel and Watson (1981), in which subjects received very extensive training in the perception of eight temporally complex (but spectrally simple) tonal patterns. Each of these patterns consisted of ten 40 ms tones of strongly varying frequencies; training was in the form of various AX discrimination paradigms, in which the standard was one of the eight tonal patterns and the comparison pattern differed from the standard only in the frequency of one of the ten tones. After the discrimination sessions (the results of which are interesting but do not concern us here), two of the subjects were tested in a remembered standard paradigm, in which only one pattern (a standard or a comparison) was presented per trial. Subjects had to indicate whether the pattern presented was one of the eight standard patterns or not. The two subjects were remarkably successful in this task, reaching discrimination levels (in $\Delta f/f$ for $d'=10$) comparable to those obtained in the same-different procedures. This shows that these subjects had managed to "learn" these patterns, so that a reasonably speech-like situation had been created; the authors themselves suggest that if one really wants to compare the perception of speech and non-speech, one must make one's subjects "learn" the non-speech signals through prolonged exposure. Obviously, the non-speech signals need to be of the same order of complexity as the speech signals; with an abrupt change every 40 ms the non-speech patterns used by Spiegel and Watson were more complex temporally than the average speech signal, but with only one spectral component they were much less complex spectrally. The most important result from Spiegel and Watson's experiments with respect to dual-process theory is that there was no effect of position in the remembered standard paradigm: it made no difference whether the changed tone occurred early or late in the pattern. The authors reason that if a pattern could be identified by its first few components (which were fairly unique across the eight standard patterns), performance would have been worst for changed tones near the onset of the pattern, before the pattern had been identified. They conclude that a top-down strategy must be at work here, in which subjects first identify a complete pattern on the basis of "overall characteristics", and that the more detailed information available in the auditory memory is only examined later.

One cannot be certain that the conclusions drawn by Spiegel and Watson with respect to dual processing may be generalised to other situations. As has already been remarked, the stimuli used by them were uncommonly complex in the time domain; moreover, the number of subjects was very low. However, we have here a very strong suggestion that stimuli are first categorised and only then examined in greater detail, and that consequently dual-processing theory may be correct, even though a particular model derived from it has turned out to be too simple. DPT should certainly not be abandoned; perhaps it should be integrated with TIR. In the next section a few tentative suggestions about such an integration will be made. As things stand now, TIR is an incomplete theory, whereas DPT has so far remained an unsuccessful theory.

4. CONCLUSIONS

This paper contains two suggestions for future research into the nature of categorical perception:

- (1) in order to demonstrate that categorical perception depends on stimulus familiarity, experiments need to be conducted in which the familiarity of complex sounds is varied systematically;
- (2) dual-process theory, which is the only theory of categorical perception incorporating long-term memory, should be tested and refined as extensively as the theory of intensity resolution has been.

So far, we have been discussing theories rather than models. We have assumed implicitly that the most plausible theory will also generate the best models, and that therefore the dual-processing models are the ones to be tested and refined. However, we do have a choice in this matter: we can either refine dual-process theory, or we can extend the theory of intensity resolution by adding long-term memory to it, either as a separate level or in the form of the new perceptual anchors. The choice does not seem difficult: if we opt for extending TIR, we will be joining a long-established theoretical tradition, in which some of the levels that are needed for a theory of categorical perception have been modelled and tested very thoroughly. The same cannot be said for DPT, which unfortunately has remained an isolated attempt to build a new theory.

As was emphasized earlier, any model of long-term memory will have to specify to what extent its contribution to the perceptual process is determined by the familiarity of the categories, the complexity of the stimuli, and the difficulty of the experimental task. When well-known sounds such as speech sounds are used as stimuli, the sensory-trace level and the context level will determine the positions of the sensation and of the criteria along the decision axis, but this will be followed immediately and overwhelmingly by categorisation. Only to the extent that categorisation is not successful, or to the extent that the experimental situation allows it, will recourse be had to the sensation itself. It is this important notion from Fujisaki and Kawashima which should be incorporated into a theory of signal detection in order for it to become more of a theory of the perception of familiar everyday sounds.

Note 1

Kewley-Port and Pisoni (1984) tried to demonstrate that Cutting's logarithmic stimuli also followed Weber's law. They carried out two experiments: in one experiment they reproduced Cutting's (1982) linear rise-time results; in a second experiment they used an adaptive tracking procedure, in order to determine JND's for linearly spaced rise times (in a 2IFC task subjects had to indicate which of two stimuli had "the more gradual onset"). They then calculated a function to relate the two sets of results to each other and used this function to predict JND's for Cutting's logarithmic data. It is a pity they did not follow a more direct route by determining JND's for logarithmically spaced stimuli in an extra experiment. Their claim that the predicted JND's for logarithmic stimuli are also in agreement with

Weber's law is based on the fact that the predicted values for Cutting's logarithmic stimuli appear to lie on a straight line, and that the correlation between these values and the corresponding segment of the adaptive-tracking curve is very high. Neither argument carries conviction - they are both based on extrapolation, and the correlation argument even lacks face validity. Cutting's (1982) results still stand, therefore, and so does his conclusion that categorical rise-time perception must be a "fickle" phenomenon, since a continuum spanning the same rise-time range can be perceived both categorically, if the stimuli are spaced logarithmically, and in accordance with Weber's law, if the stimuli are spaced linearly.

REFERENCES

1. Best, C.T., Morrongiello, B., and Robson, R., 1981. Perceptual equivalence of acoustic cues in speech and nonspeech perception. Perception and Psychophysics 29: 191-211.
2. Braida, L.D., Lim, J.S., Berliner, J.E., Durlach, N.I., Rabinowitz, W.M., and Purks, S.R., 1984. Intensity perception XIII. Perceptual anchor model of context-coding. Journal of the Acoustical Society of America 76: 722-731.
3. Broecke, M.P.R. van den, and Heuven, V.J. van, 1983. Effect and artifact in the auditory discrimination of rise and decay time: Speech and nonspeech. Perception and Psychophysics 33: 305-313.
4. Burns, E.M., and Ward, W.D., 1978. Categorical perception - phenomenon or epiphenomenon: Evidence from experiments in the perception of melodic musical intervals. Journal of the Acoustic Society of America 63, 456-468.
5. Cutting, J.E., 1982. Plucks and bows are categorically perceived, sometimes. Perception and Psychophysics 31: 462-476.
6. Cutting, J.E., and Rosner, B.S., 1974. Categories and boundaries in speech and music. Perception and Psychophysics 16: 564-570.
7. Durlach, N.I., and Braida, L.D., 1969. Intensity perception: I. Preliminary theory of intensity resolution. Journal of the Acoustical Society of America 49: 372-383.
8. Eimas, P.D., and Corbit, J.D., 1973. Selective adaptation of linguistic feature detectors. Cognitive Psychology 4: 99-109.
9. Foard, C.F., and Kemler Nelson, D.G., 1984. Holistic and analytic modes of processing: the multiple determinants of perceptual analysis. Journal of Experimental Psychology: General 113: 94-111.
10. Fujisaki, H., and Kawashima, T., 1971. A model of the mechanisms for speech perception quantitative analysis of categorical effects in discrimination. Annual Report of the Engineering Research Institute, 30, Faculty of Engineering, University of Tokyo: 59-68.
11. Green, D.M., and Swets, J.A., 1966. Signal detection theory and psychophysics. New York: Wiley.
12. Heuven, V.J.P. van, and Broecke, M.P.R. van den, 1979. Auditory discrimination of rise and decay time in tone and noise bursts. Journal of the Acoustical Society of America 66: 1308-1315.
13. Hirsh, I.J., 1959. Auditory perception of temporal order. Journal of the Acoustical Society of America 31: 759-767.
14. Kewley-Port, D., and Pisoni, D.B., 1984. Identification and discrimination of rise time: Is it categorical or non-categorical? Journal of the Acoustical Society of America 75: 1168-1176.
15. Liberman, A.M., Harris, K.S., Hoffman, H.S., and Griffith, B.C., 1957. The discrimination of speech sounds within and across

- phoneme boundaries. Journal of Experimental Psychology 54: 358-368.
16. Macmillan, N.A., Kaplan, K., and Creelman, C.D., 1977. The Psychophysics of categorical perception. Psychological Review 84: 452-471.
17. Massaro, D.W., 1974. Perceptual units in speech recognition. Journal of Experimental Psychology 102: 199-208.
18. Massaro, D.W., and Cohen, M.M., 1983. Categorical or continuous speech perception: a new test. Speech Communication 2: 15-35.
19. Massaro, D.W., and Oden, G.C., 1980. Speech perception: a framework for research and theory. In: Lass, N.J., (ed.) Speech and language: advances in basic research and practice (vol. 3). New York: Academic Press.
20. Pastore, R.E., Ahroon, W.A., Baffuto, K.J., Friedman, C., and Pulleo, J.S., 1977. Common-factor model of categorical perception. Journal of Experimental Psychology: Human Perception and Performance 3: 686-696.
21. Pisoni, D.B., 1973. Auditory and phonetic memory codes in the discrimination of consonants and vowels. Perception and Psychophysics 13: 253-260.
22. Pisoni, D.B., 1977. Identification and discrimination of the relative onset of two component tones: Implications for the perception of voicing in stops. Journal of the Acoustical Society of America 61: 1352-1361.
23. Plomp, R., 1976. Aspects of tone sensation. Academic Press.
24. Repp, B.H., 1981. Two strategies in fricative discrimination. Perception and Psychophysics 30: 217-227.
25. Repp, B.H., 1982. Phonetic trading relations and context effects: new evidence for a phonetic mode of perception. Psychological Bulletin 92: 81-110.
26. Repp, B.H., 1984. Categorical perception: issues, methods, findings. In: Lass, N.J., (ed.) Speech and language: advances in basic research and practice, vol. 9. Academic Press.
27. Repp, B.H., Healy, A.F., and Crowder, R.G., 1979. Categories and context in the perception of isolated steady-state vowels. Journal of Experimental Psychology: Human Perception and Performance 5: 129-145.
28. Rosen, S.M., 1979. Range and frequency effects in consonant categorization. Journal of Phonetics 7: 393-402.
29. Rosen, S.M., and P. Howell, 1981. Plucks and bows are not categorically perceived. Perception and Psychophysics 30: 156-168.
30. Rosner, B.S., 1984. Perception of voice-onset-time continua: A signal detection analysis. Journal of the Acoustical Society of America 75: 1231-1242.
31. Schouten, M.E.H., 1980. The case against a speech mode of perception. Acta Psychologica 44: 71-98.
32. Schouten, M.E.H., 1985. Identification and discrimination of sweep tones. Perception and Psychophysics 37: 369-376.
33. Spiegel, M.F., and Watson, C.S., 1981. Factors in the discrimination of tonal patterns. III. Frequency discrimination with components of well-learned patterns. Journal of the Acoustical Society of America 69: 223-230.
34. Werker, J.F., and Logan, J.S., 1985. Cross-language evidence for three factors in speech perception. Perception and Psychophysics 37: 35-44.

35. Zatorre, R.J., 1983. Category-boundary effects and speeded sorting with a harmonic musical-interval continuum: evidence for dual processing. Journal of Experimental Psychology: Human Perception and Performance 9: 739-752.

LEVELS OF REPRESENTATION OF PHONEMES AND BANDWIDTH OF SPECTRAL-TEMPORAL INTEGRATION*

B. Espinoza-Varas

Speech and Hearing Sciences, Indiana University, Bloomington,
Indiana, 47405, USA.

INTRODUCTION

Studies on the psychophysics of speech perception seek to discover correlations between performance in psychoacoustic tests using simple stimuli, and performance on speech tests. Such correlations may be numerical (e.g., Moore and Glasberg, this volume) or qualitative (e.g., the demonstration that "categorical perception" can be observed in both speech and psychoacoustic tests). A third group of studies correlates the output of a psychoacoustic model of auditory processing with human speech perception (e.g., Zwicker et al., 1979).

The goal of establishing simple, general relations between psychoacoustics and speech perception has been hindered by the fact that there are many perceptual tests that can be used in the correlations and no definite criteria to select one versus another. This is illustrated in Table I.

In the first place, there are many different measures of psychoacoustic performance; second, the representation of phonemes characteristically involves several perceptual attributes (e.g., phonetic label, timbre, pitch, etc.), and the response can be one of identification, similarity rating, discrimination, or detection, using a variety of alternative speech materials. Which psychoacoustic measure is the appropriate one to correlate with a specific measure of speech perception? Often, the choice of measures is somewhat arbitrary. This problem remains even if some simplifying assumptions are made about the sensory processing involved in speech. For instance, there is agreement that extraction of vowel formant frequencies involves some sort of frequency analysis by the ear. Thus, it seems almost logical to expect that vowel perception should correlate with psychoacoustic measures of frequency analysis or frequency discrimination. However, there is a variety of measures of frequency analysis: psychophysical tuning curves (PTC), critical bandwidth (CBW), critical ratio, frequency discrimination, pitch scaling, and others. Any one of these measures could be related to a variety of aspects of vowel perception. Moreover, many studies report weak or no correlations between certain measures of frequency analysis and speech perception (e.g., Moore and

*Preparation of this manuscript was supported in part by NIH and AFOSR grants to Indiana University. I thank Dr. Charles S. Watson for many stimulating discussions on several of the issues addressed in this paper. Dr. Gary Kidd provided valuable comments on earlier versions of this paper.

Glasberg, this volume; Stelmachovicz et al., 1985). For instance, abnormally wide critical bands may coexist with normal vowel intelligibility (e.g., Turner and Van Tassel, 1984).

The general aim of this paper is to describe correspondences between psychoacoustics and speech perception, which could help define some criteria for mapping the two domains. Specifically, a parallel is envisioned between different levels of perceptual representation of phonemes and different psychoacoustic bandwidths of spectral-temporal integration.

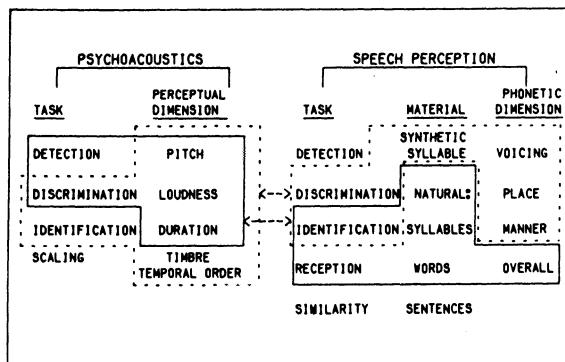


Table I, Brief catalogs of psychoacoustic measures and of speech perception measures. Performance on any psychoacoustic task and perceptual dimension can, in principle, be correlated with any speech task, material, or phonetic dimension. Approximately 1600 correlations are possible with these small catalogs. Very few of the many correlations that can be calculated have been examined experimentally. Studies searching for numerical correlations generally employ the measures enclosed by solid lines. Those searching for qualitative correlations generally employ the measures enclosed by dashed lines. The later studies have reported very close correspondences between some psychoacoustic and speech perception tests (e.g., Macmillan et al., this volume).

1. MULTIPLE PERCEPTS INDUCED BY PHONEMES

A very prominent characteristic of the representation of simple speech sounds is that it consists of multiple perceptual effects. For example, the syllable /pae/ can be "heard" in a "phonetic labelling mode", which involves primarily a decision about the appropriate phonetic label for the sound (i.e., /pae/ rather than /bae/ or /dae/). This type of decision is typically sensitive to fairly large, multidimensional changes in the stimulus. This is illustrated by the labelling function of Fig. 1A, which shows that the VOT of a prototypical /pae/ stimulus must be decreased by as much as 40-50 ms to change the label from /pae/ to /bae/¹. In natural speech, this voicing distinction is indicated by a number of other cues, besides VOT (e.g., Summerfield and Haggard, 1977). In the phonetic-label mode of "listening", very large variations in the speech sound are to a great extent ignored. For instance, large changes of the fundamental

frequency or loudness can have small effects on the label assigned to the sound.

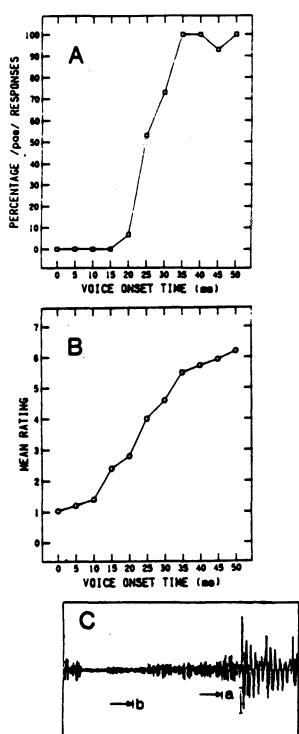


FIGURE 1. Typical identification (panel A) and rating (panel B) functions obtained with a synthetic voice-onset-time continuum. Panel (C) illustrates the JND for a base VOT of 50 ms (a). Also indicated is the amount by which the 50-ms VOT of a good /pae/ stimulus must be decreased to cause the perception of /bae/ (b). Judgements of phonetic goodness would be based on changes intermediate between (a) and (b).

At a second perceptual level, listeners can carry out intraphonemic distinctions concerning aspects of the "phonetic goodness" of the syllables relative to an internalized prototype, and decide whether it was a poor or good exemplar of a prototypical /pae/. Figure 1B shows rating of phonetic goodness of the stimuli of a synthetic VOT continuum. The rating function essentially lacks asymptotic regions, showing that at this perceptual level subjects can distinguish between stimuli that are identified with the same label. This means that decisions about the phonetic goodness of the stimuli rely on much smaller changes in VOT (relative to decisions about the phonetic label).

In addition to the above, listeners can hear the syllable in an auditory mode and decide whether it was high or low in pitch, or whether it was soft or loud. Under appropriate experimental conditions, listeners can attend to very subtle aspects, such as the amount of aspiration preceding the vowel, or the loudness of the consonant release burst, and make discriminations that approach the limits of auditory resolving power (Macmillan et al., this volume). This analytical listening characterizes the auditory level of representation. Fig. 1C summarizes the magnitude of VOT change that seems to correspond to each level of representation.

A similar multiplicity of perceptual effects, each with a different sensitivity to stimulus change, can be observed with vowels. For example, the steady state vowels /u/ and /i/ show frequency differences in all three formants. The difference in the second formant alone amounts to 1400 Hz or about six critical bands (Peterson and

Barney, 1952). Assigning a phonetic label to these vowels probably depends on the combined effect of differences in all three formants (Mermelstein, 1978; Flanagan, 1972). Decisions at the auditory level, on the other hand, are based on much subtler stimulus changes. A change of about half a critical bandwidth is sufficient to just note a difference in the frequency of a single formant (Flanagan, 1972); decisions about the phonetic goodness of these vowels must be based on changes intermediate between the JND's and the changes that cause the perception of a different vowel class, i.e., changes of 1-2 critical bands (Pols et al., 1969; Espinoza-Varas, this volume). Figure 2 summarizes various perceptual effects associated with various amounts of change in the frequency of the second formant of a synthetic vowel.

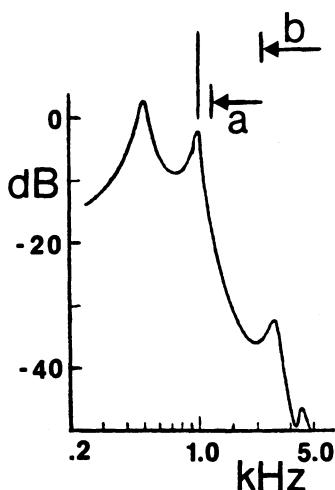


FIGURE 2, Perceptual effect associated with different amounts of change in the frequency of the second formant of a synthetic vowel. The just noticeable change in F2 frequency (a), and the change required to induce the perception of a different vowel (b) are outlined.

The interesting results are: a) subjects seem to "hear" the phonemes (or respond to the sounds) in different ways depending on the type of perceptual judgment that they are asked to make; and b) decisions about the phonetic identity of the sounds rely on large, multidimensional stimulus changes, while decisions about auditory differences are based on much subtler, unidimensional stimulus changes. Based on the sensitivity to stimulus change, the various perceptual effects elicited by speech sounds can thus be organized as a hierarchy of levels.

The multiple perceptual effects that comprise the representation of phonemes have been described earlier (e.g., Ladefoged and Broadbent, 1957), and are recognized also in recent multiple-code or dual-code models of speech perception (Fujisaki and Kawashima, 1970; Pisoni, 1973, 1975; Samuel, 1977; Soli, 1983; Repp, 1983). It appears that these various perceptual effects are simultaneously available, and that to some extent, the listener can choose to attend to any one of them. However, the phonetic labelling mode tends to be the dominant mode when listening to connected speech.

2. POSSIBLE BASIS FOR A CORRESPONDENCE BETWEEN LEVEL OF REPRESENTATION AND BANDWIDTH OF INTEGRATION

What kind of psychoacoustic mechanism would need to be invoked to account for the multilevel representation of phonemes? Flanagan (1972) has suggested the following possibility:

"speech perception is an adaptive process. It is a process in which the detection procedure probably is tailored to fit the signal and the listening task. If the listener is able to impose a linguistic organization upon the sounds, he may use information that is temporally dispersed to arrive at a decision about a given sound element. If such association is not made, the decision tends to be made more upon the acoustic factors of the moment and in comparison to whatever standard is available. The suggestion that a listener uses temporally spread information raises the question as to the size of the temporal "chunks" in which speech is perceived. Very probably the size of the perceptual element varies with the discrimination task, and the listener adjusts his processing rate to suit different types of speech information (pp. 305-306)."

Psychoacoustically, an adaptive process of this sort would be realized partly in a system with an adjustable analysis bandwidth, in both the spectral and the temporal domain. The specific bandwidth used is that which is optimal to the specific task demands.

3. BANDWIDTHS OF SPECTRAL AND TEMPORAL INTEGRATION

Several different bandwidths of spectral integration have been described in the psychoacoustic literature, ranging from very broad to narrow bandwidths.

Figure 3 shows sketches of the spectral detail that would be preserved with different bandwidths of spectral integration, ranging from very broad to very narrow. In a recent study, Spiegel (1979) reported that the maximum range of spectral integration can be as wide as 3.0 kHz. (Fig. 3a). The ear can integrate frequencies from as many as twelve critical bands, if the task requires the subject to do so. At another integration level, the auditory system seems to integrate energy distributed across only a few critical bands (Fig. 3b). Karnickaya et al. (1975) and Chistovich and Lublinskaya (1979) have proposed an integrating bandwidth encompassing 3.5 critical bands for the process of phonetic classification or labelling. A narrower bandwidth of spectral integration is the familiar critical band described by Fletcher about 50 years ago (Fig. 3c). It has been suggested (Houtgast, 1974; Karnickaya et al., 1975) that the frequency analysis achieved at the level of the critical bandwidth could be sharpened by lateral inhibition processes to obtain a still sharper frequency analysis (or integration). (Fig. 3d).

As in the frequency domain, several temporal integration constants have been described in psychoacoustic studies (Zwicker, 1973; Hirsh, 1974). Zwicker (1973) distinguished three integration constants: a long constant of about 200 ms, a medium integration constant of 20 ms, and a minimum integration constant of 2.0 ms, defined as the shortest temporal structure of the stimulus which can

influence the excitation pattern. Hirsh (1974) defined similar integration constants on the basis of studies of perception of temporal order. He concluded that the long integration constants could be involved in tasks requiring verbal labelling of temporal order. The shortest constants would be involved in discriminations based only on changes of quality of fused events, that is, events in which the listener cannot label the individual components.

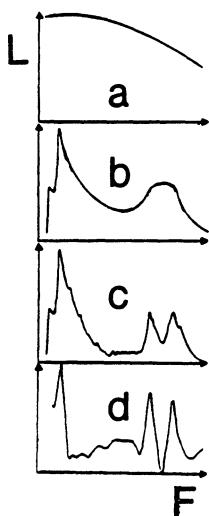


FIGURE 3, Sketch of the degree of spectral detail with four different bandwidths of spectral integration: a) maximum spectral integration. The tilt of the spectrum is preserved; b) multiple critical band integration. Formant peaks separated by several critical bands are preserved. Formant peaks separated by 1-3 critical bands are fused; c) critical band integration. All formant peaks are preserved; d) subcritical band integration. The just noticeable differences in formant frequency are preserved. Panels (C) and (D) were adapted from Karnickaya et al.(1975).

In summary, the narrower integration bandwidths appear to correspond to the sharper sensitivity of the auditory component of the representation of phonemes. The broader integration bandwidths appear to correspond to the coarser sensitivity of the phonetic-label component of the representation. It has been suggested (Zwicker, 1973) that these various integration bandwidths are also simultaneously available to listeners, and whether one or the other is engaged depends on the specific demands of the experimental task. The correspondence between bandwidth and level of representation suggests a tentative guideline for mapping psychoacoustic performance to speech perception. Analytical measures of auditory resolving power (e.g., frequency difference limen) would more likely relate to aspects of the auditory component of speech representation. Measures of coarser, global psychoacoustic processing would more likely relate to aspects of phonetic labelling (e.g., frequency proximity at which two formants are perceived as a single fused formant).

The various bandwidths of spectral and temporal integration would essentially provide something like a zoom mechanism, by means of which several degrees of stimulus detail are simultaneously available. The auditory representation contains a large amount of

detail, which is filtered out upstream in the processing. Initially, the perceptual system would attempt decoding of the speech signal on the basis of very gross features; if this is not sufficient, the more detailed representations are relied upon.

4. FURTHER EVIDENCE THAT SUPPORTS THE PARALLEL

A. Spectral-temporal distribution of acoustic cues that distinguish phonemes.

Two of the most prominent characteristics of the speech code are its redundancy and robustness. At the phonemic level, these characteristics imply that the acoustic differences or cues that distinguish phonemes are both numerous and large relative to the limits of auditory resolution.

Usually, both temporal and spectral cues coexist, and they can be distributed through almost the entire range of speech frequencies, and over intervals of up to 300 ms (Stevens, 1975). For example, the distinction /ada/ versus /ata/ may be cued by any of the following: a) duration of the initial vowel, b) presence or absence of prevoicing, c) loudness of the release burst, d) VOT, e) presence of low frequency energy at the onset of voicing, and f) fundamental frequency of the final vowel (Klatt, 1975). Temporally, these cues are distributed over an interval of 200-250 ms. Spectrally, they are distributed over a range of 4-5 octaves.

The function of the redundancy of the cues is to increase the robustness of the code, and insure rapid, accurate identification or labelling. From an informational point of view, the most efficient way of listening to such a display (i.e., the one that would transmit the greatest amount of information in the stimulus) is to attempt to listen to the entire set of cues, and to make perceptual decisions on the basis of the combined information provided by the entire set. That is, the most efficient way of listening is to integrate evidence across a wide bandwidth in frequency and in time. Perceptual studies show that listeners effectively employ all available cues, however small their perceptual magnitude (e.g., Minifie et al., 1973; Lacroix et al., 1979). Espinoza-Varas (1983), and Espinoza-Varas and Jamieson (1984) reported that listeners are able to integrate information spread over widely separated time intervals and frequency regions. Studies of information transmission with multidimensional auditory displays (Garner, 1962) show that such a broad-bandwidth listening strategy affords the greatest overall entropy, but it also causes loss of detail on each of the individual cues. In other words, the assumption of a broad-band listening strategy for accurate phonetic labelling accounts both for the subjects' ability to utilize all cues available, and for their reduced sensitivity to the details of the individual cues.

B. Changes of phonetic identity versus auditory differences in other phonetic constraints.

As noted in Section 1, the perceptual evidence consistently shows that the stimulus change required to induce a change of phonetic label is considerably larger than that required to induce the detection of a difference (when measured under optimal experimental conditions). In Section 1, the cases of voicing decisions and vowel

formant frequency were discussed. Other studies pointing to the same generalization are briefly described below.

Phonetic differences between vowel sounds can be indicated by durational cues, in addition to those of formant frequency, as occurs in the case of /bed/ versus /bad/. The duration difference between the two vowels in these syllables can be as large as 40% (Klatt, 1976). However, the smallest difference in vowel duration that is discriminable is on the order of 10% of the vowel base duration. Studies on perception of temporal order of vowel sounds show that to correctly identify (i.e. label phonetically) the temporal order in vowel sequences, listeners require component durations of no less than 125-250 ms (Thomas et al., 1970). However, component durations as short as 5.0 ms are sufficient to just discriminate the two sequences as different, though the listener is unable to describe the nature of the difference in verbal terms (Kerivan et al., 1981). Thus, durational changes that cue different phonetic labels are considerably larger than the minimum change required to detect a difference in duration. Turner and Van Tassel (1984) reported that the ability to discriminate changes in the depth of spectral notches in synthetic vowels is quite accurate compared to the typical depth of spectral notches observed in naturally produced vowels. The same discrepancy between the change required to alter the label and that required to just note a difference is observed for the case of the onset frequency of F2, which serves as a cue for place of articulation in stop consonants.

5. CONCLUSIONS

A parallel was established between levels of representation of phonemes and bandwidths of spectral-temporal integration defined psychoacoustically. The parallel was based on the following premises: a) the representation of phonemes consists of multiple perceptual effects including phonetic labelling, phonetic goodness, and auditory level; b) the levels are distinguished on the basis of the magnitude of stimulus change (either fine or gross) which determines the respective perceptual decision: at the level of phonetic labelling, decisions are based on gross, multidimensional stimulus changes, while at the auditory level they are based on much finer, unidimensional stimulus changes; c) differences in sensitivity to stimulus changes seem to correspond with psychoacoustically defined bandwidths of spectral and temporal integration; d) evidence with a variety of phonemic contrasts shows that changes of phonetic identity require a magnitude of stimulus change approaching the broader integration bandwidths, and the magnitude of stimulus change required to just notice a difference approaches the narrower integration bandwidths; e) the multiple acoustic cues that differentiate phonemes are often distributed over long time intervals and broad frequency regions. Thus, broad-bandwidth listening is required to process the entire set of cues; f) it is proposed that a correspondence between bandwidth of integration and levels of phoneme representation may be used as a criterion for mapping psychoacoustic measures to speech perception.

NOTE

1. At the boundary, the sensitivity to VOT changes is greater than at other points in the continuum. But VOT values near the phonetic boundary may occur only with synthetic speech. In English, the

distributions of VOT values for naturally produced /bae/ and /pae/ syllables are well separated (Lisker and Abramson, 1964). In addition, when the VOT cue is ambiguous, other cues are used to signal the differences (Summerfield and Haggard, 1977). The stimuli near the boundary of a synthetic continuum are not perceived as good exemplars of the phonetic categories. Therefore, it seems more valid to scale sensitivity in terms of the amount by which a prototype stimulus (clearly perceived as an exemplar of a phonetic category) must be changed to cause the perception of a different phoneme. This same criterion was employed by Flanagan (1955, p.616), when he concluded that "the DL's for detecting a quality difference in such a (vowel) sound represent a small subset of the DL's for detecting a phonemic difference."

REFERENCES

1. Chistovich, L.A., Lublinskaya, V.V. (1979). The 'center of gravity' effect in vowel spectra and critical distance between the formants: psychoacoustical study of the perception of vowel-like stimuli. Hearing Research, 1, 185-195.
2. Espinoza-Varas, B. (1983). Integration of spectral and temporal cues in discrimination of nonspeech sounds: A psychoacoustic analysis. Journal of the Acoustical Society of America, 74, 1687-1694.
3. Espinoza-Varas, B. and Jamieson, D.G. (1984). Integration of spectral and temporal cues separated in time and frequency. Journal of the Acoustical Society of America, 76, 732-738.
4. Flanagan, J.L. (1955). A difference limen for vowel formant frequency. Journal of the Acoustical Society of America, 27, 613-617.
5. Flanagan, J.L. (1972). Speech analysis, synthesis, and perception. Springer-Verlag, New York.
6. Fujisaki, H. and Kawashima, T. (1970). Some experiments on speech perception and a model for the perceptual mechanism. Annual Report of the Engineering Research Institute (Faculty of Engineering, University of Tokio), 29, 207-214.
7. Garner, W.R. (1962). Uncertainty and structure as psychological concepts. Wiley, New York.
8. Hirsh, I.J. (1974). Temporal order and auditory perception. In: H.R. Moskowitz et al. (Eds). Sensation and Measurement. D. Reidel, Dordrecht-Holland, 251-258.
9. Houtgast, T. (1974). Auditory analysis of vowel-like sounds. Acustica, 31, 320-324.
10. Karpickaya, E.G., Musnikov, V.N., Slepokurova, N.A., and Zhukov, S.Ja. (1975). Auditory processing of steady-state vowels. In: G. Fant and M.A.A. Tatham (Eds). Auditory Analysis and Perception of Speech. Acad. Press, London, 37-53.
11. Kerivan, J.E., Alfonso, P.J., and Espinoza-Varas, B. (1981). Describable acoustic cues in vowel sequence perception. Journal of the Acoustical Society of America, 69, S124.
12. Klatt, D.H. (1975). Voice onset time, frication, and aspiration in word initial consonant clusters. Journal of Speech and Hearing Research, 18, 686-706.
13. Klatt, D.H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. Journal of the Acoustical Society of America, 59, 1208-1221.

14. Lacroix, P.G., Harris, D.J., and Randolph, K.J. (1979). Multiplicative effects on sentence comprehension for combined acoustic distortions. Journal of Speech and Hearing Research, 22, 259-269.
15. Ladefoged, P. and Broadbent, D.E. (1957). Information conveyed by vowels. Journal of the Acoustical Society of America, 29, 98-104.
16. Lisker, L. and Abramson, S.A. (1964). A cross-language study of voicing in initial stops: acoustical measurements. Word, 20, 384-422.
17. Mermelstein, P. (1978). Difference limens for formant frequencies of steady-state and consonant-bound vowels. Journal of the Acoustical Society of America, 63, 572-580.
18. Minifie, F.D., Hixon, and T.J., Williams, F. (1973). Normal aspects of speech, hearing, and language. Prentice-Hall, New Jersey.
19. Peterson, G.A. and Barney, H.L. (1952). Control methods used in a study of the vowels. Journal of the Acoustical Society of America, 24, 175-184.
20. Pisoni, D.B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. Perception and Psychophysics, 13, 253-260.
21. Pisoni, D.B. (1975). Auditory short-term memory and vowel perception. Memory and Cognition, 3, 7-18.
22. Pols, L.C.W., van der Kamp, L.J.Th., and Plomp, R. (1969). Perceptual and physical space of vowel sounds. Journal of the Acoustical Society of America, 46, 458-467.
23. Repp, B.H. (1983). Categorical perception: issues, methods, findings. In: N.J. Lass, (Ed.). Speech and Language: Advances in Theory and Practice, vol. 10, Academic Press, New York.
24. Samuel, A.G. (1977). The effect of discrimination training on speech perception: noncategorical perception. Perception and Psychophysics, 22, 321-330.
25. Soli, S.D. (1983). The role of spectral cues in discrimination of voice onset time differences. Journal of the Acoustical Society of America, 73, 2150-2165.
26. Spiegel, M.F. (1979). The range of spectral integration. Journal of the Acoustical Society of America, 66, 1356-1363.
27. Stelmachowicz, P.G., Jesteadt, W., Gorga, M.P., and Mott, J. (1985). Speech perception ability and psychophysical tuning curves in hearing impaired listeners. Journal of the Acoustical Society of America, 77, 620-627.
28. Stevens, K.N. (1975). Speech perception. In: The Nervous System, Vol. 3, D.B. Tower (Ed.). Raven, New York.
29. Summerfield, Q. and Haggard, M. (1977). On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants. Journal of the Acoustical Society of America, 62, 435-448.
30. Thomas, I.B., Hill, P.B., Carroll, F.S., and Garcia, D. (1970). Temporal order in the perception of vowels. Journal of the Acoustical Society of America, 48, 1010-1013.
31. Turner, C.W. and Van Tasell, D. (1984). Sensorineural hearing loss and the discrimination of vowel-like stimuli. Journal of the Acoustical Society of America, 75, 562-565.
32. Zwicker, E. (1973). Temporal effects in psychoacoustical excitation. In: A.R. Moller (Ed.). Basic Mechanisms in Hearing. Academic Press, New York.

33. Zwicker, E., Terhardt, E., and Paulus, E. (1979). Automatic speech recognition using psychoacoustic models. Journal of the Acoustical Society of America, 65, 487-498.

GENERAL DISCUSSION OF SESSION 1: THE RELEVANCE OF PSYCHOPHYSICS FOR SPEECH PERCEPTION

Chairman: M.E.H. Schouten

Although in the past the contributors to this session held very different views on the relevance of psychophysics for research on speech perception, the most striking aspect of the session was the strong convergence that has taken place in recent years. The main reason for this is probably that the area of auditory psychophysics has become considerably wider than that of "traditional" psychoacoustics, which was concerned almost exclusively with the perception of sinusoids and noise bands. Not only have more complex signals increasingly come into the domain of psychophysics, but more importantly, so has the realization that human perception always involves memory.

Another reason for the convergence of opinion among the speakers at this session is undoubtedly the widely felt tedium at the "tennis game" (Massaro) between those who keep finding new phenomena which are special to speech, and those who then proceed to demonstrate that the same phenomena can also be observed with nonspeech stimuli.

The consensus among the speakers and, indeed, among participants generally, was that speech perception research cannot do without considering the role of learned memory categories. The differences concern the nature of these categories, the question of exactly what information about speech sounds is stored, and how it is coded. There was general agreement that auditory information plays the dominant role in the coding of speech sounds, but Repp, in his paper, also included a representation of information about articulatory manoeuvres. Massaro spoke about his concept of auditory prototypes; the perceptual anchors introduced by Macmillan et al., on the other hand, represent boundaries between categories rather than prototypes. However, Macmillan produced new data about consonant perception, indicating that here, unlike with vowels, the perceptual anchors represent the centres of the categories.

Schouten and Espinoza-Varas argued that categorical perception, as defined in the speech perception literature, is an ideal which is seldom if ever achieved, and owes its appearance to task factors (Espinoza-Varas) and stimulus factors (Schouten). This led to a general discussion, in which it was agreed that categorical perception research has outlived its usefulness: instead of chasing after an experimental artifact, we should simply concentrate on speech perception, recognizing that categories (whether they are phonemes, allophones, or syllables) play an extremely important part in speech perception.

Another set of terms whose usefulness was questioned are "top-down" and "bottom-up" processing; at least in our field of research these terms cause a great deal of confusion. A preference was expressed for the more reductionist terms "peripheral" and "central", although it was realised that (in the words of A. Fourcin) "the efferent system may be capable of modifying the nature of the transformations right down to the bottom". We want to find out first, however, how far we can usefully go without any such complicating assumptions in our models.

Another point of agreement was that speech perception and psychophysics have a very useful meeting ground in the perception of timbre: what is needed is a psychophysics of complex signals.

To sum up, since there was much more agreement than was expected, there was inevitably less discussion. We did agree, however, to drive out a number of terminological ghosts.

Chapter 2

SEPARATION OF ACOUSTIC EVENTS

THE MEANING OF DUPLEX PERCEPTION: SOUNDS AS TRANSPARENT OBJECTS*

Albert S. Bregman

Psychology Department, McGill University,
1205 Docteur Penfield Avenue, Montreal, Quebec, Canada H3A 1B1

In my previous research I have tried to discover how processes of auditory "scene analysis" help the listener to recognize patterns by sorting out those features of the signal that are likely to have arisen from the same source. In this work, I have often found that the scene analysis process seemed to be obeying this rule: each bit of acoustic evidence (such as a particular tone) is to be allocated to one or another perceptual stream, but not to more than one at a time (e.g., Bregman, 1978; Bregman & Rudnick, 1975). The rule can be called "the rule of disjoint allocation". It can be defined with reference to the familiar "vase-faces" ambiguous figure of the Gestalt psychologists. When we are aware of the faces, the line that separates a face from the vase is seen as the boundary of the face. When the vase is seen, that same line "belongs" to the vase. The same piece of sensory evidence (the line) cannot be allocated to both the face and the vase at the same time. Gestalt psychology would refer to this principle as illustrating the "belongingness" of perceived properties. One can see this belongingness principle as the natural result of the process of allocating evidence to distinct environmental objects or events. Unless the mixture of evidence is "parsed" in this way, processes of pattern recognition will attempt to recognize bundles of properties whose co-occurrence is purely accidental, the properties of one object bundled fortuitously with those of another, and it will therefore come up with erroneous descriptions of the objects. Therefore, the rule of disjoint allocation is not merely some incidental property of perception but is central to the attempt to sort evidence out. However, we must now face a phenomenon discovered by Rand (1974) that seems to challenge the belongingness rule, the phenomenon of duplex perception of speech (DPS).

Liberman (1982) has reviewed the research carried out at Haskins Laboratories on this phenomenon. For the purposes of this paper, I will refer to one of the simpler examples of it, the one in which the syllables /da/ and /ga/ are distinguished from one another by the direction of the third formant transition (as in Mann, Madden, Russell & Liberman, 1981). The stimulus pattern is shown in Figure 1 taken from Liberman (1982). In the duplex perception paradigm, the formant pattern is split into two parts, the "base" (shown in the lower

*I want to thank Bruno Repp for reading an earlier version of this paper and making many valuable suggestions. Valter Ciocca made valuable suggestions that led to Figure 3, and Pierre Abdel Ahad's ideas contributed to Figure 2. The preparation of this paper was supported by a Research Fellowship from the Killam Foundation.

left panel) and what we can call the "distinguishing information" (shown in the lower right panel). The base consists of the formant patterns that are the same in the /da/ and /ga/ syllables, and the distinguishing information consists of only the formant transition that distinguishes the two syllables. In the duplex paradigm the base is presented to one ear while the distinguishing information is presented to the other. What the listener reports hearing is this: The whole syllable, /ga/ or /da/ depending on which formant transition has been given as the distinguishing information, is heard at the ear that has received the base. At the ear that has received the formant transition, the listener hears a "chirp" that is like the sound of the transition presented alone.

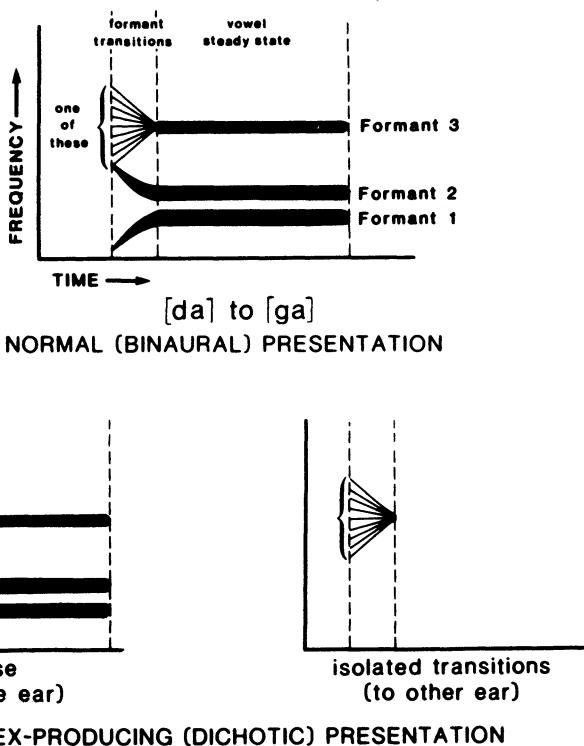


FIGURE 1, Schematic representation of the stimulus patterns used in the experiment on the integration of the time-varying formant transition by Mann, Madden, Russell & Liberman (1981). The figure (from Liberman, 1982) shows (top panel) the formant pattern used to represent /da/, /ga/, or sounds between them, (lower left) the base, and (lower right) the consonant-distinguishing formant transition.

The "duplexity" about the percept is this: it seems that one and the same bit of acoustic information, the formant transition, is being heard simultaneously in two distinct ways. As part of the speech sound it is signalling the distinction between /da/ and /ga/. At the same time it is being heard as a distinct non-verbal sound, a chirp. In other

words, the rule of disjoint allocation has failed. If the rule had been obeyed, the information from the formant transition would have either been segregated from the base and heard as a separate "chirp" stream or else it would have been fused with the base to contribute to the syllable percept. It should not have contributed to both at once.

It appears that the integration of the information from the formant transition with the base to specify a speech sound is really a genuine perceptual integration (Repp, 1984) and not merely a case where information from the chirp has somehow been used to guess which speech sound was the correct answer, as proposed by Nusbaum, Schwab & Sawusch (1983). For example the perception of the /da/ or /ga/ shows all the signs of being "categorical" in nature, showing enhanced discrimination at the phoneme boundary. The perception of /sta/ versus /spa/ shows another property that derives from the categorical property of phoneme perception, namely "trading relations". A pair of cues exhibit "trading relations" when they both signal the same phonetic distinction and they seem to have additive effects so that different combinations of strengths of the two cues yield percepts that are hard to discriminate from one another (e.g. Fitch, Halwes, Erickson & Liberman, 1980). Liberman takes the categorical quality as a distinguishing feature of speech perception as opposed to the perception of non-speech sounds. Research of the Haskins group has shown that the perception of the "speech sound" side of the duplex percept exhibits this property of categorical discrimination whereas the perception of the "chirp" side does not (Liberman, Isenberg, & Rakert, 1981). Liberman argues that this shows that the "speech sound" percept is created by a brain process that is specialized for the analysis of speech sounds and that the "chirp" percept is being dealt with by a different brain process whose function is to handle non-speech sounds. The fact that the two types of percepts can exist simultaneously is taken by Liberman to show that the brain process that underlies the perception of speech sounds is biologically distinct from, and independent of, the processes that handle the perception of non-speech sounds.

Liberman seems to argue that the duplex perception of speech (DPS) poses a great difficulty to those who argue that the perception of speech takes place through the use of a set of general purpose acoustic mechanisms as a front end, and who also insist that it is the same set of mechanisms that deal with speech perception and music perception, as well as the everyday business of getting across the street safely. He implies that if this were true, the segregation of the dichotically presented information by the non-speech mechanisms (as witnessed by the perception of two separate sounds) would have prevented any integration of evidence by the speech mechanism. Duplex perception, for Liberman, shows that speech perception is special and operates by its own laws. The strength of this argument depends on the fact that, as psychologists, we believe so firmly in the rule of disjoint allocation that when it is violated we believe that something very special is happening.

DPS also poses a problem for the scene-analysis view of perceptual grouping. However, the problem arises from concerns other than those of Liberman. The simultaneous activity of different modes of perception, each with its own particular properties does not really pose a problem for the scene analysis view. Obviously there must be

some system that we use for speech that we do not use for other purposes. So must there be for music, or humor, or skiing. We also know that more than one system is typically engaged at the same time. It is clear, for example, that we can use both the speech and the objects-in-space modes at the same time when listening to speech; we can distinguish two conversations, happening on different sides of us, both with respect to their locations and with respect to their verbal contents. The fact that we deal with the spectral differences among phonemes differently than we deal with the spectral differences among chirps is just another example of the fact that there is more than one system and that more than one of these can be used at a time.

Nor need it be problematical that speech-sound judgment is categorical while some other forms of judgment are not. I think it is likely that the categorical mode of perception will be found to be widespread in human perception and will occur whenever a continuous dimension of stimulation is coded into categories. Since the job of the coding process is to decide which category the incoming stimulus falls into, the process will likely register the category identity in the memory of the listener and will disregard the exact value of the stimulus along the dimension apart from its falling or not falling into the range that defines the category. Since for an experienced perceiver the category encoding process will be highly practiced and rapid, when the task of judging whether two stimuli are the same or different is difficult, the listener may rely on the already refined skill of category identification in making judgments of the similarity or difference between two stimuli, and it will be easier to discriminate two stimuli that fall into different categories (see Pisoni, 1973, and Crowder, 1982, for related explanations).

Since phonemes are categories in our linguistic system, the acoustic evidence for their presence will be used categorically. The same thing will hold true of the frequency information that defines the degrees of the diatonic scale (do, re, mi, etc.) or different types of chords in music (e.g., major versus minor). Experimental verification of the categorical perception of degrees of the scale has been obtained by Siegel & Siegel (1976). On the other hand, since the location and the pitch-height of unfamiliar sounds are not dealt with categorically, we will not find categorical properties in these "what-sounds-are-where" judgments. It is unlikely that all speech features are categorical in nature; the degree of emphatic stress probably is not. Furthermore, it is unlikely that categorical judgments are restricted to the modality of hearing. I would be surprised if a person's judgment, based on visual evidence, that an outline was the outline of a woman versus a man were not categorical, displaying such phenomena as trading relations (a formant around the region of the chest for a formant around the region of the hip), and enhanced discrimination of differences at the category boundaries.

The fact that there exist trading relations in speech perception and not in some other kinds of auditory judgments is interesting, but again follows the same pattern as the evidence for categorical perception: it shows that speech perception has particular properties that arise from the nature of speech. The trading property seems to arise from the fact that different forms of acoustic evidence are being combined to arrive at an interpretation of the signal as a sequence of

phonemes. Since each articulatory gesture affects more than one aspect of the speech signal, the perceptual system adds up the evidence to arrive at an interpretation. Whenever you have adding up, you have trading relations. However, the adding up of evidence is not unique to speech perception. It occurs whenever a recognition schema is used. In distinguishing a cursively written "e" from an "l", the width of the loop would trade off with the height of the letter. Furthermore, it also occurs in lower-level auditory phenomena. For example, there is a trading relation between timbre similarity and pitch similarity in controlling the formation of streams, and there is a trading relation (within certain limits) between binaural intensity relations and phase information in determining the spatial location of a source of sound. One of the apparently distinctive properties of the trading (integration) of cues that goes on in speech perception is the fact that the listener has no ready access to the individual cues that are being combined to make a discrimination between phonemes (Fitch, Halwes, Erickson & Liberman, 1980). However, I do not believe that this situation is unique to speech perception. If a person is asked to discriminate a trumpet-like onset from a violin-like onset, I doubt whether he or she could separate the cue of the monotonicity of the growth of loudness from the cue of the relative onset time of different harmonics, although both factors differentiate the two onsets.

It is purely a matter of theoretical taste whether you take the simultaneous use of two systems of judgement as evidence that our brain has a specific biological preparation for each of them. On this criterion, we would conclude that the brain has a specific biological preparation for the perception of humor and the perception of skiing skill, since we can see a skiing clown's action as humorous and as extremely skilled at the same time. It is possible that the brain has all these biologically determined compartments. Out of sheer cowardice, I will retreat and leave this question to brain scientists and philosophers.

What definitely is a serious problem for auditory theory is the apparent exemption of speech recognition from control by auditory scene analysis, in particular from the rule of disjoint allocation. I had always assumed that it was necessary to preattentively organize a mixture of sounds into disjoint streams so that recognition could take place on these already-organized streams. Numerous visual examples exist in which an incorrect parsing of the sensory input will lead to nonsensical results from the recognition process. The existence of duplex perception puts the idea of a stream-forming preprocessor into jeopardy. DPS seems to show that the parsing provided by the scene analysis process is inoperative when speech is involved. At the very least, it seems to be able to be over-ridden by integrative processes that act to group the parts of the speech sound. This is a disturbing claim because it flies in the face of the common observation that it is harder to focus attention on a particular person talking in a crowded room full of talkers if we have one of our ears blocked. This observation is interpretable only if we assume that the binaural information that we can obtain from two ears helps us to sort out the acoustic information from different talkers in exactly the same way as it does for different non-verbal sounds. What then does the phenomenon of duplex perception mean? My general strategy in dealing with this question will be to show that violations of the principle of disjoint allocation are not as rare as we think they are and are not

the mark of two distinct systems of perception operating at the same time.

Fortunately, the failure of the disjoint allocation principle has occurred in other contexts than the "base-plus-chirp" example of DPS. This allows us to examine the range of conditions under which it is found. Many examples have occurred with synthesized speech, and it is these that we shall examine first. Cutting (1976) studied a number of cases in which parts of speech signals were sent to opposite ears of a listener, and despite the fact that the resulting experiences were not all as dramatic as duplex perception, a number of these showed paradoxical percepts. In all of them, speech-sound spectral information was divided between the two ears by the experimenter but perceptually fused by the listener. Yet in the case where the fundamental frequencies of the sounds at the two ears were different, despite the fact that the listeners still fused the phonetic information they reported hearing two different sounds. This may or may not be a case of duplex perception, since it is not clear that the listener identified the phonetic percept with one of the ear-localized sounds and therefore we do not know whether only two auditory entities were heard. However, it is definitely a violation of the rule of disjoint allocation, since the formant energy was both fused to form a phonetic percept and segregated to create the experience of two sounds.

One might be tempted to conclude from these dichotic examples that spatial disparities must necessarily lead to violations of disjoint allocation. However, the earliest experiment on the splitting of formants between the ears reported that when the fundamental of the two formants was supplied by the same generator, so that the same natural pitch contour was followed by both of them, not only was there phonetic fusion, but the sound was heard in one place (Broadbent & Ladafoged, 1957). Apparently, when the cues for fusion are strong enough, a spatial as well as a phonetic fusion can take place. But even this earliest study noticed that when the fundamentals were supplied by different oscillators, either on a different fundamental or even following the same natural F0 contour, but shifting randomly in and out of phase, the two formants would segregate spatially. These observations show that before a clear segregation will take place there must be some additional information favoring segregation (such as different fundamentals or shifting phase relations) to supplement the fact that the two formants are presented to different ears. Apparently, the fact that the dichotic formants start and stop together tends to promote fusion despite the dichotic presentation. This suggests that whenever an experiment finds that a spatial segregation of dichotic information is accompanied by a fusion of the phonetic information, the paradoxical result has arisen from a conflict between acoustic factors favoring spatial fusion and spatial segregation.

The examples that have been presented so far, all involving dichotic presentation, have given the impression that the violation of disjoint allocation can only occur with dichotic presentation. While it is true that dichotic presentation introduces segregative tendencies, the source of segregation need not necessarily come from spatial cues. Darwin (1981) studied the perceptual integration of vowels synthesized from three formants and presented binaurally. He tried to affect the

segregation of the formants from one another by putting them on different fundamentals and by starting them at different times. In some conditions, introducing these changes did not prevent the formants from being phonetically integrated, but did prevent their integration into a single stream, with the result that more than one sound was heard. Darwin's example is one where the auditory system is integrating the phonetic information across the spectrum while at the same time keeping different parts of the spectrum distinct as different sounds.

In the light of these observations, the reader may have decided that the only requirement for inducing a violation of belongingness is a conflict between cues that favor the segregation and the integration of speech sound components. But even this conclusion, focussing as it does on speech sounds, is too restrictive. Violations of belongingness have also been demonstrated with musical tones. While the demonstrations were again based on dichotic presentation, they showed that speech sounds were not required. (Pastore, Schmuckler, Rosenblum & Szczesniak, 1983; Collins, 1985).

How does music fit in with Liberman's idea that DPS arises from the simultaneous activity of biologically distinct brain modules? One might be tempted to argue that music is like speech in having a special processor dedicated to it. After all, musical processing seems to be as specialized as speech processing and there are amusias as well as aphasias. We also generally believe that exceptional musical skills are the product of a special faculty or "gift" that certain people possess and others do not. Therefore it would be consistent with the evidence that has been considered thus far to believe that duplex perception and other violations of the rule of disjoint allocation will be found in those cases (and only those) where a more specialized acoustic processor (such as the one involved in speech or music) is in conflict with a less specialized, lower level one.

Such a conclusion would be tenable if it were not for the fact that such violations have been observed in cases involving pure tones that are not involved in musical patterns. The first such case occurs when two conflicting cues for spatial localization of sounds, one based on intensity and the other on timing, are presented. If there is a large discrepancy between the positions signalled by the two cues, instead of deriving a compromise location, the listener often reports hearing two sounds, one near the position signalled by the intensity cue and the other near the position signalled by the timing cue (e.g., Hafter & Jeffress, 1968). This is a form of duplex perception, in the sense that the same sounds are being used to derive two different images, and it is obviously not occurring as a conflict between a higher-order auditory system and a lower order one.

The second case, discovered at McGill by Steiger (1983), again involved a conflict between two cues for spatial location. The first was a binaural cue: the fact that two identical signals were received in synchrony at the two ears, thus specifying a tone that was in the middle. The second was a context cue: the fact that just before the binaural input, there was a tone presented at only one ear that was very similar in frequency to the binaural tone. The earlier monaural tone apparently set up an expectation to hear another monaural tone similar to it and in the same location, namely at the side of the head.

This expected position conflicted with the location cues set up by the binaural tone.

In Experiment 8 of Steiger's Ph.D. thesis, the stimulus was a rapidly repeating cycle containing a 100 ms monaural pure tone, a 100 ms binaural pure tone, and a 394 ms silence. The monaural tone was either near in frequency to the binaural tone (75 cents) or far from it (2 octaves). If the perception of the sequence had been controlled entirely by binaural cues, the listener would have heard just two tones on each cycle: a lateralized tone derived from the monaural signal which would be heard as alternating with another tone at the midline derived from the binaural signal. There were five listeners experienced in auditory experiments and they were asked to report what they heard on each cycle by marking what they heard on a two-by-five array of blanks on the answer sheet. The five horizontal blanks represented positions from left to right, and the upper and lower rows of blanks represented the first and second time slots in the cycle. The letters H and L, written in any of these blanks, indicated the pitch of the tone, high or low, heard in that position. Thus listeners essentially drew a graphic representation of their percept of each pattern on time and space axes, indicating the pitch and the number of tones they heard.

The results were quite striking. For four out of the five subjects, there were conditions under which three, not two, tones were reported. These four subjects reported three tones on 57 per cent of all trials. Three-quarters of the reports of three tones occurred when the frequency separation between the monaural and the binaural tone was small, suggesting that sequential streaming between the binaural and monaural tones was the factor that was competing with the binaural cues. When reporting three tones, listeners almost always indicated that two of them were heard in succession at the same spatial position and that the third tone was at a different location. This suggests that the auditory system tries to have frequency-based streams occupy a consistent spatial location. Steiger looked to see whether it was always the binaural tone that had split to form two images. He was able to tell which one had supplied the extra tone because the monaural and the binaural tones always had different pitches and the subjects gave a pitch label (H or L) for every reported tone. While it was the binaural tone that was often perceptually split (contributing two tones of the same pitch), the monaural tone could also be split. This latter case is duplex perception indeed: two perceived sounds derived from a single pure tone!

These last two findings of duplex perception, when there was a conflict between different types of cues for spatial location, occurred when there was no involvement of speech, music, or any other system that involves a complex encoding of conceptual elements. We are led therefore to conclude that the existence of duplex perception is not always diagnostic of a situation where a high-level system is in conflict with a low-level one. All that seems to be required is a conflict between cues that normally lead to the computation of consistent locations for streams in pitch or in space.

If all the cases of violation of the rules of belongingness, including the duplex perception of speech, involve a conflict between rules for the formation of auditory streams, what then can we make of

the original claims about the meaning of the duplex perception of speech? While it appears that speech need not be involved in duplex perception it may still be true that the particular conflict in DPS is between a speech processor and a non-speech processor. This possibility makes it necessary to look rather carefully at the DPS example to see whether an alternative explanation can be found. It would be supportive of the view of scene analysis as a preprocessor if it could be shown that even the examples of duplex perception that involved speech did not occur because of the presence of the speech, but due to conflicts entirely within the scene analysis system.

It is evident that in DPS there is a conflict of cues at the purely acoustic level. The cues that tell us that there is a separate stream (heard as a "chirp") at one ear include the fact that there is, within a certain frequency region, an asymmetry of intensity at the two ears, and the fact that when this energy stops abruptly, there is no matching abrupt cessation of energy at the other ear. The cues that tell the auditory system that there is a stream at the ear where the "base" is presented include the fact that there is low-frequency energy on that side that is not matched by energy at that frequency on the other side, and an exact continuation of that energy in the later portion of the base, unaccompanied by any energy at all on the right. These are the cues for segregation.

The cues for integration are the following: First, the energy in the two ears starts at about the same time. Second, the formant transition presented to one ear terminates at a frequency that is exactly the one at which the steady-state formant presented to the other ear begins; therefore there is a continuity cue that says that the formant transition should be connected with the formant steady state. Third, usually the formant transition is synthesized on the same fundamental frequency as the base. This conflict of cues at the purely acoustic level is something that is shared with all the examples of auditory duplex perception that I have discussed.

Given that there is a conflict of cues, why is the percept not "triplex" rather than "duplex"? Let me explain what I mean. If the phonetic process is concerned only with speech-sound identity, not with location, why could it not have simply signalled the existence of a particular phonetic sequence, leaving its location indeterminate? Then, instead of duplex perception, we might have experienced three aspects to the sound: (1) a chirp clearly on one side, (2) the base, heard as a vague, voice-like sound, clearly on the other, and (3) a conviction of the presence of a /d/ of no specific location. The fact that the /d/ sound was perceptually assigned to the side that contained the base makes it appear on the surface that not only was a phonetic identity computed (the /d/), but it was given a location in space. If duplex perception occurs because speech is special, it appears that one of the actions of the special speech-sound analyzer has been to assign the /d/ to a particular location, despite the fact that "spatial location" is a property quite outside those involved in the phonetic system. This seems a strange thing for a phonetic processor to be doing.

I would like to explore an explanation that suggests that DPS occurs as a result of a "description" process, of the type described by Bregman (1977), that is trying to build a consistent description of the

input, and which operates under certain constraints. The constraints that I am proposing are speculative but they can be supported by appeal to certain common experiences. The first step is to try to support the idea that the description-building process does not have a rule that requires a phoneme to always be assigned a spatial positon in the same way that a bicycle is. Not every /a/ has to be given a definite location by the mind. For example, suppose that a choir of baritones, whose members are distributed around the perimeter of a large reverberant hall, are all singing the sound /a/ on the same note. Suppose also that a listener is located in the centre of the hall. He or she will hear the vowel /a/ without being required to or able to assign a location to it. This example shows us that it is not obligatory to assign a location to a phoneme. So why was it done in the case of DPS?

In some circumstances, however, we do hear an /a/ as coming from a definite location. How can we explain this? Perhaps there is an indirect assignment based on a set of constraints that are used by the perceptual process to build sensible descriptions. These might include the following:

- (1) When a phoneme is detected, it is not required to be assigned to a particular location; however, it is required to be associated with a particular "voice".
- (2) A voice, in turn, is not required to have a definite spatial location; however, the "voice" concept is required to be attached to a stream.
- (3) Any stream may get attached to a location, but is not obliged to do so.

Let us also assume that the final description of the duplex perception stimulus, (adopting the language of knowledge network models to represent the perceptual structure) contains, among others, the following components:

- "stream-1 AT left-of-head" (we assume that the base is at the left)
- "stream-1 IS-A voice"
- "voice SAYS /da/"

Let us examine in more detail how this description might be built up. The construction process might go as follows: Let us assume that, as a result of parallel analyses, the pattern analyzers for "voice" and for /da/ have been successful. At this point, /da/ can be attached to "voice" ("voice SAYS /da/") since a requirement of a phonetic percept is that it be said by a voice, and there are no other voices competing for the /da/. The description "voice" gets attached to the left-hand stream ("stream-1 IS-A voice") because the information from this stream has contributed all of the information upon which the description, "voice", was based. The existence of a percept of a chirp on the right shows that the acoustic information on the right has, by itself, yielded a description that contains a location, a pitch and a timbre pattern. (Let us assume that the streams have been assigned to locations for acoustic reasons.) We end up with a description of a voice saying /da/ on the left. To recapitulate, the energy from the chirp and base have been perceptually segregated, yielding two streams, one on the right and the other on the left; the /da/

description needs a voice description to attach itself to and therefore attaches itself to the only one present; the voice description, in turn attaches itself to the stream description on the base side since it is that stream that contains the information that sustains the "voice" description.

The one remaining mystery, and the crux of the whole problem, is why it is that the information from the isolated formant is allowed to be used twice, once for the "chirp" and again for the /da/. This is the heart of the meaning of duplex perception for the scene analysis view. It is here that the rule of "belongingness", or disjoint allocation, is violated. One would think that the information from the right ear should have been given either to the chirp or to the phonetic percept, but not to both. Is there any general principle behind this violation? Is the parallel use of information a special occurrence in audition that signals the activity of a special system, or is it a normal thing that happens in certain auditory situations?

A clue to the puzzle of duplex perception may lie in the fact that sound is transparent. A sound in the foreground does not "occlude" a sound in the background in the same way as a visual object in the foreground occludes our view of objects behind it. There are three important facts about sound: (a) sound is transparent, (b) the head is partly transparent, and (c) sound bounces around in our surroundings. For these reasons, even sources of sound on opposite sides of our heads will have effects on both ears. Therefore the sound at one ear, even sound restricted to a particular frequency region, may bear the imprint, and therefore tell us about, two sources of sound. For this reason, when acoustic information gleaned from one region of the spectrum at one of our ears is used to build the mental description of one source, it should not be made totally unavailable to build the descriptions of other sources whose properties may also have affected that part of the spectrum. In other words, the rule of disjoint allocation of evidence should not always apply.

If sounds are analogous to transparent objects, we should be able to find examples of "double use of evidence" in cases of visual transparency. We have created, in our laboratory, a number of drawings that have this property. We have used, as the analogy to the two ears of the DPS listener, two planes seen at different depths, the nearer plane either looking transparent or else having a hole cut out of it. Different shapes are made to appear to lie on the two surfaces. However, there is an emergent figure that depends on information localized on both planes. To be a strict analogy to duplex perception, when this figure emerges it should appear to lie as a whole on one of the two planes. In our examples, it falls on the further surface.

The first example is a "bah-dah" distinction created by the alignment of one of the stripes of the pattern on the nearer surface with either the left-hand side or the right-hand side of a partial letter drawn on the further surface (See Figure 2). The line that completes the "b" or the "d" both serves that role and also appears as one of the stripes on the surface of the nearer form. The information is duplexed.

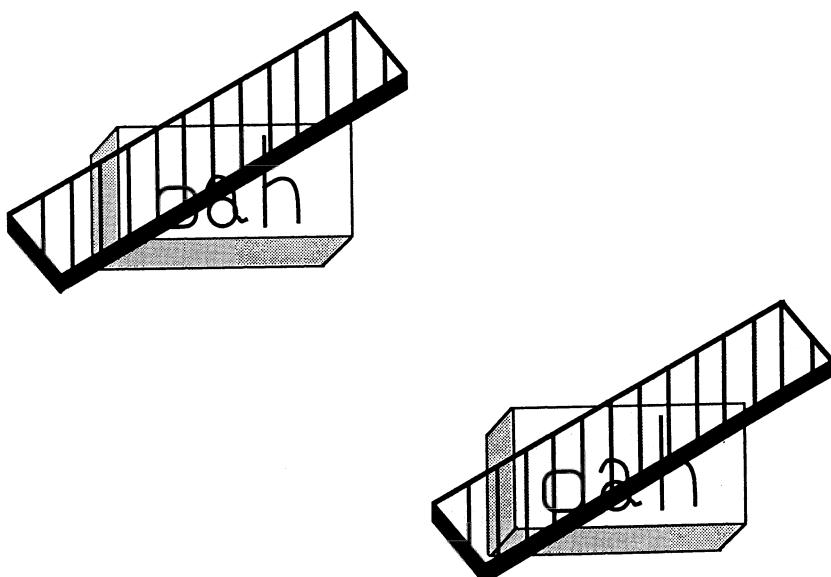


FIGURE 2, A "bah-dah" distinction in the visual domain showing duplex perception of printed words. The vertical stroke of the "b" or "d" also specifies a stripe in the nearer figure.

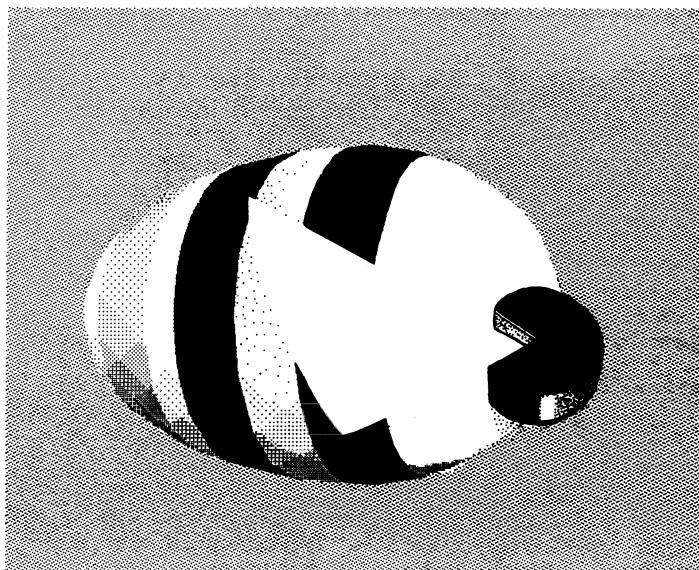


FIGURE 3, Duplex perception of Kanizsa's triangle on the surface of an Easter egg. A wedge-shaped contour specifies a wedge cut out of a nearer form and the corner of a subjective-contour triangle on the egg.

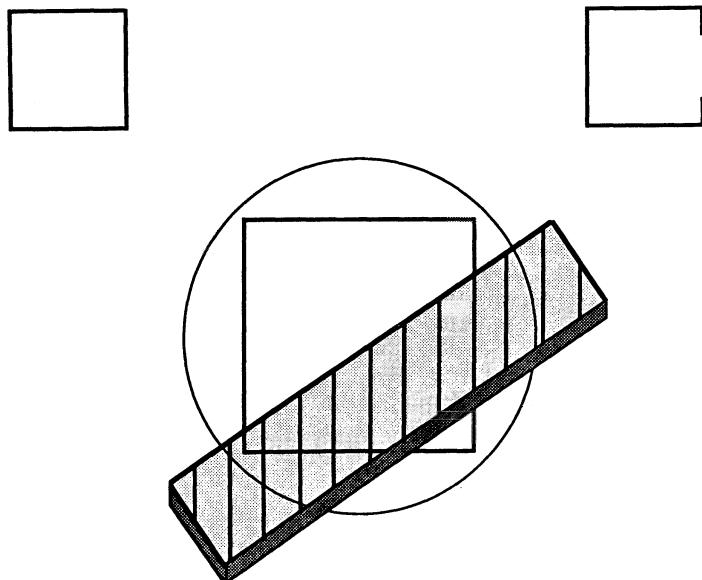


FIGURE 4, Duplex perception of a square, where one of the lines specifies both part of the square and a stripe on the nearer surface. The two figures on the top show alternative ways in which the square can be seen. Most viewers see the complete square.

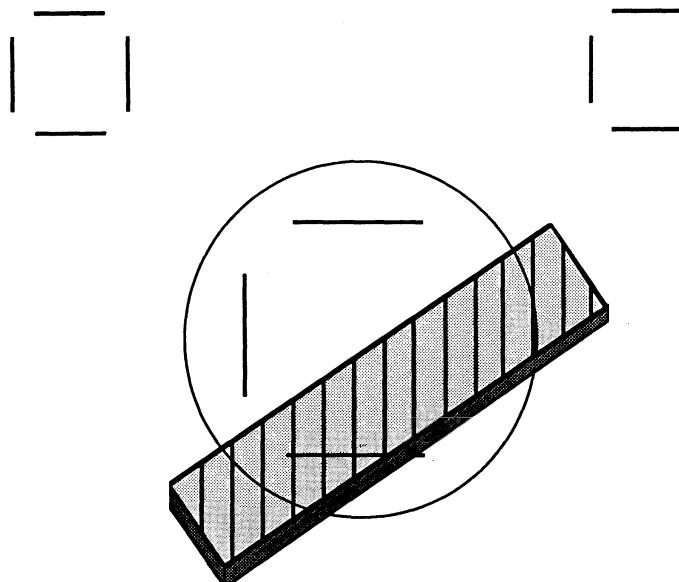


FIGURE 5, Belongingness (disjoint allocation of a shared line) supercedes duplex perception when the line is bound much more strongly into the figure on one surface than into the figure on the other. The two figures on the top show alternative ways in which the fragmented square can be seen. Most viewers see the fragmented "C" rather than the fragmented square.

The second shows a subjective contour of the type described by Kanizsa (1955). In Figure 3 we see a triangle lying on the surface of an Easter egg. The lines that define a wedge cut out of the circular form also serve to define one corner of the white triangle. Again there is a duplexing of information. The third example comes in two parts, Figure 4 and Figure 5. In both figures, a striped surface overlays a surface with a form drawn on it, with one of the stripes exactly replacing one of the lines of the underneath form. In Figure 4 the information is perceptually duplexed, with the square shape seen as a whole underneath. In Figure 5, on the other hand, the parallel stripe organization dominates and therefore the shared information is removed from the underneath form, so that the latter appears as a fragmented "C" rather than a fragmented square. In this case disjoint allocation has occurred. The two examples are similar. Why are the perceptual results so different? I think that the explanation lies in the fact there are strong integrative factors binding the shared line into the shape of the square in Figure 4, while the integrative factors are much weaker in Figure 5.

The last example implies that before information will be duplexed, it must fit very well into both organizations. This is the lesson of the visual illustrations.

Coming back to the auditory realm, we have explained the violation of the rule of disjoint allocation by the fact that sound is transparent. To that we now add the idea that the shared component was strongly associated with analyses made on both sides of the head. Does all this mean that the fact that the /da/ was a speech sound had nothing to do with DPS? Not necessarily. The possibility of transparency means that information is allowed to be duplexed but it does not force it to be. There still has to be a strong integrative tendencies pulling the information in two directions to suggest to our auditory system that it is faced with a case of transparency. Perhaps, contrary to my earlier explanation, the /da/ contributed to one of these forces of integration, because it was a highly familiar pattern. It is possible to suppose that a very highly practiced higher-order schema is allowed to revise the groupings performed by a lower-level process or to compete with them. It may be permitted to either (a) put together evidence that the lower process is trying to segregate, (b) segregate evidence that the lower process is trying to integrate.

There is strong supporting evidence for the "parsing" capabilities of familiar phoneme schemas in the following phenomenon: In a mixture of two vowels, synthesized on the same fundamental frequency (with the same phase), and starting and ending in synchrony, we can often hear both of the vowels, despite the fact that they might share significant energy in a particular frequency region. In this example there is no logical need to segregate any parts of the sound. There is no conflict of acoustic cues. We could simply hear a complex, unfamiliar sound. Furthermore, even if segregation were to occur, if the law of disjoint allocation were followed exactly, then when the information from the overlapping spectral region was given to one of the vowels, it should be unavailable to the other one and we should hear one vowel and some isolated formants. But that is not what happens instead, the common information is duplexed to the two vowel descriptions. If this example is indeed a case of duplexing of information, it shows that duplexing need not arise from the parallel

activity of the speech recognition system and some other system, or between forces entirely within the scene analysis system, but can derive from the parallel activity of two schemas, both of them within the speech system. The duplexing may be allowed because sound is transparent, but it is encouraged by the activation of two phonetic schemas. The possibility that schemas may play a role in the integration of sounds suggests that we should look for a compromise view of scene analysis that allows a role for both general environmental regularities and for highly practiced schemas.

In duplex perception, both visual and auditory, different aspects of the same data were taken as evidence for two separate perceptual "descriptions". How unusual is this? First let us look at vision. We can often find examples in normal viewing situations where a local bit of evidence (found in one region of space) can specify the features of more than one aspect of visual experience. Imagine a vase sitting on a shiny table top. Let us consider the small region below the vase where its reflection is visible. The color of the region tells the viewer about the color of the reflecting table top as well as about the color of the vase reflected in it. The boundary of the vase's reflection has a similar multiplicity of use. Its sharpness tells us how smooth the table top is and its shape contains information about how flat the table is as well as what the shape of the vase is. Beck (1975) has shown that the local information for glossiness that is contained in brightness contrasts will cause us to see the whole table top as glossy; that is, the information is applied to the description of spatial positions other than the one containing the information. The main reason that most of us have trouble drawing a realistic scene is that we are unaware of how much information is packed into local regions of the scene and of how much this evidence is duplexed (used to specify different objects and properties of the scene).

We are not troubled when the same region of light is taken as evidence for both the gloss, colour and shape of a surface. But we are perplexed when the same bit of sound contributes properties to two objects in different spatial positions. Perhaps this is because we are unaware of how often this happens in natural listening situations. Duplex perception is useful; it sometimes helps the auditory system to avoid merging different sounds inappropriately. Suppose we are presented with two sounds, one at our left and the other at our right. In this example, the two sounds have different fundamentals but both are rich in harmonics. Now if the two contain just one partial in common, that is, a partial at the same frequency and intensity, the fact that our two ears are receiving balanced energy at that frequency should, according to classical cues for localization, cause the partial to "pop out" and be heard right in front of us (after all, we can hear more than one sound at a time). If the pitches of the two tones were to move up and down independently, different partials would briefly match and pop out. As phases changed, they would go flying around the room. I have never heard this and neither has anybody I have asked. Instead the offending sounds are duplexed. The matching energy, which, if taken alone, would specify a single sound, is allocated to the two "parent" sounds instead, probably because its frequency fits well into the harmonic series of both of them. This fitting well into two organizations is what we decided was critical in the visual examples of Figures 2 to 5. We can see in the present example that the localization of sounds in space probably does not

depend only on the simple binaural matches that are the subject of classical auditory theory. Due to the fact that we never have really sat down and figured out, in a variety of real-life situations, whether or not the perceived localizations could really be predicted from classical theory, on a frequency-by-frequency basis, we really do not know how often duplexing occurs in real life.

Although it is beyond the scope of this article to pursue this much further, a principle that may force disjoint allocation in some cases is the principle of non-contradiction. One can illustrate this idea with reference to the "vase-faces" ambiguous figure of the Gestalt psychologists. When the boundary line between a face and the vase is seen as part of the face, because we see it as part of the occluding contour of the face (the shape that is visible when the face is the nearer of the two regions) we are concluding that the face region is nearer to us than the other region. If we assign the line to the vase, we are concluding that the vase region occludes the other one and is therefore nearer to the viewer. It would be a contradiction to reach both conclusions at the same time. Said more abstractly, if a line in a drawing portrays the edge of object A where it occludes object B, it cannot at the very same time portray the edge of object B where it occludes object A since the two relations are contradictory. This may be why the shared line cannot "belong" to both the vase and the faces.

It seems that our general conviction, derived from Gestalt psychology, that belongingness must hold true is wrong. How could we have been led into this error? Obviously because there are many cases in which it does hold true. It is therefore necessary that there be a principle that predicts when information will or will not be disjointly allocated. Perhaps that principle is that the information will be duplexed (or multiplexed) if it fits very well into two different descriptions or aspects of description that are being constructed for the current input.

REFERENCES

1. Beck, J. (1975). The perception of surface color. Scientific American, 233 (2), 62-75.
2. Bregman, A.S. (1977). Perception and behavior as compositions of ideals. Cognitive Psychology, 9, 250-292.
3. Bregman, A.S. (1978). Auditory streaming: competition among alternative organizations. Perception and Psychophysics, 23, 391-398.
4. Bregman, A.S. and Rudnicky, A. (1975). Auditory segregation: stream or streams? Journal of Experimental Psychology: Human Perception and Performance, 1, 263-267.
5. Broadbent, D.E. and Ladefoged, P. (1957). On the fusion of sounds reaching different sense organs. Journal of the Acoustical Society of America, 29, 708-710.
6. Collins, S. (1985). Duplex perception with musical stimuli: A further investigation. Perception and Psychophysics, 38, 172-177.
7. Crowder, R.G. (1982). A common basis for auditory sensory storage in perception and immediate memory. Perception and Psychophysics, 31, 477-483.

8. Cutting, J.E. (1976). Auditory and linguistic processes in speech perception: Inferences from six fusions in dichotic listening. Psychological Review, 83, 114-140.
9. Darwin, C.J. (1981). Perceptual grouping of speech components differing in fundamental frequency and onset-time. Quarterly Journal of Experimental Psychology, 33A, 185-207.
10. Fitch, H.L., Halwes, T., Erickson, D.M., and Liberman, A.M. (1980). Perceptual equivalence of two acoustic cues for stop-consonant manner. Perception and Psychophysics, 27, 343-350.
11. Hafter, E.R. and Jeffress, L.A. (1968). Two-image lateralization of tones and clicks. Journal of the Acoustical Society of America, 44, 563-569.
12. Kanizsa, G. (1955). Margini quasi-percettivi in campi con stimulazione omogenea. Revista di Psicologia, 49, 7-30.
13. Liberman, A.M. (1982). On finding that speech is special. American Psychologist, 37, 148-167.
14. Mann, V.A., Madden, J., Russell, J.M., and Liberman, A.M. (1981). Further investigations into the influence of preceding liquids on stop consonant perception. Journal of the Acoustical Society of America, 69, S91. (Abstract).
15. Nusbaum, H.C., Schwab, E.C., and Sawusch, J.R. (1983). The role of "chirp" identification in duplex perception. Perception and Psychophysics, 33, 323-332.
16. Pastore, R.E., Schmuckler, M.A., Rosenbaum, L., and Szczesiul, R. (1983). Duplex perception with musical stimuli. Perception and Psychophysics, 33, 469-474.
17. Pisoni, D.B. (1973). Auditory and phonetic codes in the discrimination of consonants and vowels. Perception and Psychophysics, 13, 253-260.
18. Rand, T.C. (1974). Dichotic release from masking for speech. Journal of the Acoustical Society of America, 55, 678-680(L).
19. Repp, B.H. (1984). Against a role of "chirp" identification in duplex perception. Perception and Psychophysics, 35, 89-93.
20. Siegel, W. and Siegel, J.A. (1976). Categorical perception of tonal intervals: Musicians can't tell sharp from flat. Research Bulletin No. 366, Department of Psychology, University of Western Ontario.
21. Steiger, H. (1983). Influences of sequential organization processes on the use of binaural cues. Unpublished Ph.D. Dissertation, McGill University.

PERCEPTUAL SEPARATION OF SPEECH FROM CONCURRENT SOUNDS*

C.J. Darwin and Roy B. Gardner
Laboratory of Experimental Psychology, University of Sussex,
Brighton BN1 9QG

THE PROBLEM

Hearing has evolved in a hostile environment. Between the sound source and the ear, the different frequencies of a sound are differentially absorbed by the air, reflected by surfaces and mixed in with other sounds. Despite all these distortions and additions, the brain achieves a remarkable constancy of percept, especially in the perception of speech. If we take this ability seriously, we are led to difficult but crucial issues concerning the relationship between phonetic knowledge and the representation of sound in the auditory system. These issues never arise if we only consider the perception of speech as the perception of the sound produced by a single speaker, or synthesiser, heard in a sound-proof room, yet they are crucial in understanding the computational problem faced by hearing, and of great practical significance in constructing robust speech recognition devices or hearing aids. Speech perception - come out of the closet and join the cocktail-party!

Phonetic Knowledge and Auditory Features

Phonetic knowledge is knowledge about a speaker's articulations and sounds. It relates phonetic categories, such as [b] or high front vowel, to acoustic events such as first formant frequency, or silence, or burst spectrum, or fundamental frequency. These acoustic events are properties of the sound from a particular source, and only translate simply into auditory features in the ideal world of the sound-proof room. In the real world the relationship is much more complex, since the sound from the speaker in question is intermingled with that from other sound sources. The most obvious example is silence, which we need to elicit an adequate percept of a stop. When more than one sound source is present the silence of one speaker will be filled with the sounds of others. Yet it is experimentally clear that we can hear perfectly good stops in a signal which has no physical silence in it. If a continuously repeating formant pattern is given a pitch contour that alternates rapidly between a high and a low value, listeners hear it as two voices and take the apparent silence of one voice during the presence of the other as a cue to a stop consonant whose place of articulation depends on the trajectory of the formant pattern into and out of the apparent silence (Darwin and Bethel-Fox, 1977). The waveform reaching the listeners' ears does not have any silence in it,

*Gardner's salary was provided by SERC grants GR/C 8522.1 and GR/C 65930. Research facilities were provided by SERC grant GR/D 6009.9

the silence is perceptually inferred. But how? One question, which is particularly germane to this meeting, is: to what extent such perceptual grouping is based on general auditory principles and to what extent on constraints which are specific to speech sounds (see Darwin, 1981, 1984; Bregman, this volume). In addition, we can ask to what extent properties of the peripheral auditory system contribute to perceptual organisation (cf. Summerfield, this volume).

There are obvious parallels between principles of auditory organisation and Marr's (1982) analysis of the computational problems facing vision. Marr sees visual information as being translated between a set of different levels of representation, each of which makes explicit different properties of the image, and uses different types of constraints. The notion of an edge illustrates the different levels of description. At a low level, local edges consist of changes in luminance over a particular scale. These local edges are then grouped together to give higher-level, extended edges, using constraints that derive from general properties of surfaces and objects. Edges may also be derived from changes in texture or by grouping together collinear features (such as a string of crosses). By making an edge a more abstract concept than a simple change in luminance, vision can achieve greater generality in the recognition of objects. By analogy in hearing we might expect similar benefits from more abstract definitions of auditory properties that were based on the grouping together of elementary properties according to general constraints on sound. The fact that phonetic properties can be heard in stimuli in which formant peaks are signalled by unconventional means (Remez, Rubin, Pisoni, and Carrell, 1981) argues for an abstract representation of the auditory features that make contact with phonetic knowledge.

The challenge to hearing from Marr's work is to understand sufficiently the computational problem posed by hearing. Only then will we be able to identify what representations of the acoustic signal we need, and what constraints could mediate between them.

Auditory Streams and Speech

The experiments of Bregman and his collaborators have demonstrated a number of different acoustic dimensions that can be used by perception to group together the components of a complex sound. The experiments have generally used repeating cycles of simple sounds (a useful summary from a musical perspective is given by McAdams and Bregman, 1979). Properties such as frequency proximity (Bregman and Campbell, 1971) and continuity (Bregman and Dannenbring, 1973), similar onset/offset times (Bregman and Pinker, 1978; Dannenbring and Bregman, 1978), harmonicity and common amplitude or frequency modulation (McAdams, 1984; Bregman, Abramson, Doehring, and Darwin, 1985) encourage the perceptual system to treat different tones as part of a perceptual whole. It is an empirical question, which we have addressed, whether and to what extent similar grouping principles might apply in the perception of the phonetic properties of speech sounds.

We have been able to show that perceptual organisation based on properties similar to those used in Bregman's experiments can have a very marked effect on perceived phonetic categories (Darwin and Sutherland, 1984; Darwin and Gardner, 1986). But equally there have

been a number of occasions where a manipulation that is very effective at changing one measure of perceptual organisation, such as the perceived number of sound sources, has had no detectable effect on phonetic categorisation (e.g. see Darwin, 1981, for pitch differences between formants, and compare Gardner and Darwin, 1986, with McAdams, 1984, for FM effects on harmonics). Such a dissociation between different measures of perceptual organisation gives a possible way to identify different levels of perceptual organisation.

It seems unlikely both on theoretical and experimental grounds that general auditory grouping constraints can determine which sounds form part of the speech of a particular speaker. It seems more likely that low-level grouping mechanisms weakly constrain the groups of sounds that phonetic mechanisms can pool together into a speech percept (Darwin, 1981; Weintraub, 1985; Bregman, this volume). Indeed, in terms of simple sound producing mechanisms, speech can be regarded as consisting of multiple sound sources exciting a system of cavities that can change discontinuously. The excitation source can change abruptly from noise to buzz or be a mixture of buzz and noise in different frequency regions, and the transfer function changes discontinuously as different systems of cavities are coupled into or removed from the system (as in nasals or laterals).

REVIEW OF PERCEPTUAL EXPERIMENTS

This section will review our own and others' experiments on perceptual grouping effects on speech.

Fundamental Frequency

Most voiced speech has a clear harmonic structure. Does the perceptual system exploit harmonic structure in determining which frequency components should contribute to phonetic categorisation?

Broadbent and Ladefoged (1957) played the two formants of a synthetic sentence to different ears of their listeners. When the formants had the same fundamental, a large majority of subjects reported hearing a single voice, but when the formants had different fundamentals, almost all the listeners heard more than one voice. A difference in fundamental between different parts of the spectrum gives rise to a percept of more than one speaker. But curiously it does not necessarily prevent the two contributing parts from being integrated by phonetic decisions. Cutting (1976) showed that subjects' judgement of the place of articulation of stops were unimpaired when the second formant was led to one ear on a different fundamental from the first formant which went to the other ear. Cutting's claim that the speech perception mechanism can integrate information across ears and across fundamentals has been the subject of some debate (cf Nusbaum, Schwab, and Sawusch, 1983; Repp, Milburn, and Ashkenas, 1983) within the original stop-consonant paradigm. It has received some independent support from experiments by Darwin (1981, expts 2 and 3) using two different F1 and two different F2 trajectories that could be combined to give four different diphthong percepts. Although subjects consistently heard more than one sound when different pitches were present, there was very little evidence of subjects tending to hear diphthongs corresponding to formants that shared a common fundamental. However, a subsequent experiment (Darwin, 1981

expt 4) did show a clear effect on phonetic category of grouping by fundamental.

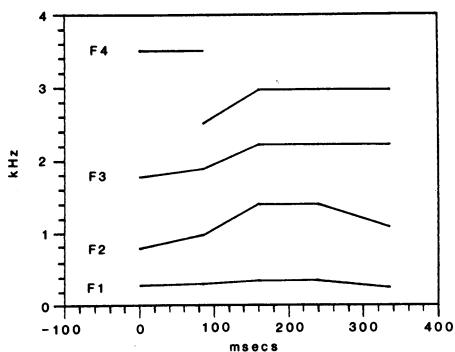


FIGURE 1, Formant tracks for /ru/-/li/ composite syllable.

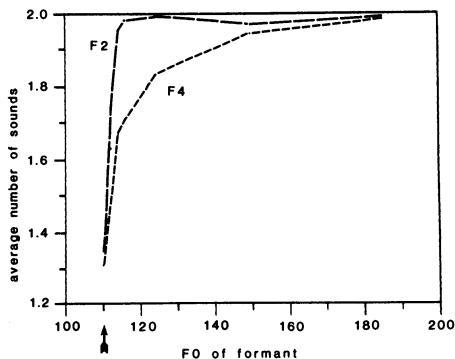


FIGURE 2, Average number of categories reported by 15 subjects when the indicated formant was on the fundamental frequency given on the abscissa, and the remaining formants were on a fundamental of 110 Hz.

maximum of about 2 with a pitch difference of only 4 Hz (Figure 2). When the fourth formant was on a different fundamental the maximum was reached by 14 Hz (about 2 semitones). However, the phonetic segregation of the second formant to give a /li/ response approached its maximum more slowly, asymptoting at 60% with about 2 tones difference. Her results are shown in Figure 3. The left hand panel shows the mean number of times each category was reported when the second formant had a different F0, the right shows the corresponding data when the fourth formant had a different F0. Notice that when the second formant is at 116Hz (a semi-tone above the other formants), the predominant response is /ru/ plus F2, indicating that the second formant is both being heard as a separate sound source and being incorporated into the phonetic percept, a "duplex" percept.

Subjects heard a four-formant speech-like sound (Figure 1). The first three formants taken together are readily perceived as /ru/, while the first, third, and fourth give /li/. The relative number of /ru/ and /li/ responses could be changed radically by altering the fundamental frequency contours of the different formants. When the second formant had a different F0 contour from the others, subjects heard /li/ on 54% of trials, as against 18% of trials when the fourth formant had a different F0 contour. Recently, this result has been replicated and extended by Sally Gaskill in this laboratory. She has examined the extent of grouping by F0 as a function of the pitch difference between either the second or the fourth formants and the remainder. The subjects were trained to identify not only /li/ and /ru/, but also the isolated second and fourth formants. They could respond to the test sounds with either /li/ or /ru/ alone, or in combination with the F2 or F4 category. She found that the total number of categories that subjects reported increased very rapidly as the pitch difference between the second formant and the remainder increased, reaching a

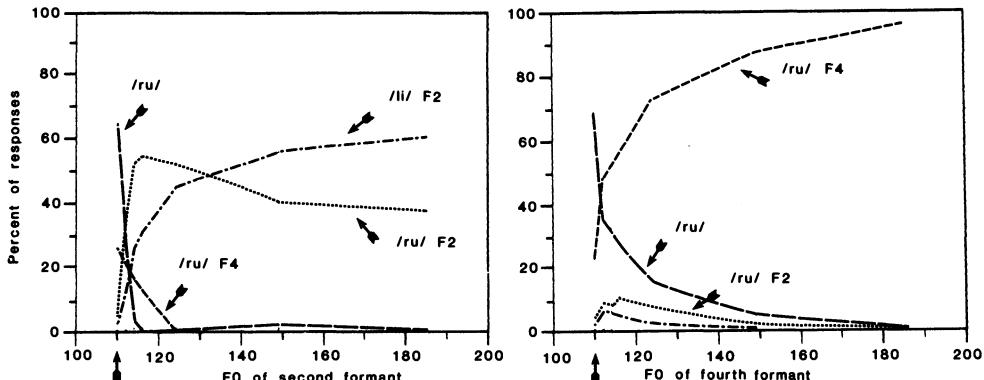


FIGURE 3. The left hand panel shows the mean reports of different categories when the second formant in figure 1 had the fundamental frequency shown on the abscissa. The right-hand panel shows the corresponding data when the fourth formant had a different fundamental.

Using less contrived stimuli, experiments from Nooteboom's group at IPO have addressed similar issues. Scheffers (1983) asked subjects to identify pairs of Dutch vowels (taken from a set of 8) differing in fundamental by between zero and 4 semitones. When the vowels were on the same fundamental, subjects identified about 68% of the vowels correctly. A one-semitone or more difference increased the score to about 79%. Perceptual grouping of different formants by a common fundamental is one possible cause of this 11% increase, although differences in F0 within a particular spectral region may also have made easier the task of identifying the position of the formant peaks. Scheffers' results have been extended, using different F0 contours, by Halikia and Bregman (1984). Using even more natural stimuli, Brokx and Nooteboom (1982) asked subjects to listen to semantically anomalous sentences against a background of speech. The intelligibility of the content words from the sentences was higher (60% correct) when the interfering passage had a different pitch than when it had the same pitch (40% correct) as the sentences.

Pitch perception and vowel colour

Scheffers attempted to model his results on vowel identification using an approach to pitch perception developed at IPO, called the "harmonic sieve" (Duifhuis, Willems, and Sluyter, 1982) and based on an extension of Goldstein's (1973) optimal processor theory. The idea, briefly, is of a sieve, with holes at the harmonic frequencies of a particular fundamental. If the spectral peaks of a periodic sound are fed to the sieve, then its harmonics will drop through. Moreover, if that sound is now mixed with another sound that has a different period, the sieve will block the harmonics of the second sound unless they are sufficiently close to the first sound's harmonics. The multiple pitches of a mixture of sounds can then be obtained by finding, say, the best-fitting harmonic series to those peak frequencies that pass through the sieve. To extend this model to vowel perception we can propose that only those harmonics that emerge through the sieve can contribute to the colour of the perceived vowel.

Moore, Glasberg, and Peters (1984) have compared the performance of a harmonic sieve model with the pitch judgements made by subjects listening to complex tones in which one harmonic has been mistuned by varying amounts. According to the sieve model, a mistuned harmonic should continue to contribute to the pitch of a complex sound as long as it is mistuned by less than the width of the sieve's holes. Moore et al. found that a harmonic continued to make a full contribution to the pitch of the complex provided that it was mistuned by less than about 3%, but that it was progressively excluded from the calculation of pitch as the mistuning increased from 3% to 8%. By 8% the mistuned harmonic made no contribution to pitch.

If the perception of vowel colour also depends on the harmonics that pass through the harmonic sieve (at least in the first formant region, where they are well-resolved by the cochlea), then we should find that, as a harmonic is mistuned, a vowel's colour changes as if the harmonic were being physically removed.

We have recently tested this prediction (Darwin and Gardner, 1986) by mistuning the 500 Hz and 375 Hz components of vowels with a fundamental of 125 Hz differing in F1 along an /I/-/e/ continuum. The phoneme boundary for the original continuum lies at around 450 Hz, so physically removing or attenuating the 500 Hz component makes the vowel sound more /I/-like and moves the phoneme boundary to a higher nominal F1 value. Similarly, physically removing or attenuating the 375 Hz component gives a more /e/-like vowel and moves the boundary to a lower nominal F1 value. If mistuning a harmonic causes it to be perceptually removed from the computation of vowel colour according to the harmonic sieve model we can make the following predictions:

First, large amounts of mistuning (c. 8%) should give phoneme boundaries similar to those found when the mistuned component is physically removed;

Second, small amounts of mistuning (c. 3%) should have no effect on the phoneme boundary;

Third, intermediate amounts (from 3% to 8%) should give intermediate boundary shifts.

The results of the experiment fulfilled the first prediction. Large amounts of mistuning, around 8%, gave large shifts in the phoneme boundary that in some instance were as large as those found when the component in question was physically removed. However, with small amounts of mistuning there were significant, although smaller, shifts in the phoneme boundary, which turned out to be due to the phase changes introduced by slight mistuning. The effect of phase changes on vowel quality has interesting neuro-physiological correlates that are the subject of a companion paper (Palmer, Winter, Gardner, and Darwin, this volume). Related perceptual results are also reported by Traunmüller (this volume).

Summary of fundamental frequency effects

(i) Under suitable conditions, there are clear perceptual grouping effects on phonetic categorisation due to a common fundamental frequency. They have been found for perceptual grouping both across formants (Darwin, 1981, expt 4) and within a formant

(Darwin and Gardner, 1986). In addition, it is likely that the increase in intelligibility produced by putting two competing speech signals on different fundamentals is also due to some type of perceptual grouping (Brokx and Nooteboom 1982; Halikia and Bregman, 1984; Scheffers, 1983).

(ii) Grouping by common harmonic structure within a formant region behaves in a similar way for the perception of vowel quality as for the perception of pitch (Moore et al., 1984; Darwin and Gardner, 1986).

(iii) A different fundamental does not preclude two formants from contributing to the same phonetic category. A subject may be aware of more than one sound source, but still hear a phonetic category that integrates information from the two apparent sources (Darwin, 1981; Gaskill, pers. comm.). Such an ability is necessary in natural speech where vocal tract excitation can be noisy in the whole or part of the spectrum.

Frequency Modulation

Apart from grouping effects due to instantaneous fundamental frequency, there is also some evidence that grouping may be based on dynamic, frequency modulation properties. McAdams (1984) has found that if one harmonic of a 16 partial complex is given frequency modulation (in the form of vibrato and jitter) that is incoherent with the remaining 15 partials, then subjects hear more than one sound source. But McAdams failed to find any effect on the perceptual prominence of one vowel presented with two others when its fundamental was modulated with different vibrato and jitter from that of the other two vowels.

A similar lack of effect of frequency modulation differences on phonetic categorisation has been found by us (Gardner and Darwin, 1986), using the paradigm described in the previous section. We looked for phoneme boundary shifts along the /I/-/e/ continuum when the 500 Hz harmonic was frequency modulated at a different rate or phase from the other harmonics of the vowel. The depth (2%) and frequencies (6 and 10Hz) of modulation were similar to those used by McAdams. The depth was not sufficient to give any perceptual grouping due to instantaneous mistuning. Although subjects could hear the additional sound source produced by an incoherently modulated harmonic, we failed to find any phoneme boundary changes due to grouping by frequency modulation.

In a non-speech context, Bregman and Doehring (1984) have examined the effect of different directions of linear frequency modulation on the grouping of simultaneous components. Their results do not separate effects of frequency modulation as such from concomitant effects of instantaneous mistuning.

In summary, there is no clear evidence that the dynamic aspects of frequency modulation influence perceptual grouping for the formation of phonetic categories.

Onset- And Offset-time Differences

Two experiments from Bregman's group drew attention to the importance of onset- and offset-time as a perceptual grouping heuristic. A tone that is not exactly synchronous with another tone is more likely to be captured by a competing auditory stream than when it is exactly synchronous (Bregman and Pinker, 1978). This conclusion was refined by Dannenbring and Bregman (1978), who showed that for asynchronies to help a tone to be removed from a complex, that tone had either to start before the remainder, or finish after them - conditions that make that tone more audible (Rasch, 1978).

The important question for speech perception, though, is what happens to the timbre or vowel colour of a complex, after a component tone has either been heard out, or become part of another perceptual stream. As we saw in the section on fundamental frequency, an integrated phonetic percept can arise from two distinct fundamental frequencies. It is conceivable that vowel quality could be left unimpaired after one of the component frequencies has been heard as a separate sound. Analogous effects are encountered with the phenomenon of "duplex" perception (Liberman, Isenberg, and Rakert, 1981), when part of one formant is led to a different ear and is both heard as a separate sound and incorporated into the phonetic percept.

It is clear that the vowel quality is changed when a component tone is perceptually separated out by onset or offset-time differences (Darwin, 1984; Darwin and Sutherland, 1984). These experiments used a similar paradigm to that described above in the sections on fundamental frequency. An F1 continuum is set up between /i/ and /e/, and the phoneme boundary located. Then energy is added to, say, the 500 Hz harmonic of the vowel resulting in a shift of the phoneme boundary. If this added energy is then made to start either a few tens of milliseconds before the vowel or to continue a little after it, it is heard out as a separate tone and the phoneme boundary reverts back to its original value.

Interpretation of onset-time effects are complicated by the possibility that they may in part be due to adaptation. Adaptation has been used as an explanation of a similar effect reported by Summerfield and his colleagues (Summerfield and Assmann, this volume; Summerfield et al. 1984). Here, a spectrally flat noise acquires a vowel-like quality when it is preceded by a noise that has troughs in place of a vowel's formant peaks. It is unlikely that all the effects of tone onset-/offset-time on vowel quality can be attributed to such an adaptation effect. First, the effect occurs with offsets as well as with onsets. Second, the tendency for a leading tone to influence vowel colour is reduced when the leading part of the tone forms a new perceptual group with another tone that also precedes the vowel but which stops as the vowel starts (Darwin and Sutherland, 1984). Perceptual grouping must be playing some role.

Although onset- and offset-times clearly can play a role in perceptual grouping for phonetic categorisation, sounds that start at different times can clearly still be integrated into a phonetic percept (e.g. F1 cutback in aspirated stops), and sounds that start and stop at the same time (and are harmonically related) may not necessarily be integrated. An example of the latter effect is given in Darwin (1984,

expt 2) where varying amounts of extra energy are added to first formant harmonics. Small amounts of added energy are incorporated into the vowel percept and change the phoneme boundary between /ɪ/ and /e/, but large amounts are discounted and heard as a separate tone. One explanation for this effect is that the speech perceptual system knows what are plausible shapes for formant envelopes and takes what it needs from the input. A similar point, at a higher level of analysis is made by the demonstration described by Liberman and Studdert-Kennedy (1978) that additional formants added to speech can be perceptually separated even when they are on the same fundamental and have the same onset- and offset-times as the speech.

Non-local Effects

So far we have only considered perceptual grouping of simultaneous (or almost simultaneous) frequency components of speech sounds on the basis of local properties such as harmonic structure, frequency modulation and onset-time. Can we also demonstrate perceptual separation of simultaneous components of speech on the basis of non-local properties? The following experiment demonstrates grouping effects based on a preceding and following sequence of tones. It was resuscitated by conversations with Richard Hammersley at the 1985 BASICS meeting in Banff.

The experiment is based on the same paradigm as our previous experiments, using a change in the phoneme boundary along an F1 continuum of 56 ms duration, 125 Hz fundamental vowels to investigate the extent to which added energy (at 500 Hz) is incorporated into the perception of vowel quality. The added energy at 500 Hz could, however, be interpreted as a tone that forms part of a sequence of 1,2,4, or 8 steady, ascending or descending tones preceding the vowel and of 2 tones continuing after the vowel (one set of steady conditions had no tones coming after the vowel). The configurations of the various stimuli are shown in Figure 4. The ascending and descending tones were spaced either at 62.5 Hz intervals (half-harmonic) around 500 Hz or at whole tone intervals around 500 Hz (396.9, 445.4, 561.2, 630.0 Hz, etc). The 56 ms tones were presented at a rate of 10/s.

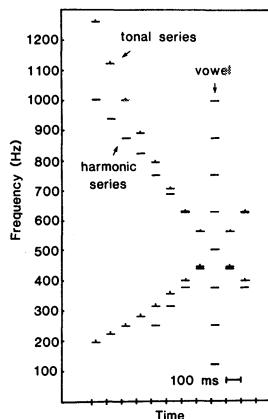


FIGURE 4, Stimulus configurations used to examine the effect on vowel quality of embedding a vowel in a musical or harmonic series of tones.

The results of the experiment are shown in Figure 5. When 6dB of energy at 500 Hz is added to the original continuum, the phoneme boundary moves to a lower F1 value from about 480 Hz to about 460 Hz. When sounds of this new continuum are preceded by a steady sequence of tones at 500 Hz, there is a significant shift back towards the original phoneme boundary (diamond symbols joined by dashed line). The shift is significant with a single preceding tone, and asymptotes with 4 preceding tones. The shift is greater when there are two tones following the vowel (diamonds with solid line), than when there are not, a difference that could be due to perceptual grouping but not to adaptation. With the ascending and descending tone sequences, we found no difference between the half-harmonic and the tonal series, but there were marked differences between the ascending and descending series. When the vowel is preceded and followed by a descending series of tones, the phoneme boundary shifts (open squares and triangles) are very similar to those found for the steady tones. On the other hand when the vowel is preceded and followed by ascending series of tones, there is a small upwards shift with one preceding tone, and then no evidence of any shift at all. Although it is not at all clear what causes the difference between the ascending and descending tone sequences, the movement of the phoneme boundary in the steady and descending tonal conditions gives clear evidence for some perceptual grouping mechanisms removing part of the 500 Hz component from the vowel on the basis of non-local previous and succeeding events.

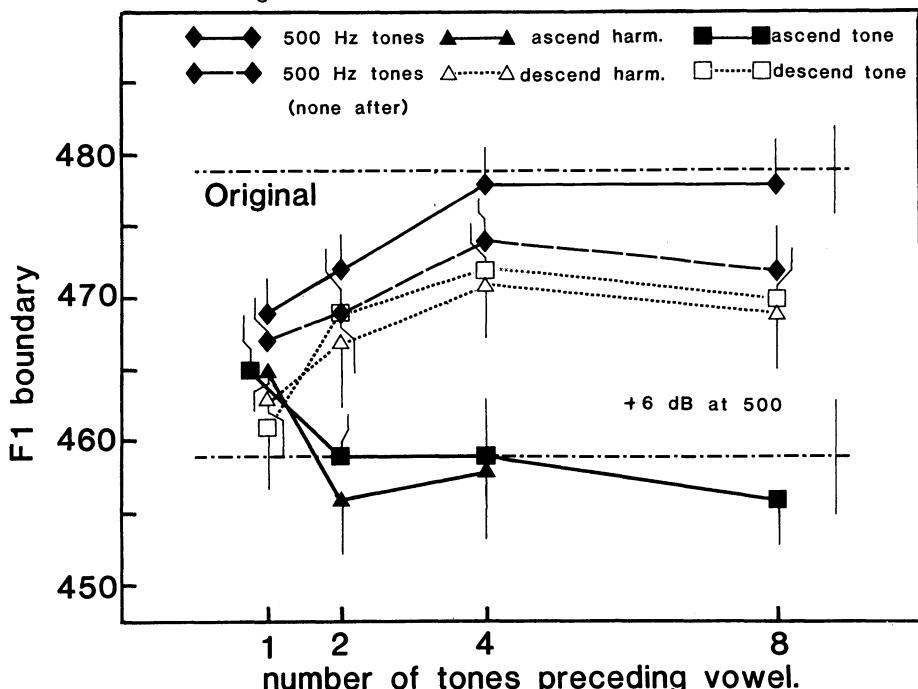


FIGURE 5. Phoneme boundary between /i/ and /e/ for a vowel placed in a context that causes additional energy at one tone of the vowel's harmonics to be treated as a part of a sequence of tones.

AUDITORY REPRESENTATIONS

We return now to the problem raised at the beginning of this paper: how to characterise in computational terms processes and representations that mediate between the acoustic input and phonetic knowledge.

The experiments reviewed above have demonstrated quite clearly that phonetic properties cannot be matched directly against a simple representation (such as a spectral template or even a description of the formant peaks) of the input sound. Parts of the sound must be selectively removable and new abstract properties must be derivable from the selective representation. Since the parts that can be discounted are local in frequency and time, the initial representation must be based on local properties. This representation must also make explicit those properties which can be used for perceptual organisation and must also allow new properties to emerge after re-organisation has occurred. Weintraub's computer model of sound separation allows such re-organisation to occur on the basis of common peaks in a function similar to the autocorrelation function, but it is likely that other simple auditory dimensions involving temporal relations and continuity will also be necessary (see Weintraub, this volume). Groupings made on the basis of general auditory principles should be able to be over-ridden by speech-specific mechanisms, and (as Weintraub implements) speech-specific mechanisms should be able to group together objects that have been identified as separate by auditory processes.

REFERENCES

1. Bregman, A.S., Abramson, J., Doebring, P., and Darwin, C.J. (1985). Spectral integration based on common amplitude modulation. *Perception and Psychophysics*, 37, 483-493.
2. Bregman, A.S. and Dannenbring, G.L. (1973). The effect of continuity on auditory stream segregation. *Perception and Psychophysics*, 13, 308-312.
3. Bregman, A.S. and Doebring, P. (1984). Fusion of simultaneous tonal glides: the role of parallelness and simple frequency relations. *Perception and Psychophysics*, 36, 251-256.
4. Bregman, A.S. and Pinker, S. (1978). Auditory streaming and the building of timbre, *Canadian Journal of Psychology*, 32, 19-31.
5. Broadbent, D.E. and Ladefoged, P. (1957). On the fusion of sounds reaching different sense organs. *Journal of the Acoustical Society of America*, 29, 708-710.
6. Brokx, J.P.L. and Nooteboom, S.G. (1982). Intonation and the perceptual separation of simultaneous voices. *Journal of Phonetics*, 10, 23-36.
7. Cutting, J.E. (1976). Auditory and linguistic processes in speech perception: inferences from six fusions in dichotic listening. *Psychological Review*, 83, 114-140.
8. Dannenbring, G.L. and Bregman A.S. (1978). Streaming vs fusion of sinusoidal components of complex tones. *Perception and Psychophysics*, 24, 369-376.
9. Darwin, C.J. (1981). Perceptual grouping of speech components differing in fundamental frequency and onset time. *Quarterly Journal of Experimental Psychology*, 33A, 185-207.

10. Darwin, C.J. (1984). Perceiving vowels in the presence of another sound: constraints on formant perception. Journal of the Acoustical Society of America, 76, 1636-1647.
11. Darwin, C.J. and Bethell-Fox, C.E. (1977). Pitch continuity and speech source attribution. Journal of Experimental Psychology: Human Perception and Performance, 3, 665-672.
12. Darwin, C.J. and Gardner, R.B. (1986). Mistuning a harmonic of a vowel: grouping and phase effects on vowel quality. Journal of the Acoustical Society of America, 79, 838-845.
13. Darwin, C.J. and Sutherland, N.S. (1984). Grouping frequency components of vowels: when is a harmonic not a harmonic? Quarterly Journal of Experimental Psychology, 36A, 193-208.
14. Duifhuis, H., Willems, L.F., and Sluyter, R.J. (1982). Measurement of pitch in speech: An implementation of Goldstein's theory of pitch perception. Journal of the Acoustical Society of America, 71, 1568- 1580.
15. Gardner, R.B. and Darwin, C.J. (1986). Grouping of vowel harmonics by frequency modulation: absence of effects on phonemic categorisation. Perception and Psychophysics, 40, 183-187.
16. Goldstein, J.L. (1973). An optimum processor theory for the central formation of pitch, Journal of the Acoustical Society of America, 54, 1496-1516.
17. Halikia, M.H. and Bregman, A.S. (1984). Perceptual separation of simultaneous vowels presented as steady states and as parallel and crossing glides. Journal of the Acoustical Society of America, 75, S1, S83.
18. Liberman, A.M., Isenberg, D., and Rakert, B. (1981). Duplex perception of cues for stop consonants. Perception and Psychophysics, 30, 133-143.
19. Liberman, A.M. and Studdert-Kennedy, M.G. (1978). Phonetic Perception. In R. Held, H. Leibowitz and H.L. Teuber (Eds). Handbook of Sensory Physiology VIII: Perception. Heidelberg: Springer-Verlag.
20. Marr, D. (1982). Vision. San Francisco: Freeman
21. McAdams, S. and Bregman, A.S. (1979). Hearing Musical Streams. Comp. Mus. J., 3(4), 26-43,60.
22. McAdams, S. (1984). Spectral fusion, spectral parsing and the formation of auditory images. Unpublished Ph.D. dissertation, Stanford University.
23. Moore, B.C.J., Glasberg, B.R., and Peters, R.W. (1985). Relative dominance of individual partials in determining the pitch of complex tones. Journal of the Acoustical Society of America, 77, 1853-1860.
24. Nusbaum, H.C., Schwab, E.C., and Sawusch, J.R. (1983). The role of 'chirp' identification in duplex perception. Perception and Psychophysics, 33, 323-332.
25. Rasch, R.A. (1978). Perception of simultaneous notes such as in polyphonic music. Acustica, 40, 21-33.
26. Remez, R.E., Rubin, P.E., Pisoni,C.B., and Carrell, T.D. (1981). Speech perception without traditional speech cues. Science, 212, 947- 950.
27. Repp, B.H., Milburn, C., and Ashkenas, J. (1983). Duplex perception: confirmation of fusion. Perception and Psychophysics, 33, 333-337.

28. Scheffers, M.T.M. (1983). Sifting vowels: auditory pitch analysis and sound segregation. Doctoral dissertation, Groningen University.
29. Summerfield, A.Q., Haggard, M.P., Foster, J., and Gray, S. (1984). Perceiving vowels from uniform spectra: phonetic exploration of an auditory after-effect. Perception and Psychophysics, 35, 203-213.
30. Weintraub, M. (1985). A theory and computational model of auditory monaural sound separation. Ph.D. thesis, Stanford University.

SOUND SEPARATION AND AUDITORY PERCEPTUAL ORGANIZATION

Mitchel Weintraub

SRI International, Sensory Sciences Research Laboratory,
333 Ravenswood Avenue, Menlo Park, CA. USA 94025

This paper describes a computer model that attempts to separate two simultaneous talkers. Using such a model, a speech recognition system can improve its ability to recognize separate utterances made simultaneously by two different talkers. In order to achieve this goal, the computer model's algorithms were designed to functionally simulate the major steps that are performed by the human auditory system when processing multiple sound sources monaurally. It is based on my belief that the human auditory system computes similar representations as the computer model, even if the physiological processes and representations that the auditory system uses differ slightly from those presented here.

The computer model's primary source of information for separating the two simultaneous talkers is periodicity information. Although the psychoacoustic literature has discussed many other information sources that the auditory system can use for separating sounds (Bregman, 1984, Darwin, this volume, Weintraub, 1985), computer simulations to date have primarily used periodicity information to separate sounds. I plan to add additional acoustic information (e.g. pitch dynamics, common amplitude modulation characteristics between different frequency regions) that would make the current computational model more realistic and possibly improve its performance.

This paper focuses on the need for an intermediate representation in the separation of simultaneous sounds. It outlines an approach where there are three different levels of acoustic representation. The computational mechanisms that determine the characteristics of each level are also reviewed. Some conceptual issues regarding levels of representation in the auditory system will also be discussed. These issues have arisen out of my computer modeling work, and a resolution of them will hopefully allow the model's mechanisms to accurately reflect the actual mechanisms that are being used by the auditory system. It is my aim to make predictions with this model that can be verified or contradicted by comparing the results with psychoacoustic experiments.

1. OVERVIEW OF AUDITORY PERCEPTUAL ORGANIZATION

The three basic questions motivating this research are:

1. What information is used by the auditory system to separate sounds?

2. What transformations and representations are used by the peripheral auditory system in order to derive this information from the incoming sound?

3. How is this information used by the auditory system to separate and interpret the sounds that it hears?

Regarding the first question, psychoacoustic experiments have shown that the auditory system uses many different types of information for the separation of the sounds that it hears. The monaural information cues which have been discussed in the literature are: periodicity, periodicity dynamics, the onset and offset of sounds, spectral continuity, local amplitude modulation fluctuations, visual information (e.g., lip-reading cues), and linguistic information (phonetic and word transitional probabilities, phrasal and message content), (Weintraub 1985). The computer model presented in this paper tries to answer the remaining two questions of how the auditory system computes these information cues (especially periodicity cues) and how this information is used to separate simultaneous sounds.

In some instances, a relaxed conversation between two people will not consist of two simultaneous voices. Instead, the two people will take turns talking to each other. When listening to this type of conversation, the original goal of a sound separation system (computing a spectral estimate of each of the two talkers' voices) was to determine which person is currently speaking, and to assign the observed acoustic energy to this talker's internal representation. When viewed from this perspective, the objective of the computer model is to assign local regions of the acoustic input to the correct talker.

The computational model for separating sounds is based on the following principle: The information computed by the auditory system (e.g. periodicity, amplitude modulation, ...) is a function of both frequency and time, and is a property of a local acoustic frequency-time region (abbreviated as LAFTR). A local acoustic frequency-time region (LAFTR) is assigned to a sound source based on the consistency (or goodness of fit) between the local properties of this region and the properties of the sound source.

By describing a sound as consisting of many different LAFTR's, a system can analyze the incoming sound by finding relationships between the different regions. Similarities in the features of the different regions imply a relationship between them and allow them to be interpreted as coming from the same sound source.

Research on selective attention has dealt primarily with how a person focuses on one sound in the presence of other sounds. Some theories of attention are characterized by the 'early filtering' models (filtering here refers to the separation of one sound from other sounds) of Broadbent (1957) and of Treisman (1960). A listener filters or separates by focusing on different functional channels (e.g., an internal channel that represents the location of the desired sound source, pitch channel, etc.). In other theories of attention, such as the model of Deutsch and Deutsch (1963), sound separation does not occur until late in the processing (until at least the semantic level is reached).

Focusing attention on the output of a single functional channel cannot account for how the human auditory system monaurally separates simultaneous sounds. Since one can view speech as consisting of periodic segments, nonperiodic segments, bursts, and periods of silence, the speech of a single talker cannot be monaurally analyzed along any single dimension. It is therefore not possible for the auditory system to separate sounds by focusing its attention on a single functional channel.

Suppose that one is listening to a male and a female voice. The male voice says the digit "three", and the female voice then says the digit "seven". The waveforms for these two words can be spliced together digitally so that the word "seven" starts as soon as the word "three" ends. When this sequence is played to a listener, he hears a male voice saying "three", followed immediately by a female voice saying "seven". However, if the "even" part of the digit "seven" is deleted, what is left is the male talker's "three", followed by the /s/ of the female's "seven". When people listen to this waveform they will hear the word "threes". Since the voiced part of the female's speech is missing, the auditory system interprets the /s/ as belonging to the male voice. It is only after the listener hears the voiced part of the female's digit seven that the /s/ is correctly interpreted as originating with the female talker.

Two information cues aid the auditory system in determining which speaker is responsible for saying the /s/. The presence of smooth spectral transitions between the /s/ and the surrounding voiced regions is one cue that can help the auditory system determine which speaker said the /s/. Another source is linguistic constraints. Knowledge about phonemes and phonetic transitions can be helpful for determining which speaker produced the fricated segment. Linguistically, the interpretation of the /s/ as forming a part of the digit "seven" is a better explanation of the incoming sound than assigning the /s/ to the digit "three".

The computer model describes the auditory system's interpretation as follows:

1. When the auditory system hears the male vowel /iy/ from the digit "three", all the local acoustic frequency-time regions (LAFTR) are periodic, and have the same period. The auditory system knows that these LAFTR's belong together because of their acoustic similarity. After the vowel /iy/, the auditory system encounters an onset of fricated energy, followed by a group of LAFTR's that have no periodic structure to them. These nonperiodic frequency-time regions are also grouped together based on their acoustic similarity: they are all nonperiodic. However, the LAFTR's of the vowel are distinguished from the LAFTR's of the /s/ since the former are periodic and the latter are aperiodic. When the periodic segment "even" from the female digit "seven" is heard, the LAFTR's of this region are also grouped together based on the similarity of their periodic properties.
2. The acoustic input in this example is treated as consisting of three different groups, with each group-object (abbreviated as G-O) consisting of many LAFTR's with similar periodic properties. These group-objects are then assigned to the different sound streams based on spectral continuity constraints as well as on linguistic constraints

(phonetic transitional probabilities, word transition probabilities, phrasal and message content).

In the above example, LAFTR's with similar periodicity properties are grouped together into an intermediate representation called a group object (G-O). The reason for the intermediate level of G-O is that LAFTR's are not assigned to sound streams independently, but act as a cohesive unit. In some respects, the G-O is a conceptual abstraction, since it is possible to devise computational algorithms that achieve the same sound separation results without ever explicitly having any intermediate representation.

The presence of an intermediate representation can also be seen in Bregman's streaming experiments. Here, collections of pure tones are assigned by the auditory system to either one sound stream or another. A tone is never split where one part of the tone is assigned to one sound stream and another part is assigned to a different sound stream. Each tone behaves as a cohesive 'unit' or 'group' where all its events (which result from the response of the cochlear model to that tone) are assigned to the same sound stream.

A G-O can be reassigned to a different sound stream, thus reassigning all the LAFTR's in the G-O. In the 'three-seven' example discussed earlier, all the LAFTR's that compose the frication sound /s/ were initially assigned as a unit to one sound stream, and then later (when the second vowel was analyzed) to a different sound stream.

This author hypothesizes that there are three different levels of the auditory system's representation of sounds at different points in the separation processing. These levels are called the 'Local Acoustic Frequency-Time Region', the 'Group Object', and the 'Sound Stream', defined as:

LAFTR: A local acoustic frequency-time region consists of all the neuron firings in a local region of acoustic space. In the current cochlear model used in this simulation, there are 85 frequency channels (with overlapping responses). In each frequency region, all the neural events in a 10 ms interval make up the LAFTR.

Group Object: A group object is a collection (across both frequency and time) of LAFTR's having similar properties, that are perceived (by certain similarity measures: e.g. common periodicity) as a unit. It is an intermediate level in the representation of sounds and corresponds to a segmentation of the incoming sound into regions (over time) that have similar properties.

Sound Stream: A sound stream is an internal representation of a particular sound source, which consists of a temporal succession of group objects. Group objects that are assigned to a given sound stream are thought to originate from the same sound source.

Various LAFTR's are grouped together into the same G-O based on the degree of similarity between their acoustic features (e.g. periodicity, periodicity dynamics, onset, offset, and amplitude modulation). Currently, the only acoustic feature that is used to group LAFTR's together is periodicity. Group objects are assigned to sound streams as follows:

If there are two group objects that overlap in time, they must be assigned to different sound sources. If it can be determined that one of the group objects belongs to one sound stream, then the other group object must be assigned to the other sound stream. Periodic G-O's are assigned to a sound stream based on the compatibility between the pitch contour of the G-O and the average pitch of the talker.

There is additional information which I believe the human auditory system uses in the assignment of group objects to sound streams, namely:

1. If two group objects are sequential in time (with one group object starting after the other group object has ended), the continuity in the spectrum between the end of one group object and the start of the next group object can link them together as coming from the same sound source. Thus if one group object has been assigned to a sound stream, a following G-O with a smooth spectral transition into it is also likely to be assigned to that sound stream.
2. Visual information about when a person starts and stops speaking can be used to assign group objects with similar onset and offset times to the appropriate sound stream.
3. Linguistic information about what a person is expected to say can be used to assign a group object to the appropriate sound stream. The group object must first be transformed into a phonetic representation, and then linguistic information can help in assigning the group object to a sound stream, based on which phonemes and words would fit in with the 'train of thought' from the desired speaker.

2. A COMPUTATIONAL MODEL OF SOUND SEPARATION

The input to this model of sound separation is the output of a computer model of cochlear processing developed by Lyon (1982, 1984). In this cochlear model, an incoming sound signal (sampled at 16 kHz) is filtered by an 85-channel cascade filter-bank. The output of each filter is maintained at the full sampling rate (16 kHz). To compress the tremendous dynamic range of the input, the different auditory nonlinearities are lumped together and modeled by the use of a coupled automatic gain control (AGC) mechanism. Finally, since auditory neurons respond only when the basilar membrane moves towards the scala media, the compressed filter-bank output is passed through a half-wave rectifier before the neural encoding stage.

The goal when performing neural encoding is to preserve the timing and intensity information in the output of the cochlear model. Rather than model neural encoding by a stochastic process that uses many neurons to probabilistically encode the amplitude and timing information of a signal, a deterministic neuron model is used. Each positive waveform peak in each frequency channel (at the output of the cochlear model) is encoded as a single neural event. The location of the neural event corresponds to the local peak in the output of that channel. The amplitude of the cochlear waveform at the peak location and the area under the cochlear waveform between zero crossings are stored along with the peak time as properties of the neural event. Two important differences between this neural event

encoding and a probabilistic neural model are: (1) the event encoding output resembles the behavior of an array of auditory fibers instead of an individual auditory neuron, and (2) the event encoding does not have a 'refractory period' (a minimum time between neural firings) and will encode all the peaks in each frequency channel. A more realistic neural model would fire at a rate below some maximum rate, and the synchrony of its firings would decrease at frequencies above 1 kHz. In each frequency region, all the neural events in a 10 ms interval make up the LAFTR.

Computation of a local periodicity feature in this model is based on Licklider's theory (1951) of pitch processing in the auditory system. According to his theory, a neural structure computes an ongoing short-time autocorrelation function of the output of each place along the cochlea. The neural structure delays each frequency channel's output through a tapped delay line, and at each tap detects the coincidence of a pulse at the delay-line output with an undelayed pulse. Every 10 ms, the **COINCIDENCE REPRESENTATION** (which computes the coinciding of neural events with previous neural events in a delay line) is computed. The coincidence representation is a two-dimensional array, parameterized by cochlear place (frequency), and the delay period (repetition period between neural firings). An example of this coincidence representation is shown in figure 1.

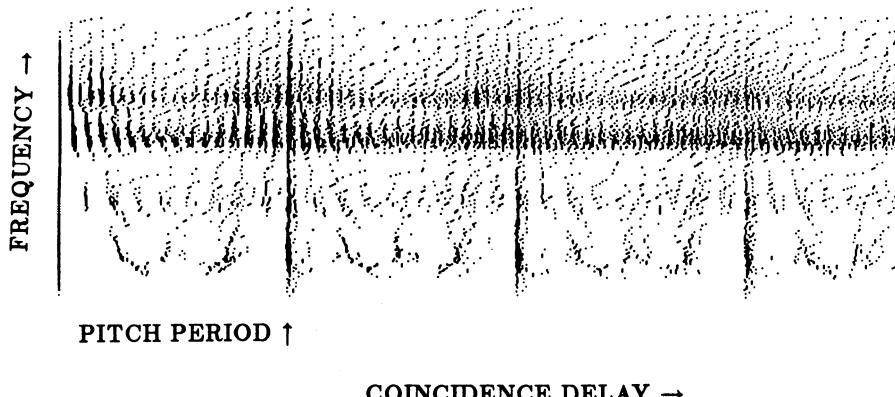


FIGURE 1, The coincidence function of the vowel /i/ in the digit /six/.

This coincidence representation is modified for use by the pitch tracking algorithm in the following stages: (1) each frequency channel is smoothed in the coincidence dimension, (2) the smoothed coincidence representation is normalized (by scaling the value at zero delay to 1.0), which allows each of the 85 channels to contribute equally to the determination of the pitch period, and (3) the values of the smoothed and normalized coincidence function near zero delay are eliminated by subtracting the smoothed and normalized coincidence function obtained from white noise. This modified coincidence representation (MCF) is then used in the dynamic programming algorithms to track the pitch period of each of two talkers.

Some important points about the way that periodicity is computed are:

1. Differences in neural firing rate information across frequency are not used to aid the computation of periodicity.
2. All frequency channels receive equal weight in the periodicity computation. This is in disagreement with Moore's recent evidence that certain harmonics are more important in the computation of the pitch period.
3. Harmonics are not resolved or detected. Each frequency region contributes equally to the computation of periodicity, whether it is responding to a single harmonic component or to several harmonic components in the same critical band.
4. The periodic signal does not have to consist of neighboring harmonic components in order for the periodicity of the sound to be computed.

Based on the results of a Markov model (which makes voiced-unvoiced-silence decisions for each of the two sounds present), all the spectral energy is assigned to that sound source if the system decides that only one sound is present. If there is more than one sound source present, the system uses a two step procedure to estimate the spectrum of each talker, as follows: (1) an initial spectral estimate of each sound source is computed using only local periodicity information, and (2) an iterative algorithm is used that locally maximizes the probability of the spectral estimate given the local periodicity information and spectral continuity constraints. The spectral continuity constraints limit the rate of spectral change (over both time and frequency) in the spectral estimate of each sound source.

The spectral estimation algorithms determine how to assign the energy in a local frequency-time region to the two sounds present, based on the observed periodicity values in that LAFTR. For example, if the coincidence representation in one frequency channel had a large peak at the pitch period of the first talker, but had no peak at the pitch period of the second talker, the estimation algorithm would decide that the energy in this frequency channel should be assigned to the first talker. The system is statistically trained (using a training database) to assign the energy in a single frequency channel on the basis of the local periodicity information contained in the coincidence representation. When either two periodic sounds are present or one talker's voice is periodic and the other's is nonperiodic, probability histograms are computed of the amplitude ratio of the two sounds (A_1/A_2), conditioned on the values of the coincidence representation in each frequency channel. For example, using a training database (of two simultaneous talkers), a histogram of the actual amplitude ratio between the two sound sources is computed when two periodic sounds are present, which is conditioned on the values of the smoothed coincidence function at the two pitch periods and on the difference between the two pitch periods. These probability histograms are used by the system to compute the probability that the energy in a certain frequency channel is most correctly assigned to one of the two periodic sound sources.

This separation algorithm requires only that frequency channel have a high value in the coincidence representation at the location of the pitch period for the energy in that region to be assigned to a sound source. It does not require that the peak (in the coincidence representation) at the pitch period be the largest peak. It does require that the peak at the actual pitch period be much larger than the peak at the other pitch period.

The above spectral estimation procedure can be viewed in the following way: (1) the pitch periods of the two sounds are computed, and then (2) an assignment (of spectral energy) is made based on the consistency of the local periodicity information contained in a LAFTR's coincidence representation with the pitch periods of a G-O. Using this approach, the spectral estimate is conditioned only on the local coincidence representation. Therefore, whether or not two different harmonic components (analyzed by two different LAFTR's) have the same onset or offset time makes no difference to the assignment (LAFTR's to G-O) of the harmonic components after the initial onset transient. This seems to conflict with the recent experimental results of Darwin and Gardner (1985). Here, the onset or offset time of two harmonic components is very important in the determination of how a component is assigned to a sound stream.

Psychoacoustic experiments (Darwin et al., 1985, Bregman, 1978) appear to indicate that whenever the auditory system can track a sound's components over time (using its acoustic properties), it will never split these components between sound sources. Certain acoustic properties such as pitch, pitch dynamics, onset, and offset can be viewed as properties of a "sound component", and these components are then assigned to "group-objects" based on the similarity of these properties. The continuity of a sound's components through time seems to be an extremely powerful piece of information that is heavily used by the auditory system.

These experiments seem to indicate that a sound's organization over time may be more important than the organization across the frequency dimension. In particular, I hypothesize that group-objects are organized as follows:

1. A sound's components are tracked over time. In the lower frequency regions, harmonics are tracked through time by following the phase-locking of neurons. In the higher frequency regions, local bands of energy are tracked through time.
2. Those sound components that were tracked through time having the same dynamics over time are grouped together into group-objects.

This problem of deciding what LAFTR's at one time are responding to the same acoustic component at the next time instant is known as the correspondence problem (Ullman 1979). In order for all the LAFTR's which have neurons that phase lock onto a particular harmonic to be assigned to a different G-O from the LAFTR's that contain harmonic components that start at a different time, the auditory system must maintain a correspondence through time (using the local acoustic features) of the LAFTR's that are responding to the same acoustic component.

The current computational model is being modified so that it can explain the separation results of Darwin (1984, 1985). In addition to these psychoacoustic modeling efforts, the separated output of the computer model is also fed into a speaker-independent continuous-digit-recognition system (Kopec and Bush, 1985). The results (described in more detail in Weintraub, 1985, 1986) indicate that the separation system results in a small improvement in recognition accuracy. It is my belief that the addition of additional acoustic information cues (e.g. pitch dynamics, common amplitude modulation characteristics between different frequency regions), as well as changes in how this information is used (by maintaining the continuity of a sound through time), could make this computational model of auditory sound separation more realistic (by closely mimicking the functional stages in human sound separation) and possibly improve its performance. In addition to these modifications in the separation algorithms, modifying a speech recognition system to deal with noisy speech signals might also lead to improvements in recognition accuracy.

REFERENCES

1. Bregman, A.S. (1984). Auditory Scene Analysis. Proceedings of the 7th International Conference on Pattern Recognition, 168-175.
2. Bregman, A.S. and Pinker, S. (1978). Auditory Streaming and the building of timbre. Canadian Journal of Psychology, 32, 19-31.
3. Broadbent, D.E. and Ladefoged, P. (1957). On the fusion of sounds reaching different sense organs. Journal of the Acoustical Society of America, 29, 708-710.
4. Darwin, C.J. (1984). Perceiving vowels in the presence of another sound: Constraints on formant perception. Journal of the Acoustical Society of America, 76, 1636-1647.
5. Darwin, C.J. and Gardner, R.B. (1985). Mistuning a harmonic of a vowel: grouping and phase effects on vowel quality. Submitted to Journal of the Acoustical Society of America, July 1985.
6. Deutsch, J.A. and Deutsch, D. (1963). Attention: Some theoretical considerations. Psychological Review, 70, 80-90.
7. Kopec, G.E. and Bush, M.A. (1985). Network-based isolated digit recognition using vector quantization. To be published in IEEE Transactions on Acoustics, Speech, and Signal Processing.
8. Licklider, J.C.R. (1951). A duplex theory of pitch perception. Experiientia, 7, 128-133.
9. Lyon, R.F. (1982). A computational model of filtering, detection, and compression in the cochlea. Proceedings of the 1982 IEEE ICASSP.
10. Lyon, R.F. (1984). Computational models of neural auditory processing. Proceedings of the 1984 IEEE ICASSP, 36.1.
11. Treisman, A.M. (1960). Contextual cues in selective listening. Quarterly Journal of Experimental Psychology, 12, 242-248.
12. Ullman, S. (1979). The interpretation of visual motion. MIT Press.
13. Weintraub, M. (1986). A computational model for separating two simultaneous talkers. IEEE International Conference on Acoustics, Speech, and Signal Processing, Paper 3.1.
14. Weintraub, M. (1985). A theory and computational model of monaural auditory sound separation. Stanford University Ph.D. Thesis.

15. Weintraub, M. (1984). The GRASP sound separation system. IEEE International Conference on Acoustics, Speech, and Signal Processing, Paper 18A.6

ON THE SIGNIFICANCE OF SPECTRAL SYNCHRONY FOR SIGNAL DETECTION

T. Houtgast

TNO Institute for Perception,
Soesterberg, The Netherlands

INTRODUCTION

Many sounds (pulses, bursts, also plosives) have a brief and broad-band character, showing a high degree of temporal synchronization along the frequency scale. For instance, after 1/3-octave band filtering, the resulting envelopes will peak at approximately the same instant. Thus, in general, such sounds will lead to synchronized stimulation in many critical bands. This paper is concerned with the significance of this stimulus parameter - the degree of spectral synchrony - for signal detection.

The experimental approach is the following. Consider a compound stimulus which consists of a combination of nine individual Gaussian-shaped tone pulses, each tone pulse covering a well defined and restricted area in the frequency-time domain. All nine individual tone pulses have the same masked threshold (in pink noise). The masked threshold of the compound stimulus is measured as a function of the configuration of the nine tone pulses in the frequency-time domain. Various such configurations are considered, including the case of perfect spectral synchrony: nine tone pulses with different carrier frequencies and coinciding peaks of the Gaussian envelopes. Of all configurations considered, the latter appears to lead to optimal detectability.

STIMULI

The elementary signal, a Gaussian-shaped tone pulse, is described in fig.1. Parameters are the nominal frequency and temporal values f_0 and t_0 . The multiplication factor $f_0^{1/2}$ makes the energy of a tone pulse independent of f_0 .

The value of a in the argument of the exponent determines the effective duration and bandwidth. For $a=0.2$, the compromise in limiting both duration and bandwidth leads to the values for Δt and Δf as indicated. Note that $\Delta f=0.2f_0$ implies that the effective bandwidth is a constant proportion of the nominal frequency (thus, the bandwidth is slightly smaller than 1/3 octave). Note also, that $\Delta t=5/f_0$ implies that the effective duration is five periods of the nominal (carrier) frequency, which amounts to 10 ms at 500 Hz and shortens with increasing f_0 . The lower part of fig. 1 indicates how the elementary signal can be characterized in the frequency-time domain.

Fig. 2 displays several configurations for a compound stimulus. The grid in the frequency-time domain is defined by nine positions along the frequency scale (at 1/3-octave intervals) and nine positions

along the time scale (at 10-ms intervals). Note that by the definition of the elementary signal, the grid ensures that the elementary signals are well separated in frequency and/or time.

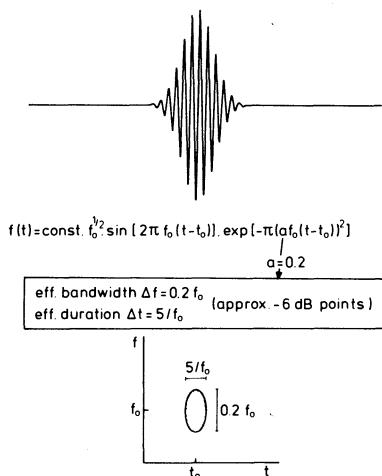


FIGURE 1, Description of the elementary signal, covering a restricted area in the frequency-time domain. A compound stimulus consists of nine such signals, with various combinations of (f_0, t_0) .

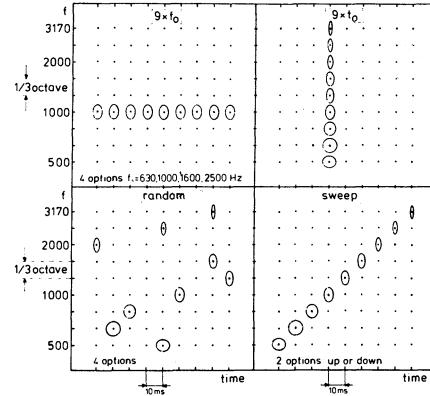


FIGURE 2, Four examples of compound stimuli. The nine elementary signals are distributed on a 9×9 grid such that they are well separated in frequency and/or time.

Table I, Listing of the six types of stimuli involved. Since most stimulus types include several options the total number of stimuli amounts to 16.

Stimulus	Symbol in fig. 3
Single	o
$9 \times f_0$	F
$9 \times t_0$	S
RANDOM	R
SWEEP	U, D
9×9	■

The configuration indicated in the upper-right panel in fig. 2 is the one of special interest, i.e. the case of spectral synchrony. Besides the ones illustrated in fig. 2, several other stimulus conditions were considered as well. A complete listing is given in Table I. All stimuli were calculated and stored digitally (40- μ s samples, 12-bit D-A conversion).

METHOD

Subjects listened on headphones (Beyer DT-48). The masker was continuous pink noise (spectral level -3 dB/octave) at a Sensation Level of about 50 dB. In a pilot experiment it was verified that for the pink-noise masker, each elementary signal with f_0 in the range of 500-3170 Hz led to the same masked threshold (recall that the energies of the elementary signals were the same).

Detection thresholds were determined in a four-interval forced-choice procedure, with feedback. The four intervals in a trial were indicated by signal lights. During a run, four of the 16 possible stimuli (Table I) were considered, and at each individual trial a random selector determined which of those four stimuli was to be presented. The presentation levels of each of the four possible stimuli were kept within the relevant range by an adaptive procedure (stimulus level +/- 2 dB after correct/incorrect responses). A run consisted typically of 400 trials, 100 for each stimulus. Each of the 16 possible stimuli was included in at least four runs. Responses were pooled for each stimulus and each stimulus level, and are plotted as psychometric functions.

The procedure of combining several possible stimuli within a run prevents the subject from tuning in to one particular stimulus; the subject has to keep an "open mind" in both frequency and time. Especially in case of a random configuration, it may well be that threshold decreases after prolonged training on one particular frozen sample. The present procedure leads to what might be called the "spontaneous" masked threshold.

RESULTS

The results of two subjects are presented in fig. 3. All levels are presented relative to the masked threshold (62.5% score) for the single elementary signals. Of the compound stimuli consisting of nine elementary signals, the one labeled S (synchrony) yields the lowest masked threshold.

In terms of total energy, the summation of nine elementary signals leads to an increase of 9.5 dB, and 81 corresponds to 19.1 dB. When this is taken into account, the masked threshold depends on the stimulus configuration as presented in fig. 4. (These data represent the mean of nine subjects, and were obtained with a method somewhat different from the one described before.)

DISCUSSION

Fig. 4 leads to the following observation. Given a continuous pink-noise masker and given a frequency-time window of 500-3170 Hz along the frequency scale and 100 ms along the time-scale, the distribution of stimulus energy within that window influences the

detection threshold. Concentrating the energy in frequency and/or time (single, $9 \times f_0$, $9 \times t_0$) leads to better detectability than distributing the energy over both frequency and time ($9 \times \text{RANDOM}$, 9×9).

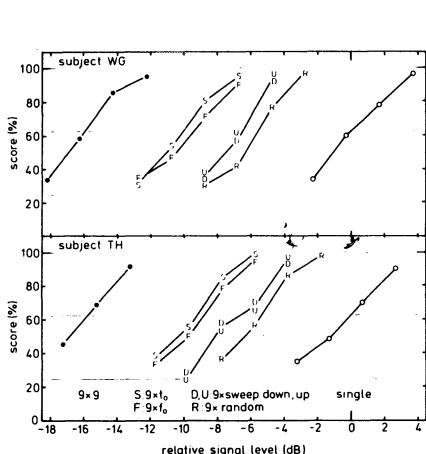


FIGURE 3, Scores obtained for two subjects in a four-alternative forced choice procedure for various signal levels for the six types of stimuli involved (specified in Table I). Levels are expressed relative to the masked threshold level of the single elementary signals (define by a score of 62.5%).

The close relation between "single" and " $9 \times f_0$ " is a traditional result: both stimuli are restricted in bandwidth to 1/3-octave and thus the total energy essentially determines masked threshold (at least for a duration up to 100 ms). The equally important effect of concentrating stimulus energy in the time domain ($9 \times t_0$) illustrates the significance of spectral synchrony. As indicated in fig. 4, the effect amounts to about 5 dB: for a broad-band stimulus, with energy distributed from 500 to 3170 Hz, a brief synchronized presentation yields a 5 dB lower threshold than a non-synchronized presentation within a 100 ms interval. The finding of synchrony-facilitated broad-band detection has interesting theoretical implications. Traditional detection theories are strongly critical-band oriented, and spectral synchrony is not acknowledged as a relevant parameter. In case of broad-band detection (i.e., the threshold being approached in a number of individual critical bands), some process of across-frequency integration must be involved. Apparently, the efficiency of that process benefits from spectral synchrony.

equal-energy thresholds

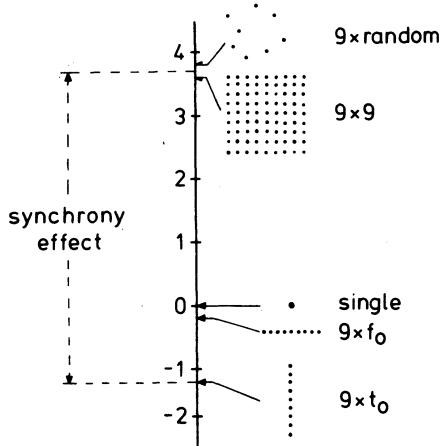


FIGURE 4, Relation among the masked threshold levels of five types of stimuli after compensation for the energy increment when adding nine (+9.5 dB) or 81 (+19.1 dB) elementary signals. Mean data of nine subjects.

Various further pilot experiments are being performed, concerning the following questions:

1. Does synchrony-facilitated detection improve even further when the dispersion in the peripheral auditory system is compensated for by a slight delay of the high-frequency parts of the stimulus?
2. What is the influence of the duration of the elementary signals (i.e., is the synchrony effect restricted to the present brief stimuli)?
3. To what extent is speech intelligibility in noise affected by introducing some degree of de-synchronization across the frequency scale?

AUDITORY ENHANCEMENT IN SPEECH PERCEPTION

Quentin Summerfield and Peter Assmann
MRC Institute of Hearing Research, University of Nottingham, U.K.

1. INTRODUCTION

Summerfield et al. [22] described a perceptual effect of temporal spectral contrast which they called the 'Flat-spectrum Vowels' effect. They summed the first 50 harmonics of 100 Hz with equal amplitudes, creating a signal with a flat spectrum. Three sets of adjacent harmonics were then omitted at frequencies centred on the first three formants of a vowel, creating the 'spectral complement' of the vowel. Stimuli were constructed in which 500-ms segments of a vowel complement preceded and followed a 500-ms segment of the flat spectrum. Listeners identified the flat spectrum as the vowel whose complement surrounded it, despite the absence of formant peaks. The effect reveals the existence of processes that enhance newly-arriving acoustical energy in relation to pre-existing energy. These processes could improve the effective signal-to-noise ratio of intermittent signals in background noises. The aims of this paper are to discuss factors that may contribute to this effect and to illustrate roles it may play in the perception of speech.

Subsequent experiments demonstrated that it was the complement preceding the flat spectrum that was important for producing the effect, and revealed similarities to effects reported previously [2,24,25,27,28]. These results can be summarised by saying that when the spectral valleys in a signal with a complex spectral envelope are filled, the resulting signal with a uniform spectrum may be perceived to possess a complementary spectral envelope. In discussing these results we shall refer to segments playing the role of vowel complements as 'precursors', to segments playing the role of flat spectra as 'test stimuli', and to the general phenomenon revealed in [2,22,24,25,27,28] as the 'enhancement' effect.

Three types of explanation can be put forward to account for the enhancement effect, founded, respectively, in (i) peripheral auditory adaptation, (ii) perceptual grouping, and (iii) comparison of spectral profiles. Depending on the configuration of the stimuli, each may contribute, but to different degrees.

2. PERIPHERAL AUDITORY ADAPTATION.

The effect could originate in a form of peripheral auditory adaptation qualitatively akin to the form displayed by primary auditory nerve fibres [9] (Figure 1, left-hand panel). The auditory excitation in channels stimulated by components in a vowel complement diminishes over time because of adaptation. When the previously missing

components are re-introduced, they excite relatively unadapted channels, produce more excitation than the pre-existing components, and stand out perceptually in relation to them. Peripheral auditory adaptation provides powerful concepts which may explain a range of perceptual phenomena obtained with speech sounds [e.g. 3,4, and Lacerda, this volume].

Five observations are broadly consistent with an origin for the enhancement effect in peripheral auditory adaptation: that it occurs only when precursor and test stimulus are presented to the same ear [22,23,24], that it is more potent the more abruptly components are re-introduced [24], that its size increases with the intensity of the stimuli [24,27], that listeners judge the intensity of the pre-existing components to diminish [23], and that estimates of the time constant of the development of the effect (ranging from 20 ms [3] to 300 ms [24]) and that of adaptation in primary auditory nerve fibres (e.g. 50-60 ms [26]) are similar.

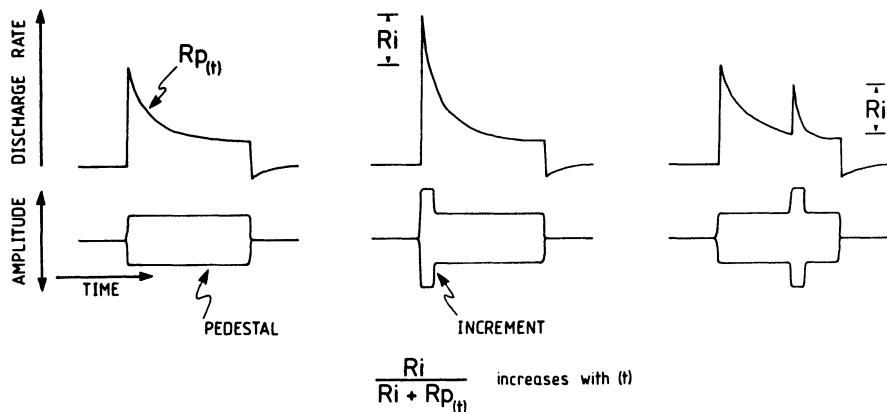


FIGURE 1, Schematic discharge patterns of primary auditory nerve fibres in response to pedestal tones [$R_p(t)$] and intensity increments [R_i], after [13]. Because [R_i] is time-invariant for given amplitudes of pedestal and increment, the ratio [$R_i/R_p(t)$] increases with time, potentially increasing the prominence of energy at the frequency of the increment.

Summerfield et al. [23] suggested a slight refinement to this account: the effect is more likely to reflect the incremental response of adapted primary auditory nerve fibres [19], than differences in levels of activity between adapted and unadapted fibres. There are two parts to this argument. First, because the frequency resolution of the peripheral auditory system is finite, the valleys in the auditory excitation pattern of a vowel complement are not infinitely deep. Thus, when the missing components are re-introduced, increments in excitation occur in frequency channels where there has been some prior stimulation, rather than in channels where previously there was none. Compatibly, it is not necessary to omit components completely to produce an effect; the valleys in a vowel complement need only be a few dB deep [22,23]. Second, Smith [19] demonstrated that the increase in discharge rate of primary auditory nerve fibres in response to a brief increment in the level of a pedestal tone is time-invariant (Figure 1). As a result, the ratio of the additional discharge produced

by the increment [Ri] to the short-time discharge produced by the pedestal [Rp(t)] is greater, the later the increment occurs. This ratio may determine the auditory prominence of the increment. Its increase over time leads to the prediction that increments should be easier to detect the later they occur. This prediction is born out psychoacoustically for the detection of brief sinusoids in broad-band noises [29]. If an incremental response occurs when a component is re-introduced into a harmonic series, there will be more excitation in channels tuned to its frequency than in channels tuned to the frequencies of the pre-existing components, causing the re-introduced component to stand out perceptually, and giving rise to the enhancement effect.

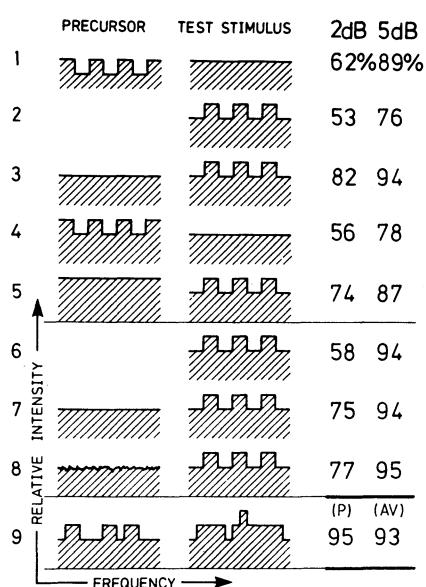


FIGURE 2, Schematic spectra of precursors and test stimuli, and %-correct accuracy of identification of (i) test stimuli for 2-dB and 5-dB changes in spectral amplitude (conditions 1-8), and (ii) precursors (P) and 'added vowels' (AV) (condition 9). (Conditions 1-5, n=6; Conditions 6-8, n=3; Condition 9, n=4).

Figure 2 shows results that are broadly compatible with this account. Stimuli were synthesised digitally by summing the first 50 harmonics of 100 Hz with randomised phases. Onsets and offsets were shaped by a 10.7 ms Kaiser function. Durations between the -6 dB points were 988.6 ms (precursors) and 188.6 ms (test stimuli). Stimuli were presented monaurally to the left ears of listeners with audiometrically normal hearing. Stimulus level was such that the 2-kHz component in stimuli with a flat spectrum was 41 dB SPL. Test stimuli could be identified as /i/, /a/, /u/, /ɔ/, or /ɛ/ as a result of changes in the levels either of 6 harmonics (3 pairs of adjacent harmonics straddling the centre frequencies of the first 3 formants of the vowels [Table I]) or of the remaining 44 harmonics. Results are given in terms of %-correct vowel identification for two instances of each condition with, respectively, 2 dB and 5 dB changes in harmonic amplitude between precursors and test stimuli. For example, Condition 1 in Figure 2 shows a version of the flat-spectrum vowels effect; accuracy of vowel identification was 61.7% when the valleys in the vowel complement were 2 dB deep, and 88.5% when they were 5 dB deep. Both performance levels were significantly above chance (20%).

Table I: Frequencies of harmonics defining formants

Vowel	F1	F2	F3
EE	200,300	2200,2300	3000,3100
AH	600,700	900,1000	2900,3000
OO	200,300	800,900	1900,2000
OR	300,400	700,800	2800,2900
ER	400,500	1200,1300	2600,2700

In Condition 2, isolated test stimuli were presented in which the 3 pairs of harmonics had been raised above the level of the flat spectrum. When the same test stimuli were introduced by a precursor with a flat spectrum in Condition 3, performance improved, demonstrating that increments in spectral amplitude are enhanced, irrespective of whether they fill spectral valleys. We shall refer to the improvement in performance between Conditions 2 and 3 as the 'Peaked-spectrum Vowels' effect. Smith et al. [20] demonstrated that the decrease in discharge rate in response to a brief decrement in the intensity of a pedestal is time-invariant, like the incremental response. Conditions 4 and 5 show analogous enhancement effects. Condition 4 is a variant of the flat-spectrum vowels effect, but now the uniform spectrum was created by lowering the levels of the 44 harmonics corresponding to the peaks in the vowel complement. Performance was poorer than in Condition 1, but still significantly above chance. Condition 5 is the corresponding peaked-spectrum vowels effect. The test stimuli were identified more accurately here than when presented in isolation in Condition 2. Collectively, the results of Conditions 1-5 are consistent with an explanation for the enhancement effect in which changes in spectral amplitude are enhanced when they occur in frequency regions where there has been prior stimulation and thus adaptation.

However, there are reasons for caution in concluding that all aspects of the enhancement effect can be explained in terms of processes akin to adaptation in primary auditory nerve fibres. First, enhancement may involve not only a reduction in the effective level of pre-existing components [23], but also a genuine increase in the effective level of the re-introduced component [23,25]. In [25] it was suggested that the ability of the pre-existing components to suppress energy at the frequency of the re-introduced component might diminish because of adaptation, allowing the re-introduced component to produce more excitation than normal. Higher-level processes could also be involved. Some units in the cochlear nucleus produce responses to the changes in intensity in amplitude-modulated tones [6] that are enhanced relative to the modulations in the waveform envelope, particularly in the presence of higher-frequency inhibitors of fixed amplitude [13].

Some estimates of the time constant of the decay of the enhancement effect (e.g. 300 ms [2,22]) are similar to the time constant of the recovery of discharge rate in primary auditory nerve fibres (e.g. 170 ms in gerbil [26]) and of the compound action

potential in man (380 ms [5]). However, a second reason for doubting whether adaptation can explain all aspects of the effect is that other estimates (e.g. 1 s [27] and 2 s [24]) exceed the duration of the recovery of any process of physiological adaptation known to occur at low sound-pressure levels. Thus, it is possible that processes with longer memories involving attention and pattern processing are partly responsible for enhancement effects.

3. ATTENTION AND PATTERN PROCESSING

(a) Perceptual Grouping

In the flat-spectrum vowels effect, the re-introduced components might be grouped together because they share a common amplitude envelope (Bregman, this volume) and so be segregated from the pre-existing components. Compatible with this argument are the observations that other manipulations such as brief changes in the level [16] or phase [12] of an incompletely resolved component in a harmonic series can draw a listener's attention to that component, and increase its perceptual prominence. This explanation can account for both the flat- and the peaked-spectrum vowels effects by arguing that in Conditions 1 and 3 of Figure 2 the 6 harmonics defining the formants of the vowel possess a different amplitude envelope from the remaining 44 harmonics and so might be segregated from them, whereas in Condition 2 all 50 harmonics share the same amplitude envelope and would be interpreted together.

However, this explanation has difficulty in accounting for two other observations. First, it is not clear why the effects shown in Figure 2 are not eliminated once a detectable silent interval is introduced between precursor and test stimulus, since we should expect all 50 harmonics in the test stimulus then to be grouped together. Second, Condition 8 in Figure 2 shows that enhancement of the vowel in a harmonic test stimulus (similar to that in Condition 2) occurs when the precursor is a broad-band noise, where again we should not expect the 6 harmonics defining the vowel in the test stimuli to be grouped separately from the other 44 harmonics. This result, together with the demonstration of increased forward masking by a re-introduced component [25], suggest, that processes responsible for perceptual grouping may benefit from the enhancement effect, but that they are not its cause.

(b) Comparison of Spectral Profiles

For enhancement to occur, a precursor must precede a test stimulus [24]. The precursor provides a spectral profile with which the profile of the test stimulus can be compared. Possibly spectral profiles can be measured, stored, and compared in ways that highlight differences between successive spectra. Running counter to this suggestion is the observation that the effects do not occur if precursor and test stimuli are presented to different ears. However, the ability of listeners to detect changes in the level of one or more components of a complex tone between the intervals of a two-alternative forced-choice experiment [7,21] shows some parallels with the results summarised in Figure 2. Listeners can maintain a high level of performance even when the ISI is extended to several seconds, and when the overall levels of the stimuli to be compared in the two

intervals are randomised. We have yet to determine whether the effects shown in Figure 2 occur when precursors and test stimuli have appreciably different overall levels and, in particular, whether they occur when level is unpredictable.

4. PRECURSORS WITH VOWEL-LIKE SPECTRA

Condition 9 of Figure 2 suggests that increments to components in a precursor that itself possesses spectral peaks can be enhanced. Here the precursor was one of the five vowels, and test stimuli were created by increasing the levels by 5 dB of 3 pairs of harmonics corresponding to the formants of one of the other 4 vowels. Each precursor and test stimulus sounded like a sequence of two vowels. Potentially, the result demonstrates that the enhancement effect is sufficiently potent to allow a newly arriving signal to be distinguished from pre-existing energy even when both have similar, complex, spectral envelopes.

However, it might be wrong to attribute the accuracy of identification of the 'added' vowel entirely to the enhancement effect. It is possible to identify both members of a pair of vowels that are presented concurrently with the same amplitude envelope, fundamental frequency, and overall amplitude, with an accuracy significantly above chance [17,30]. In Condition 9, the test stimulus was a pair of concurrent vowels. The precursor in effect told the listener the identity of one of them. Possibly, this knowledge guided pattern recognition processes in identifying the added vowel. It might do this by defining a spectral pattern which could be subtracted from the spectral pattern of the test stimulus to reveal the added vowel [30].

To dissociate contributions of pattern processing from the enhancement effect, test stimuli were created by summing the waveforms of two of the five vowels, including the case where a vowel was added to itself. The precursor was a single vowel. The subjects' task was to identify both vowels in the test stimulus. In different conditions the precursor was (i) omitted, (ii) displayed on a VDU as its orthographic approximation ('EE', 'AH', 'OO', 'OR', or 'ER') indicating its phonemic identity, but providing no immediate evidence of its acoustical structure, (iii) presented acoustically to the ear opposite the test stimulus indicating both its phonemic identity and its acoustical structure, or (iv) presented acoustically to the same ear as the test stimulus. Only in Condition (iv) could the enhancement effect contribute to identifying the added vowel.

Three sets of stimuli were constructed to allow the influence of the naturalness of the spectral shape of the component vowels to be assessed. In each set the vowels were constructed from the first 50 harmonics of 100 Hz. Set 1 was created from a signal with a uniform spectrum by raising the levels of three pairs of adjacent harmonics by 9 dB. Set 2 was similar except that the other 44 harmonics were omitted. Set 3 was created using cascade formant synthesis [10]. Waveform-envelope shaping, presentation conditions, and subject-selection criteria, were the same as those described previously.

Figure 3 shows line spectra of the three versions of the vowel /i/ along with their auditory excitation patterns computed according to the formulae suggested by Moore and Glasberg [14]. Table II displays

results for each stimulus set expressed as the accuracy with which both vowels in the test stimuli were identified correctly. Although there were small differences in performance among the three sets, the pattern of differences across conditions was the same for each set. All subjects confirmed a role for the enhancement effect in helping to identify the added vowel by producing better performance in Condition (iv) (acoustic precursor in same ear as the test stimulus) than in Condition (ii) (orthographic precursor) or (iii) (acoustic precursor in opposite ear to the test stimulus). Conditions (ii) and (iii) did not produce reliably different performance, though each was superior to Condition (i) (no precursor). When the results were analysed in terms of the accuracy with which the 'enhanced' vowel in the test stimulus (i.e., the vowel not identical to the precursor) was identified correctly, condition (iv), but not conditions (ii) and (iii), showed a significant improvement over condition (i).

Table II: %-correct identification of both vowels in 'double' vowels [conditions (i)-(iv)] and single vowels [condition (v)] for three stimulus sets ($n=5$, set 1; $n=6$, sets 2 and 3).

		1	2	3
(i)	Isolated double vowels	33.5	54.2	49.2
(ii)	Orthographic Precursor	55.8	67.9	67.1
(iii)	Contralateral Precursor	55.0	72.9	68.7
(iv)	Ipsilateral Precursor	67.8	87.7	83.8
(v)	Isolated single vowels	96.8	98.3	99.0

Following suggestions in [1], [15], [16], and [17], we have attempted to model the perceptual data from conditions (i) and (iv) by means of spectral distance metrics applied to auditory excitation

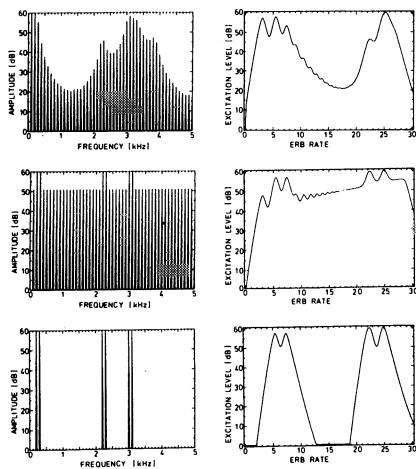


FIGURE 3, Line spectra (left) and erb-rate auditory excitation patterns (right) for the vowel /i/ from each stimulus set. Set 1 (top), Set 2 (middle), Set 3 (bottom).

patterns. We have assumed that the frequency with which a particular response category is assigned to a double vowel can be predicted by the degree of similarity between the auditory spectrum of the double vowel and that of the response vowel. One metric which holds promise is a modification of the weighted spectral slope metric proposed by Klatt [11]. The modified spectral representation is the negative portion of the second differential of the excitation pattern. This representation highlights spectral peaks and edges relative to spectral valleys, as shown in Figure 4 for the vowel /i/ from Set 3 (cascade formant synthesis). The product-moment correlations between the predicted distances and response profiles from Condition (i) are -0.90 (Set 1), -0.83 (Set 2), and -0.88 (Set 3). These high correlations, and the fact that perceptual performance was similar with the three acoustically different stimulus types, reflect the importance of spectral peaks and shoulders for judgements of phonetic quality in vowels [cf. 1, 11]. Finally, we found that the response profiles to the 'added' vowel in Condition (iv) are best predicted by assuming that the listener responds to the added vowel with no contribution from the background vowel. This outcome is compatible with the subjective experience of hearing the precursor and test stimulus as a sequence of two vowels.

5. SUMMARY: ROLE OF THE ENHANCEMENT EFFECT IN SPEECH PERCEPTION

In the flat- and peaked-spectrum vowels effects the precursor can be regarded as a noise and the test stimulus as the same noise to which a signal has been added. The signal is easier to detect when it begins some time after the start of the noise, rather than at the same time as the noise. In the experiments summarised in Figure 2, the task of distinguishing signal from noise was made artificially difficult by giving each the same harmonic structure. However, analogous effects

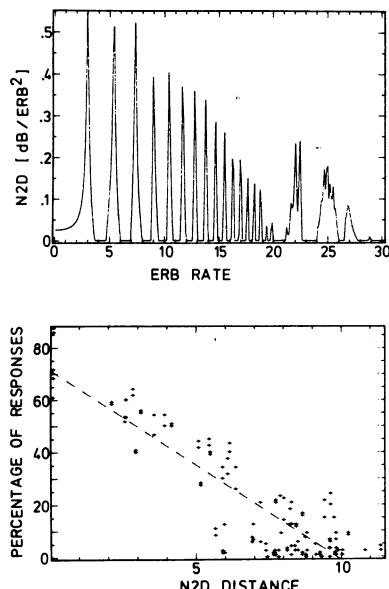


FIGURE 4, (Top) Negative portion af the second differential of the erb-rate excitation pattern (N2D) of /i/ (cascade synthesis) (compare Fig. 3). (Bottom) Relationship between the probability of each of 5 alternative responses (/i/, /a/, /u/, /ɔ/, /ɜ/) to a 'double' vowel from Set 3 (cascade synthesis) and the computed distance (after [15]) between the N2D representations of the double vowel and each of the 5 single vowels. 125 data points (25 double vowel combinations with 5 response alternatives per combination) averaged over 6 listeners ($r=-0.88$).

occur with more natural noises and signals. For example, Scheffers [18] has shown that the threshold for distinguishing Dutch vowels presented in pink noise is 3-4 dB lower when the vowels start 300 ms after the onset of the noise than when the two begin together. Similarly, informal observations of our own suggest that consonants in isolated CV syllables can be identified more accurately when they start a few hundred milliseconds after a broad-band noise than when the two start together.

The effects summarised in Figure 2 can also be interpreted as showing that the auditory system highlights changing acoustic patterns in preference to spectrally-static patterns. This is likely to have general utility in representing natural time-varying signals, and specific value when signals must be detected against low levels of noise through communication channels such as reverberant rooms and hearing aids that have non-uniform frequency responses. In emphasising frequencies at which the spectral amplitude is changing, the auditory system will tend to reduce the effects of spectral colouration from the communication channel [8].

REFERENCES

1. Assmann, P.F. (1985). The role of harmonics and formants in the perception of vowel quality. Unpublished Ph.D. Thesis, University of Alberta.
2. Cardozo, B.L. (1967). Ohm's Law and masking. IPO Annual Progress Report, No. 2, 59-64 (Institute for Perception Research, Eindhoven, Netherlands).
3. Chistovich, L.A., Lublinskaya, V.V., Malinnikova, T.G., Ogorodnikova, E.A., Stoljarova, E.I., and Zhukov, S.Ja. (1982). Temporal processing of peripheral auditory patterns of speech. In The Representation of Speech in the Peripheral Auditory System, R. Carlson and B. Granström (Eds.). Elsevier, Netherlands. Pp. 165-180.
4. Delgutte, B. (1982). Some correlates of phonetic distinctions at the level of the auditory nerve, in The Representation of Speech in the Peripheral Auditory System, R. Carlson and B. Granström (Eds.). Elsevier, NL. Pp. 131-149
5. Eggermont, J.J. and Odenthal, D.W. (1974). Action potentials and summating potentials in the normal human cochlea, Acta Otolaryngologica, Supplement, 316, 39-61.
6. Frisina, R.D. (1983). Enhancement of responses to amplitude modulation in the gerbil cochlear nucleus: Single-unit recordings using an improved surgical approach. Special Report. ISR-S-23. (Institute for Sensory Research, Syracuse University, NY.)
7. Green, D.M. and Kidd, G. (1983). Further studies of auditory profile analysis. Journal of the Acoustical Society of America, 73, 1260- 1265.
8. Haggard, M.P. (1977). Mechanisms of formant frequency discrimination. In Psychophysics and physiology of hearing, E.F. Evans and J.P. Wilson (Eds.). Academic Press. Pp. 499-507
9. Kiang, N.Y.S., Watanabe, T., Thomas, E.C., and Clark, L.F. (1965). Response patterns of single fibers in the cat's auditory nerve. MIT Press, Cambridge, MA.
10. Klatt, D.H. (1980). Software for a cascade/parallel formant synthesizer. Journal of the Acoustical Society of America, 67, 971- 995.

11. Klatt, D.H. (1982). Speech processing strategies based on auditory models. In The Representation of Speech in the Peripheral Auditory System, R. Carlson and B. Granström (Eds.). Elsevier, Netherlands. Pp. 181-196.
12. Kubovy, M. and Jordan, R. (1979). Tone segregation by phase: On the phase sensitivity of the single ear, Journal of the Acoustical Society of America, 66, 100-106.
13. Møller, A. (1975). Dynamic properties of excitation and inhibition in the cochlear nucleus, Acta Physiologica Scandinavica, 93, 442-454.
14. Moore, B.C.J. and Glasberg, B.R. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns, Journal of the Acoustical Society of America, 74, 750-753.
15. Nocerino, N., Soong, F.K., Rabiner, L.R., and Klatt, D.H. (1985). Comparative study of several distortion measures for speech recognition. Speech Communication, 4, 317-331.
16. Plomp, R. (1976). Aspects of Tone Sensation. Academic, London.
17. Scheffers, M.T.M. (1979). The role of pitch in perceptual separation of simultaneous vowels, IPO Annual Progress Report, 14, 51-54 (Institute for Perception Research, Eindhoven, Netherlands).
18. Scheffers, M.T.M. (1983). Identification of synthesized vowel sounds in a noise background. IPO Manuscript No. 450/II (Institute for Perception Research, Eindhoven, Netherlands).
19. Smith, R.L. (1979). Adaptation, saturation, and physiological masking in single auditory nerve fibres, Journal of the Acoustical Society of America, 65, 166-178.
20. Smith, R.L., Brachman, M.L., and Frisina, R.D. (1985). Sensitivity of auditory-nerve fibres to changes in intensity: A dichotomy between decrements and increments. Journal of the Acoustical Society of America, 78, 1310-1316.
21. Spiegel, M.F., Picardi, M.C. and Green, D.M. (1981). Signal and masker uncertainty in intensity discrimination. Journal of the Acoustical Society of America, 70, 1015-1019.
22. Summerfield, Q., Haggard, M.P., Foster, J. and Gray, S. (1984). Perceiving vowels from uniform spectra: phonetic exploration of an auditory after-effect. Perception and Psychophysics, 35, 203-213.
23. Summerfield, Q., Sidwell, A., and Nelson, T. (1986). Auditory enhancement of changes in spectral amplitude. Journal of the Acoustical Society of America (in press).
24. Viemeister, N.F. (1980). Adaptation of masking. In Psychophysical, Psychological, and Behavioural Studies in Hearing, G. v.d. Brink and F.A. Bilsen (Eds.). Delft University Press. Pp. 190-199.
25. Viemeister, N.F. and Bacon, S. (1982). Forward masking by enhanced components in harmonic complexes. Journal of the Acoustical Society of America, 71, 1502-1507.
26. Westerman, L.A. (1985). Adaptation and recovery of auditory nerve responses. Special Report. ISR-S-24. (Institute for Sensory Research, Syracuse University, NY).
27. Wilson, J.P. (1970). An auditory after-image. In Frequency Analysis and Periodicity Detection in Hearing, R. Plomp and G.F. Smoorenburg (Eds.) A.W. Sijthoff, Leiden, Netherlands.
28. Zwicker, E. (1964). 'Negative afterimage' in hearing, Journal of the Acoustical Society of America, 36, 2413-2415.

29. Zwicker, E. (1965). Temporal effects in simultaneous masking and loudness. Journal of the Acoustical Society of America, 38, 132-141.
30. Zwicker, U.T. (1984). Auditory recognition of diotic and dichotic vowel pairs. Speech Communication, 3, 265-277.

GENERAL DISCUSSION OF SESSION 2: SEPARATION OF SIMULTANEOUS EVENTS

Chairman: Brian C.J. Moore

With regard to Bregman's paper, there seemed to be general agreement with Bregman's account of duplex perception, indicating that it was not unique to speech perception. The question remained, nevertheless, whether it gave any special insight into mechanisms of speech perception. It was suggested, that, if it were possible to investigate duplex perception in infants, it would be interesting to determine whether it occurred at a particular age, linked to the development of linguistic knowledge.

One objection to Bregman's account was that "transparency" is not in fact a commonly occurring property of sounds. Rather, one sound would only be audible "behind" another if the levels of the sounds were comparable. More normally, the more intense sound would mask the weaker one. Bregman's response was that the methods used by the auditory system to separate simultaneous sounds do not guarantee success. For example, temporal synchrony may be used to group sounds together, but this does not always work reliably in reverberant rooms. Hence the auditory system is prepared to make use of rules which work some of the time. In general, we operate in a redundant way, making use of several different methods to separate acoustic events. With regard to duplex perception, the key question is: are there ever situations where it is justified to use the same piece of information twice? The answer is "yes", and this explains why the auditory system is able to do this.

The question was raised as to whether the auditory system used methods of separating acoustic events which are specific to speech. The answer seems to be "yes". For example, we can achieve a reasonable degree of accuracy in identifying two simultaneous vowels which are synthesised with exactly the same envelope and fundamental frequency. In this case, there are no obvious acoustic cues which could be used to achieve the separation of the two sounds, and knowledge about the spectral shapes of the vowels must be used. Of course, the use of knowledge of this type is not restricted to speech; our knowledge of the typical spectral composition of notes played by musical instruments might allow a similar separation of simultaneous notes.

A question was raised about the relative importance for sound separation of temporal modulation patterns in different frequency bands, and of the tracking of individual harmonics over time. Some models for separation of simultaneous sounds assume a rather low degree of frequency selectivity, and make use mostly of the temporal pattern of modulation. Other models assume greater frequency selectivity and place a greater emphasis on tracking harmonics over

time. There was general agreement that both types of cues exist and are used by the auditory system.

It was pointed out that when listening to synthetic speech it is often possible to hear the intended speech sounds, while at the same time certain extraneous sounds are audible, not "belonging" to the speech. This seems to be an example of duplex perception. The relative levels of the different acoustic elements seem to be important in this effect, and in the demonstration of duplex perception in which the formant transitions only are presented to one ear. The formant transitions are both integrated with the steady-state portion of the sound in the opposite ear and are heard as a chirp. If the level of the transitions is reduced, however, then a point is reached when they are not heard as a separate sound, but nevertheless they still form an integrated percept with the sound in the opposite ear. It is as if the phonetic processor takes its "share" of the sound first, and the auditory processor takes what is left (if anything).

Chapter 3

DYNAMIC ASPECTS

TRADING RELATIONS, ACOUSTIC CUE INTEGRATION, AND CONTEXT EFFECTS IN SPEECH PERCEPTION*

D.B. Pisoni and P.A. Luce

Speech Research Laboratory, Indiana University, Department of Psychology, Bloomington, Indiana 47405, USA

The study of speech perception differs in several very important ways from the study of general auditory perception. First, the signals typically used to study the functioning of the auditory system have been simple, discrete and well defined mathematically. Moreover, they typically vary along one perceptually relevant dimension. In contrast, speech sounds involve very complex spectral relations that typically vary quite rapidly as a function of time. Changes that occur in a single perceptual dimension almost always affect the perception of other attributes of the signal. Second, most of the basic research on auditory perception over the last four decades has been concerned with problems surrounding the discriminative capacities of the sensory transducer and the functioning of the peripheral auditory mechanisms. In the perception of complex sound patterns such as speech, the relevant mechanisms are, for the most part, quite centrally located. Moreover, while many experiments in auditory perception, and sensory psychophysics have commonly focused on experimental tasks involving discrimination of both spectral and temporal properties of auditory signals, such tasks are often inappropriate for the study of more complex signals including speech. Indeed, in the case of speech perception and probably the perception of other complex auditory patterns, the relevant task for the observer is more nearly one of absolute identification rather than differential discrimination. Listeners almost always try to identify, on an absolute basis, a particular stretch of speech or try to assign some label or sequence of labels to a complex auditory pattern. Rarely, if ever, are listeners required to make fine discriminations that approach the limits of their sensory capacities.

Given the published literature on the perception of simple auditory signals, it is generally believed, at least among researchers in the field of speech perception, that a good deal of what we have learned from traditional auditory psychophysics using simple sinusoids is only marginally relevant to the study of speech perception. Perhaps some of what is currently known about speech perception might be relevant to the perception of other complex auditory patterns which have properties that are similar to speech. At the present time, there are substantial gaps in our knowledge about the perception of complex signals that contain very rapid spectral changes such as those found in speech. And, there is little if any research on the perception of complex patterns that have the typical spectral peaks and valleys that

*This research was supported, in part, by NIH research grant NS-12179 to Indiana University in Bloomington

speech signals have. Finally, our knowledge and understanding of patterns containing amplitude variations like the complex temporal patterns found in speech is also quite meager at this point in time. Obviously, there is a lot of basic research to do.

A voiced (periodic) speech signal is typically thought to be produced by excitation of a time-varying filter with a source spectrum which has harmonics at multiples of the fundamental. For unvoiced (aperiodic) signals, the situation is somewhat more complicated because the source spectrum is continuous and may contain energy at all frequencies and the location of the energy in the vocal tract can occur at a number of different locations between glottis and lips. However, in considering only voiced sounds, it has been convenient to assume, for modeling purposes, that the interactions between source and filter are minimal, and thus it is theoretically convenient to dissociate properties related to the source spectrum from properties imposed by the vocal tract transfer function. Thus, the relevant perceptual attributes for the perception of segmental sounds of speech are closely associated with changes in spectral shape over time. In contrast, the relevant perceptual attributes for the perception of suprasegmental or prosodic attributes of speech are related to the changes in the temporal properties of speech, such as duration and variations in pitch and amplitude as a function of time. Considering only the segmental properties of speech sounds in patterns of word-length size, it is possible to generate an enormously rich set of highly distinctive acoustic patterns (i.e., words) that can be identified and responded to very rapidly by human listeners. When interest is directed to prosodic attributes of speech and some of the properties related to source characteristics, it immediately becomes apparent that an even richer and more distinctive set of complex signals can be generated by the combination of only a small number of variations on a larger set of perceptually relevant dimensions.

As Pollack (1952) demonstrated over thirty years ago, speech sounds represent a class of signals that are able to transmit relatively high levels of information with only gross variations in perceptually distinctive acoustic attributes. In other words, speech is an efficient signaling system because of its ability to exploit fundamental processing strategies of the auditory system. This theme has been taken up and expanded recently by Stevens (1980), who argues that speech signals display a certain set of general properties that set them apart from other signals in the listener's auditory environment. According to Stevens, all speech signals have three general properties or attributes in common. First, the short-term power spectrum sampled at specific points in time always has "peaks" and "valleys". That is, speech signals display up and down alternations in spectrum amplitude with frequency. These peaks in the power spectrum arise from the peaks observed in the vocal tract transfer function and correspond to the formants or vocal resonances that are so prominent in vowel and vowel-like sounds. The second general property that speech sounds display is the presence of up and down fluctuations in amplitude as a function of time. These variations in amplitude correspond to the alternation of consonants and vowels occurring in syllabic-like units roughly every 200-300 msec. Finally, the third general property that speech signals display is that the short-term spectrum changes over time. The peaks and valleys of the power spectrum change; some changes occur rapidly, like the formant transitions of stop consonants,

whereas other changes are more gradual, like the formant motions of semi-vowels and diphthongs. According to Stevens (1980), speech sounds have these three general attributes and other sounds do not, and it is these attributes that distinguish speech sounds from other complex nonspeech sounds.

It should also be mentioned here in addition to some of the differences in the signal characteristics between speech and nonspeech noted above, there are also very marked differences in the manner in which speech and nonspeech signals are processed (i.e., encoded, recognized and identified) by human listeners. For the most part, research over the last thirty-five years has demonstrated that when human observers are presented with speech signals, they typically respond to them as linguistic entities rather than simply as random auditory events in their environment. The set of labels used in responding to speech is intimately associated with the function of speech as a signalling system in spoken language. Thus, speech signals are categorized and labeled almost immediately with reference to the listener's linguistic background and experience. And, a listener's performance in identifying and discriminating a particular acoustic attribute is often a consequence of the functional role this property plays in the listener's linguistic system. It is possible to get human listeners to respond to the auditory properties of speech signals with some training and the use of sensitive psychophysical procedures. But one of the fundamental differences between speech and nonspeech signals lies in the linguistic significance of the patterns to the listener and the context into which these patterns may be incorporated.

In the sections below, we discuss several recent findings that deal with the dynamic or time-varying aspects of speech perception. The topics to be considered in this paper include findings on trading relations, perceptual integration of acoustic cues, and context effects. The findings from these studies point to significant gaps in our current understanding of the perception of speech and nonspeech sounds in isolation and in context. At the present time, we do not have a psychophysics of speech, nor do we have a psychophysics of complex sounds. Current theoretical efforts represent only a very meager beginning and, in some cases, an unsatisfactory attempt to understand a wide variety of phenomena in the field of speech perception. Our discussion of these selected topics in speech perception is designed to emphasize the wide separation that currently exists between researchers working in the mainstream of speech perception and those attempting to develop a psychophysics of speech and other complex sounds. It is hoped that this presentation will generate a great deal of discussion at the workshop about future directions for research and theory in speech perception and the research goals of investigators who are currently interested in the psychophysics of both speech and nonspeech signals.

Cue Trading and Acoustic Cue Integration. It has been well-known for many years that several cues may signal a single phonetic contrast (e.g., Delattre, Liberman, Cooper & Gerstman, 1952; Denes, 1955; see Repp, 1982, for a review). Thus, it is possible to demonstrate that when the perceptual utility of one cue is attenuated, another cue may take on primary effectiveness in signalling the contrast under scrutiny because both cues, it is assumed, are equivalent. This is called a phonetic trading relation (Repp, 1982). In

recent years, phonetic trading relations have been cited as evidence for a specialized speech mode of perception. There appear to be two reasons for this view. First, some demonstrations of phonetic trading relations involve both spectral and temporal cues that are distributed over a relatively long temporal interval. Repp (1982) has argued that it is hard to imagine how such disparate cues arranged across relatively long time windows could be integrated into a unitary percept if specialized (i.e., non-auditory) processes were not in operation. Repp proposes, furthermore, that the basis of this specialization lies in the listener's abstract knowledge of articulation. In other words, because we as listeners know (implicitly) how speech is produced, we are able in some way to integrate acoustically different cues that arise from an articulatory plan into a single unified phonetic percept. The second line of evidence for specialization of speech perception involves demonstrations that phonetic trading relations do not apparently arise for nonspeech sounds. Such evidence is therefore taken to be proof that the integration of multiple cues giving rise to trading relations is somehow peculiar to processing speech signals.

One frequently investigated trading relation involves the so-called stop manner contrast in word pairs such as "say"- "stay" or "slit"- "split." The presence or absence of a stop in such minimal pairs may be signalled by one of two cues: (1) silent closure duration between the offset of /s/ frication and the onset of voicing, and (2) the first formant transition onset. Fitch, Halwes, Erickson, and Liberman (1981) examined the degree to which these two cues are phonetically equivalent in perception. A demonstration of the phonetic equivalence of these two diverse cues would suggest the operation of specialized processes that "ignore" the acoustic diversity of these cues and integrate them into a unitary phonetic percept.

Fitch et al. synthesized two syllables, one having formant transitions biasing perception of the syllable /lIt/ and another the syllable /plIt/. /s/-frication was appended to the beginnings of each syllable and two series of stimuli were generated by varying the closure interval between the /s/ frication and the vocalic portion of the syllable. One series of stimuli was thus composed of /s/ + /lIt/ and another of /s/ + /plIt/, with both series varying in the duration of the closure interval. Fitch et al. presented these sets of stimuli to subjects for identification. For both series, stimuli with sufficiently short closure durations were heard as /sllt/ and stimuli with sufficiently long closure durations were heard as /spllt/. Thus, Fitch et al. demonstrated that in spite of the formant transitions, the duration of the closure interval could induce identification of the stimuli from both series as either /sllt/ or /spllt/. However, their results also showed that, on the average, relatively more silence (approximately 20 ms) was required for identification of /spllt/ for the /s/ + /lIt/ series than for the /s/ + /plIt/ series. These findings demonstrate that formant transition cues and closure duration trade off in producing perception of the presence or absence of the stop /p/.

To determine more precisely if formant transitions and closure duration are perceptually equivalent, Fitch et al. carried out a second experiment on the discrimination of /sllt-spllt/ stimuli containing either only one cue, two cooperating cues, or two conflicting cues. The logic behind this experiment was as follows: If formant transitions

and closure duration are equivalent, their perceptual effects should be additive. Thus, relative to a baseline condition, adding a cooperating cue should enhance discriminability, whereas adding a conflicting cue should decrease discriminability due to the fact that the perceptual effects of these two cues should cancel one another out. This result is precisely what Fitch et al. found. Discrimination was best for the cooperating cue stimuli, intermediate for the single cue stimuli, and worst for the conflicting cue stimuli.

To further buttress the claim that phonetic trading relations (and the concomitant notion of phonetic equivalence) are peculiar to speech processing, Best, Morrongiello, and Robson (1981) performed an experiment using sine-wave analogs of "say" and "stay," a contrast for which they demonstrated a similar trading relation to that of "slit"- "split." [Sine-wave analogs were constructed by imitating the center frequencies of formants of natural speech tokens with pure tones (Remez, Rubin, Pisoni, and Carrell, 1981).] Two versions of stimuli were constructed: In one, the sine-wave portion of the stimulus had a low onset of the lowest tone (simulating /dɛl/ or, in Best et al.'s terms, "strong" [dɛl]); in the other version there was a high onset of the lowest tone (simulating /eɪl/ or "weak" [dɛl]). Noise was then appended to the beginning of each stimulus to simulate /s/-frication, and test continua were generated by varying the closure interval.

Best et al. presented these stimuli to subjects for identification using an AXB procedure. In this procedure, A and B are endpoints of the continuum and X is more like A or B. According to post-hoc interviews the subjects were partitioned into two groups, "speech" listeners and "nonspeech" listeners. [For sine-wave stimuli modelled after natural speech, some listeners spontaneously hear the stimuli as speech (although somewhat unnatural speech). Other listeners, however, hear the stimuli as nonspeech whistles (see Remez et al., 1981).] Identification functions for the "speech" or "say"- "stay" listeners revealed a trading relation; those who failed to hear the stimuli as speech, however, failed to display identification functions indicative of a trading relation. In addition, the subjects who heard the stimuli as nonspeech were further subdivided into two groups, one group which attended to spectral cues (i.e., onset frequency of the lowest tone) and one which attended to temporal cues (duration of the closure interval). Thus, the nonspeech listeners were unable to trade the two cues and attended to either the spectral cue or the temporal cue. Apparently, subjects who heard the stimuli as speech perceived the stimuli in a phonetic mode in which the temporal and spectral cues were somehow integrated into a unitary percept, thus giving rise to the observation of a trading relation; those subjects hearing the stimuli as nonspeech were presumably perceiving the stimuli in an auditory mode in which integration of the two cues was impossible.

The demonstration of trading relations constitutes the newest source of evidence for the existence of a specialized speech mode in which knowledge of articulation comes to bear on the perception of speech. According to Repp (1982), "trading relations may occur because listeners perceive speech in terms of the underlying articulation and resolve inconsistencies in the acoustic information by perceiving the most plausible articulatory act. This explanation requires that the listeners have at least a general model of human vocal tracts and of

their way of action" (p.95). Thus, based in part on demonstrations of phonetic trading relations, researchers, particularly those associated with Haskins Laboratories, have once again renewed their efforts to argue for articulation-based specialized phonetic processing. It is not clear, however, that such a position is entirely unassailable.

Massaro and his colleagues (Massaro and Oden, 1980; Oden & Massaro, 1978; Massaro & Cohen, 1977) offer an alternative account of trading relations that explicitly denies any specialized processing. Instead, in their model, speech perception is viewed as a "prototypical instance of pattern recognition" (Massaro and Oden, 1980, p. 131). Briefly, Massaro and Oden argue that multiple features corresponding to a given phonetic contrast are extracted independently from the waveform and then combined in the decision processor according to logical integration rules. These rules operate on fuzzy sets so that information regarding a given feature may be more-or-less present or "sort of" present. This aspect of their model, then, stresses continuous rather than all-or-none information. Thus, features are assigned a probability value between 0.0 and 1.0 indicating the extent to which a given feature is present. Subsequently, the degree to which this featural information matches a stored prototype is determined according to a multiplicative combination of the independent features. The fact that multiple features are evaluated independently, and that these features can assume ambiguous values (eg., 0.5), can account for the finding that the perceptual utility of two cues may trade off in rendering a given phonetic percept.

Although Massaro's model can handle phonetic trading relations without reference to articulation or specialized phonetic processing, the results of Best et al. demonstrating cue trading for speech stimuli but not for sine-wave analogs of speech present a problem. If speech perception is simply a prototypical example of pattern recognition, why are some patterns (eg., speech) processed differently than other patterns (eg., sine-wave analogs)? One additional, reasonable assumption can be invoked to account for the speech-nonspeech findings, namely that experience with speech stimuli sensitizes the listener to the existence of many possible redundant cues to a given phonetic contrast. For speech stimuli, then, the listener is biased toward evaluation and integration of all possible cues. For sine-wave analogs, with which the listener has presumably had little experience, the listener may hold no expectations of the possible dependencies among cues. Thus, the absence of expectations of cue dependencies for nonspeech stimuli may have produced the differences in cue trading effects of speech and nonspeech observed in the Best et al. study. Indeed, the very fact that some listeners in this study attended to spectral characteristics of the stimuli and others to temporal characteristics attests to the fact that both cues were available to the nonspeech listeners. However, because these subjects did not treat these stimuli as speech-like, they may not have applied certain overlearned strategies for evaluating and integrating diverse cues to the sine-wave analogs (see Grunke & Pisoni, 1982; Schwab, 1981).

Repp, Liberman, Eccardt, & Pesetsky (1978) dismiss a similar explanation of cue trading and integration on a priori grounds. However, Repp (1983b) has recently argued that cue trading results do not in fact support the claim that speech is processed by specialized mechanisms. Much in the spirit of our discussion of the cue trading

literature, Repp concludes that cue trading effects "are not special because, once the prototypical patterns are known in any perceptual domain, trading relations among the stimulus dimensions follow as the inevitable product of a general pattern matching operation. Thus, speech perception is the application of general perceptual principles to very special patterns" (p. 132).

In short, we believe, as was shown several years ago with categorical perception effects, that reasonable alternative explanations are possible for the cue trading evidence reported thus far. Whether these explanations will stand the test of time is, of course, an empirical question. The arguments proposed for the existence and operation of specialized speech processing mechanisms and the mediation of articulation in speech perception have quite broad implications for linguistics, psychology, and the philosophy of mind. Thus, because of the important ramifications of these claims, it is probably best to err on the side of caution in evaluating the available evidence. The cue trading evidence is one of the most compelling sources of evidence to date for a speech mode of perception. However, the historical lesson taught by the failures of previous lines of research (e.g., categorial perception, selective adaptation) to demonstrate specialized speech processing emphasize the importance of maintaining a healthy skepticism in evaluating any new evidence for specialization of speech processing with reference to articulatory mediation. In short, we are sympathetic to the position being advocated here, but we are not yet convinced from the experimental data used to support these claims.

Context Effects in Speech Perception. Much, if not all, of the research on speech perception over the last thirty-five years has been concerned with the minimal cues, features, or acoustic attributes that support perception of segmental phonetic contrasts in highly restricted environments (e.g., CV syllables). Although this reductionism has made scientific investigation more tractable, it has led many researchers to ignore, or at least postpone consideration of the perceptual problems posed by the production of speech in the context of fluently articulated sentences or passages of connected discourse. At the level of fluent continuous speech, the problems of invariance and segmentation appear to become even more imposing. Not only are segments coarticulated within syllables in continuous speech, but coarticulatory effects are spread across words, making isolation of words within sentences a seemingly insurmountable task for the listener. In addition, many suprasegmental effects found in continuous speech introduce other sources of variability that need to be accounted for in the perceptual process. For example, phrasal and sentential contexts introduce variations in fundamental frequency, stress placement and timing, and duration, all of which fall under the rubric of "prosodic phenomena." Finally, context effects produced by differences in speaking rate and speaker characteristics (e.g., sex, age, dialect) introduce further sources of variability that affect the acoustic-phonetic encoding of the linguistic message in the speech waveform.

Traditionally, context-conditioned variability has been viewed as a source of "noise" in the acoustic signal from which phonetic segments are extracted. Recently, however, research efforts have

focused on discovering the systematic effects of context, giving rise to the notion of "lawful variability" (Elman & McClelland, 1983). Basically, this conception of contextually-conditioned variability treats context effects as sources of important acoustic-phonetic information rather than simply noise in the signal (Church, 1983; Elman & McClelland, 1983; Nakatani & O'Connor-Dukes, 1980). The notion of "lawful variability" stems from a number of diverse demonstrations of the orderliness and predictability of context effects in the production and perception of speech. What was once thought to be noise that must be filtered out in recovering phonetic segments from the waveform is now coming to be thought of as a source of useful information arising from systematic, rule-governed contextual effects.

Inherent in the idea of "lawful variability" is the growing tendency to view the speech waveform as a rich source of acoustic-phonetic information. Previously, it was thought that acoustic-phonetic information was so impoverished that higher-levels of knowledge must continually be brought to bear on the perception of speech. Recent approaches that take advantage of rule-governed variability, however, emphasize the richness and informativeness of the acoustic-phonetic information in the waveform. Thus, a number of researchers have begun to advocate more bottom-up approaches to the speech perception process, an understandable turn of affairs in light of the claim that "rather than a bane, phonetic variability may be a boon in speech perception" (Nakatani and O'Connor-Dukes, 1980, p. 13).

In the sections below, we discuss four types of context effects that have been of recent interest: (1) local phonetic context, (2) phonological and lexical context, (3) phrasal and sentential context, and (4) speaking rate. Each of these areas has proven to be highly amenable to experimental investigation and has advanced our knowledge considerably about systematic context effects in speech perception.

Local Phonetic Context Effects. One of the most pervasive effects of local phonetic context is that of allophonic variation. Allophonic variation refers to the fact that a given phoneme may have many different acoustic-phonetic realizations, depending on the context in which it is produced. For example, a /t/ in syllable-initial position, such as in /tEd/, is aspirated in English (i.e., accompanied by a short burst of noise associated with release). However, a /t/ occurring in syllable-final position is rarely released (e.g., in /bEt/), and a /t/ occurring in the cluster /st-/ is never aspirated (e.g., in /stap/). All three phonetic realizations of [t] are said to be allophones of the phoneme /t/. Although [t]'s occurring in clusters and in syllable-final position have acoustic attributes different from [t]'s occurring in syllable-initial position, we nevertheless perceive every phonetic realization of a [t] as the phoneme /t/.

A number of years ago, Nakatani and his colleagues (Nakatani and Dukes, 1977; Nakatani and Schaffer, 1978; Nakatani and O'Connor-Dukes, 1980) and Church (1983) proposed that allophonic variation should be viewed as a source of information in parsing words and syllables in sentences (see also Oshika et al., 1975). Consider the following phonetic transcription of the question "Did you hit it to Tom?", discussed by Klatt (1977):

[dɪʃəfɪtɪtam]

This transcription is meant to represent a "normal" articulation of the question in fluent casual speech. A spectrogram of this utterance is shown in figure 1 along with spectrographic representations of the same words produced in isolation. As is apparent from this example, the "ideal" (or citation) forms of the words, [dId], [yu], [hlt], [It], [tu], [tam], undergo many phonetic changes when produced in sentential context. These changes, if viewed simply as noise imposed on the canonical phonetic transcriptions of the words, would appear to make the lexical retrieval process quite difficult for the listener. In particular, where does one word end and another begin? An analogous situation in printed text would arise if the spaces were removed from a sentence, such as "CATSATEASEARERARELYEARNESTOPPONENTS" ("Cats at ease are rarely earnest opponents").

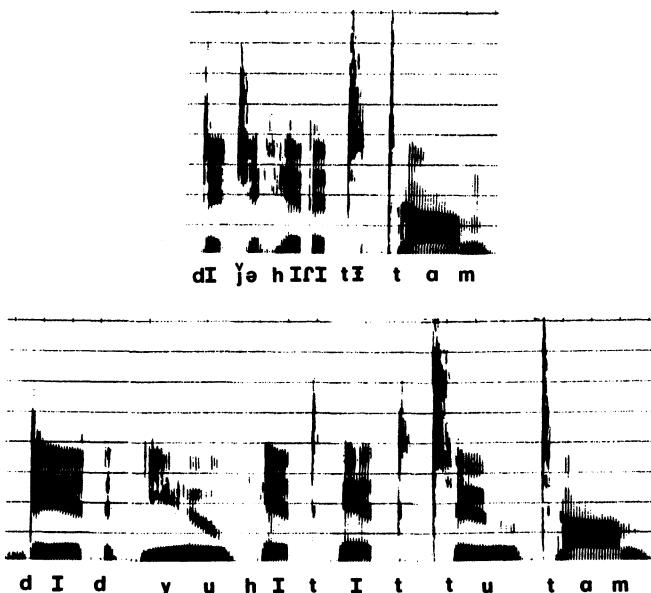


FIGURE 1.

The problem of parsing "Did you hit it to Tom?" into its constituent lexical items may be overcome, however, by appealing to at least two sources of information: allophonic variation and phonological constraints (Church, 1983). Church, building on the earlier work by Klatt (1977), points out that five allophonic rules are operative in the example: (1) /d/ before /y/ in "did you" palatalizes, rendering /dɪʃy/; (2) unstressed /u/ reduces to /ə/ in "you", rendering /dɪʃə/; (3) intervocalic /t/ in "hit it" flaps, rendering /hɪʃɪt/; (4) /u/ in "to" reduces and devoices, rendering /tɪ/; and (5) /t/ in "it to" geminates, rendering /tɪt/. Thus, many of the phonetic changes observed in sentential context are highly predictable, and thus highly informative, if one assumes that the listener has access to implicit knowledge

concerning the way allophonic variations operate. Applying these five allophonic rules, we can recover a great deal of information about the underlying phonemic representation of the sentence:

[d̪jəh̪ʃt̪it̪tam] becomes /d̪idyuh̪t̪ittutam/

Another source of constraint pointed out by Church is imposed by the operation of phonological rules. If we expand, as Church suggests, the original transcription to include the presence of aspiration and glottal stops, and subsequently apply a few general phonological rules, hypothesization of lexical items is further simplified. Including aspiration and glottalization renders the following transcription:

[d̪jəh̪ʃt̪iθiθam]

Syllable boundaries are now predictable given the following four phonological rules: (1) /h/ always occurs in syllable-initial position, (2) [ʃ] always occurs in syllable-final position, (3) [?] always occurs in syllable-final position, and (4) [θ] always occurs in syllable-initial position (Church, 1983). (As Church points out, [θ] may be found in syllable final position, although aspiration in syllable-final position is very different from that in syllable-initial position.)

Applying the rules governing allophonic variation to recover the underlying phonemes and the phonological rules to identify syllable boundaries, we obtain the following transcription (syllable boundaries indicated by a ≠):

/d̪idyu≠h̪t̪i≠t̪i≠t̪u≠t̪am/

It is clear from this one example that exploiting the information in a relatively fine-grained phonetic transcription allows recovery of a great deal of information concerning syllable boundaries and the underlying phonemic representations. By inference, it seems reasonable to suppose that the listener makes use of his implicit knowledge of allophonic variation and phonological rules to parse continuous speech into words. What is perhaps most compelling, however, is the degree to which the acoustic cues to allophones and syllable boundaries are differentially encoded in the signal. By treating the manifestations of allophonic variation and phonological processes in the speech waveform as sources of important information in the signal rather than simply noise, we are able to take advantage of the systematic variability contained in the speech waveform. Moreover, we are able to focus more closely on acoustic-phonetic information in resolving ambiguities, rather than having to appeal to higher level knowledge sources (such as syntax and semantics) for "hypotheses" about what may or may not be present in the signal (Chomsky & Halle, 1968). In short, such an approach to speech perception emphasizes the richness of the speech signal instead of touting the impoverished and highly variable nature of the acoustic cues to phonetic segments and word boundaries.

One of the first researchers to advocate the notion that allophonic variation aids in parsing lexical items in sentences was Nakatani (Nakatani and Dukes, 1977). Nakatani and his colleagues have conducted a number of important perceptual and acoustic studies aimed

at identifying cues to word juncture in order to specify how allophonic variation as well as prosodic information are used in identifying the beginning and ends of words. Nakatani and Dukes (1977) examined possible allophonic cues to word juncture in pairs of words such as "no notion" and "known ocean." Such words pairs are phonetically identical except for the locus of the word juncture (see also Bolinger & Gerstman, 1957). Nakatani and Dukes excised portions of word pairs that immediately preceded and followed the word juncture and cross-spliced these excised portions between words in a pair. They then presented the spliced and original versions of these word pairs to subjects for identification. In this way, Nakatani and Dukes were able to determine whether the offset of the first word, the onset of the second word, or both contributed to perception of a word boundary.

Nakatani and Dukes' results showed that word junctures were almost entirely cued by the onset of the second word in the pair, except for words ending in /r/ or /l/. Because /r/ and /l/ have distinctly different allophones at the beginnings and endings of words, these allophones constituted strong cues for word juncture both word-initially and word-finally. In addition, Nakatani and Dukes found that allophonic variations at the beginning of the second word in the pair provided cues to word juncture even in the absence of /r/ or /l/. In particular, they found that glottalization and/or laryngealization cued word junctures when the second word began with a vowel. Finally, they showed that aspiration of voiceless stops, which is most evident for word initial allophones, aided in identifying word junctures.

The findings of Nakatani and Duke provide strong empirical support for Church's claim that allophonic variation is an important source of information in segmenting words in sentences. In another study of word juncture cues, Nakatani and O'Connor-Dukes (1980) extended their previous experiment to include a number of other allophonic and segmental differences that cue word juncture. They found: (1) that gemination of consonants helped to distinguish such pairs as "drunk converse" (in which a doubling or gemination of the /k/'s is present) and "drunken verse" (in which no gemination occurs); (2) that flapped apical stops distinguish pairs such as "hardy feat" (which contains a flapped /d/) and "hard defeat" (which contains the geminate /d+d/); (3) that the presence of a syllabic /n/ distinguishes pairs such as "maiden forced" (which contains a syllabic /n/) and "maid enforced"; (4) that deletion of the unstressed vowels in words such as "bakery," pronounced "bakry," distinguishes word pairs such as "bakery guarded" and "bake regarded"; and finally (5) that vowel reduction in prefixes such as "de-" distinguishes word pairs such as "hard defeat" and "hardy feat." In short, Nakatani and O'Connor-Dukes have shown that listeners take much allophonic variation into account in parsing word strings into their constituent lexical items.

Nakatani and Schaffer (1978) and Nakatani and O'Connor-Dukes (1980) have also shown that stress patterns and rhythm can also aid in identifying word boundaries (see also Nakatani, O'Connor & Astor, 1981). Using reiterant speech, which preserves prosodic information but eliminates allophonic variation and other segmental differences (see Liberman and Streeter, 1978), these researchers demonstrated that listeners could correctly parse reiterant speech versions of adjective-noun phrases such as "malformed nose" and "long stampede."

Taken together, these studies demonstrate that much information resides in the speech signal that can significantly affect segmentation of sentences into words. What was considered by some to be noise and random variation (e.g., allophonic variation) appears to have quite important ramifications for the identification of words in fluent speech from the bottom-up analyses of the speech waveform.

Phonological and Lexical Context Effects. We have seen how local phonetic context may serve to guide a listener's parsing of sentences into words. We now turn to the issue of how somewhat higher level linguistic constraints can influence a listener's perception of phonetic segments. In the preceding section, we suggested that knowledge of phonological rules may aid the listener in recovering underlying phonemic representations and in identifying syllable boundaries. In this section we turn to a somewhat more abstract role of phonology in speech perception, namely the role of knowledge of phonologically permissible sequences in speech perception (sometimes called phonotactics). In addition, we examine some evidence that relates to the effects of lexicality on the perception of phonemes. Both phonological and lexical context effects illustrate the degree to which the listener's knowledge of permissible sound sequences and words in the lexicon influence his or her perception of phonetic segments.

Massaro and Cohen (1983) have recently reported the results of an experiment aimed at evaluating the degree to which phonological context can affect listeners' perception of phonemes. In one of their conditions, Massaro and Cohen generated a synthetic continuum ranging from /ri/ to /li/. The manipulation of crucial interest was the consonant preceding the /ri-li/ syllables. Massaro and Cohen placed each of the /ri-li/ stimuli after one of four consonants: /p/, /t/, /s/, and /v/. In English, both /ri/ and /li/ are permissible after /p/; only /ri/ is permissible after /t/; only /li/ is permissible after /s/; and neither /ri/ nor /li/ are permissible after /v/. Massaro and Cohen were interested, in part, in determining if phonological context (permissible and non-permissible) would affect subjects' labelling of the /ri-li/ continua. In particular, they hypothesized that more /r/ responses would be obtained for the /tri-tli/ continuum and more /l/ responses for the /sri-sli/ continuum.

As predicted, Massaro and Cohen found that phonological context did affect listeners' labelling of the stimuli. Their subjects produced more /r/ responses than /l/ responses in the context of /t/ and more /l/ responses than /r/ responses in the context of /s/. The identification functions for the /pri-pli/ and /vri-vli/ continua fell between the two other functions, as expected. Massaro and Cohen furthermore showed that their effect was in fact due to phonological context and not to auditory interactions between the initial stops and the following /ri/ or /li/ syllables.

In a similar experiment, Ganong (1980) examined the effect of lexical context on the identification of word-initial stops. Ganong varied the VOT of word-initial stops to generate continua ranging from a word to a nonword (eg., "dash" to "tash") and from a nonword to a word (eg., "dask" to "task"). He then presented these stimuli to subjects for identification. Ganong found that lexicality (i.e., whether the stimulus was perceived as a word or a nonword) strongly affected

subjects' labeling of the word initial stop. Subjects produced more "dash" responses for the "dash"- "tash" continuum and more "task" responses for the "dask"- "task" continuum.

The Massaro and Cohen and Ganong studies both demonstrate the effects linguistic knowledge can have on the categorization of speech sounds. These results show that perception of phonetic segments is heavily influenced by what listeners know about permissible sequences of speech sounds in English and by their knowledge of words in the lexicon. Thus, phonological and lexical context further serve to constrain the perceptual analysis of the speech signal. These studies, in conjunction with those by Nakatani and his colleagues, also demonstrate that relatively early in the perceptual processing of speech, many ambiguities may be resolved by employment of allophonic rules, phonological rules, and lexical constraints. We view these studies as important new demonstrations of the influences of phonological and lexical context on the perception of phonetic segments. This work furthermore represents a sharp departure from the earlier views that assumed the speech perception process was strongly driven by top-down knowledge of syntax and semantics.

Sentence-level context effects. Thus far, we have discussed the systematic variability of phonetic segments in various local phonetic environments and how the listener's knowledge of allophonic variation, phonological rules, and lexical items may serve to support our perception of phonemes and words. Another source of systematic variation is introduced, however, when our attention is focused beyond the phonetic segment or word to the study of speech produced in sentential contexts. At this level of analysis, sentence-level effects come into play. We use the term "sentence-level context effects" to refer to those changes in the acoustic-phonetic structure of speech that arise not from the effects of the articulation of adjacent segments, but from the production of fluent speech in sentences.

One of the most widely studied effects of sentence-level context concerns the changes in fundamental frequency (F0), duration, and amplitude that occur at phrase boundaries. The presence of a major syntactic boundary (eg., the boundary between an initial subordinate clause and a main clause) may be signalled by any one of a number of possible cues: a marked fall in the slope of F0 preceding that boundary (Cooper & Sorenson, 1981; Maeda, 1976; Pierrehumbert, 1979), a "resetting" of F0 following the boundary (Cooper & Sorenson, 1981; Maeda, 1976), a pronounced lengthening of segments immediately preceding a boundary (Klatt, 1975; Oller, 1973; Luce & Charles-Luce, 1985), a decrease in amplitude at the boundary (Streeter, 1978), and a pause at the boundary (Goldman-Eisler, 1972). The question addressed by many of the studies investigating syntactic boundary phenomena concerns which of these cues are most important for the listener in identifying phrase boundaries.

Much of the work on the perception of phrase boundaries has employed ambiguous utterances that are manipulated in such a way as to allow assessment of a single potential cue (eg., Lehiste, 1973; Lehiste, Olive and Streeter, 1976; Lehiste, 1983). The most comprehensive of these studies was performed by Streeter (1978). Streeter examined the relative importance of phrase-final lengthening, F0 declination, and changes in amplitude for the identification of

phrase boundaries in ambiguous algebraic expressions. In one condition, Streeter electronically manipulated duration, F0, and amplitude of utterances of the phrase [A plus E times O], which may be read as [(A plus E) times O] or [A plus (E times O)], and required subjects to identify which of the two possible readings was intended for a given stimulus. She found that both duration and F0 served to cue phrase boundaries, whereas amplitude had little effect. (See Luce and Charles-Luce (1983) for similar findings obtained from a reaction time task.) Moreover, she found that duration and F0 were additive, not interactive, cues. Streeter's study thus demonstrates that changes in F0 and duration induced by the presence of a phrase boundary are important independent cues for listeners in the identification of phrase boundaries.

Cutler (Cutler and Darwin, 1981; Cutler and Foss, 1977) has examined the extent to which prosodic information enables the listener to predict where sentence stress will fall. Because sentence stress is usually placed on words of primary semantic importance in a sentence, the ability to predict sentence stress would presumably enable the listener to focus in on those words in a sentence most crucial to the message. Thus, prosodic variations may help direct the listener to high information centers in fluent speech.

It is clear that listeners rely on variability introduced by suprasegmental context effects to extract syntactic and semantic information from the speech waveform. Thus, duration and pitch changes caused by the occurrence of syntactic boundaries and by sentence stress placement provide valuable information for the parsing and comprehension of sentence-length utterances. Moreover, it is clear that a listener's processing of prosodic information is quite complex, in that no single cue has yet been shown to be necessary in identifying phrase boundaries or stress placement (see Cutler & Ladd, 1983). Although the recent interest in the role of prosody in speech perception is certainly a welcome trend, much work is needed to specify more precisely how the listener takes advantage of this obviously important source of information in the perception of fluent speech.

Effects of Speaking Rate. One final effect of suprasegmental context that deserves discussion is that induced by changes in speaking rate. Effects of speaking rate on the perception of speech are not, in the strict sense of the term, "prosodic" effects. Instead, the issue of the effects of speaking rate on the perception of speech relates to the issue of "perceptual normalization." Just as we may ask how a listener compensates or normalizes for the acoustic consequences of changes in the vocal tract sizes of different speakers, we may also ask how the listener normalizes for changes in speaking rate (within and between speakers). In short, how do listeners normalize for speech produced in the context of many different speaking rates?

In a comprehensive review of the effects of "global" speaking rate on the production and perception of phonetic segments, Miller (1981) discusses a number of changes at the phonetic level induced by changes in speaking rate. For vowels, both spectral and durational changes may be observed as speaking rate is increased. In particular, vowels tend to reduce at faster rates of speech so that target formant

frequencies are rarely achieved. For consonants, cues to voicing of syllable-initial (VOT) and intervocalic stops (closure duration) undergo systematic changes as speaking rate is speeded or slowed. In addition, manner class distinctions between consonants are likewise affected by changes in speaking rate.

One of the most interesting findings concerning the effects of speaking rates on the perception of segmental contrasts concerns voicing of stop consonants in syllable-initial and intervocalic position. In a series of studies, Summerfield (1974, 1975a, 1975b; Summerfield & Haggard, 1972) examined the effects of speaking rate on the identification of stimulus continua varying along the dimension of VOT. He found that the rate of articulation of the carrier sentence in which the stimuli were embedded affected the voicing boundaries for the continua in systematic ways. In particular, for a carrier sentence produced at a fast speaking rate, shorter VOT's were required to identify a stimulus as voiceless than when the carrier sentence was produced at a slower speaking rate.

On the basis of these and other studies, it appears that listeners adjust their judgments of phonetic contrasts to compensate for perceived speaking rate. Moreover, the adjustments are highly systematic and predictable. Although it is not the case that all of the effects of speaking rate heretofore demonstrated are so straightforward as those demonstrated by Summerfield (see also Port and Dalby, 1982), it is probably true that the variability introduced by changes in speaking rate are automatically compensated for by the listeners (Miller, Green, & Schermer, 1982) and have highly predictable effects on listeners' perceptions. Unfortunately, we do not have a good theoretical account of these findings yet, nor a deep understanding of the perceptual mechanisms responsible for this form of perceptual compensation (see, however, Pisoni, Carrell & Gans, 1983).

Conclusions. Despite these recent findings and their immediate impact on theoretical efforts in speech perception, there are still very large gaps in our understanding of the auditory/perceptual processing of speech signals by human listeners. In the past, it has been very easy to account for a set of findings in speech perception by appealing to the existence and operation of specialized speech processing mechanisms. Unfortunately, such global explanatory accounts are becoming more and more unsatisfactory as we begin to learn more about the psychophysical and perceptual properties of speech and complex nonspeech signals, and about how the auditory system encodes these types of signals. It is clear to us that theoretical accounts of specific phenomena in speech perception such as trading relations, cue integration, and context effects can no longer be couched in terms of vague descriptions of articulatory mediation by specialized perceptual mechanisms. We have not carried out all the appropriate nonspeech control experiments yet, but we are certain that more precise and testable explanations of these findings will be forthcoming in the years ahead.

REFERENCES

1. Best, C.T., Morrongiello, B. and Robson, R. (1981). Perceptual equivalence of acoustic cues in speech and nonspeech perception. Perception and Psychophysics, 29, 191-211.
2. Bolinger, D. & Gerstman, L.J. (1957). Disjuncture as a cue to constructs. Word, 13, 246-255.
3. Chomsky, N., and Halle, M. (1968). The Sound Pattern of English. New York: Harper and Row.
4. Church, K.W. (1983). Phrase-structure parsing: A method for taking advantage of allophonic constraints. Bloomington, Ind.: Indiana University Linguistics Club.
5. Cooper,W.E.,and Sorensen, J.M. (1981).Fundamental Frequency in Sentence Production. New York: Springer-Verlag.
6. Cutler, A., and Darwin, C.J. (1981). Phoneme-monitoring reaction time and preceding prosody: Effects of stop closure duration and of fundamental frequency. Perception and Psychophysics, 29, 217-224.
7. Cutler, A., and Foss, D.J. (1977). On the role of sentence stress in sentence processing. Language and Speech, 20, 1-10.
8. Cutler, A., and Ladd, D.R. (1983).Prosody: Models and Measurements. New York: Springer-Verlag.
9. Delattre, P.C., Liberman, A.M., Cooper, F.S., and Gerstman, L.J. (1952). An experimental study of the acoustic determinants of vowel color: Observations of one- and two-formant vowels synthesized from spectrographic patterns. Word, 8, 195-210.
10. Denes, P. (1955). Effect of duration on the perception of voicing. Journal of the Acoustical Society of America, 27, 761-764.
11. Elman, J.L., and McLelland, J.L. (1983). Exploiting lawful variability in the speech waveform. Paper presented at the Symposium on Invariance and Variability, M.I.T., Cambridge, Mass.
12. Fitch, H.L., Halwes, T., Erickson, D.M., and Liberman, A.M. (1980). Perceptual equivalence of two acoustic cues for stop-consonant manner. Perception and Psychophysics, 27, 343-350.
13. Ganong, W.F. (1980). Phonetic categorization in auditory word perception. Journal of Experimental Psychology: Human Perception and Performance, 6, 110-125.
14. Goldman-Eisler, F. (1972). Pauses, clauses, sentences. Language and Speech, 15, 103-113.
15. Grunke, M.E., & Pisoni, D.B. (1982). Some experiments on perceptual learning of mirror-image acoustic patterns. Perception and Psychophysics, 31, 210-218.
16. Klatt, D.H. (1975). Vowel lengthening is syntactically determined in a connected discourse. Journal of Phonetics, 3, 129-140.
17. Klatt, D.H. (1977). Review of the ARPA speech understanding project. Journal of the Acoustical Society of America, 62, 1345-1366.
18. Lehiste, I. (1973). Phonetic disambiguation of syntactic ambiguity. Glossa, 7, 107-121.
19. Lehiste, I., Olive, J.P., and Streeter, L.A. (1976). The role of duration in disambiguating syntactically ambiguous sentences. Journal of the Acoustical Society of America, 60, 1199-1202.
20. Liberman, M.Y., and Streeter, L.A. (1978). Use of nonsense-syllable mimicry in the study of prosodic phenomena. Journal of the Acoustical Society of America, 63, 231-233.

21. Luce, P.A. and Charles-Luce, J. (1983). The role of fundamental frequency and duration in the perception of clause boundaries: Evidence from a speeded verification task. Journal of the Acoustical Society of America, 73, S67.
22. Luce, P.A., and Charles-Luce, J. (1985). Contextual effects on vowel duration, closure duration, and the consonant/vowel ratio in speech production. Journal of the Acoustical Society of America, 78, 1949- 1957.
23. Maeda, S. (1976). A characterization of American English intonation. Unpublished doctoral thesis. Cambridge, Mass.: M.I.T.
24. Massaro, D.W., and Cohen, M.M. (1977). The contribution of fundamental frequency and voice onset time to the /zi/-/si/ distinction. Journal of the Acoustical Society of America, 60, 704-717.
25. Massaro, D.W. and Cohen, M.M. (1983). Phonological context in speech perception. Perception & Psychophysics, 34, 338-349.
26. Massaro, D.W., & Oden, G.C. (1980). Speech perception: A framework for research and theory. In N.J. Lass (ed.), Speech and Language: Advances in Basic Research and Practice. Vol. 3. New York: Academic Press, 129-165.
27. Miller, J.L. (1981). Effects of speaking rate on segmental distinctions. In P.D. Eimas & J.L. Miller (Eds.), Perspectives on the Study of Speech. Hillsdale, N.J.: Lawrence Erlbaum Associates.
28. Miller, J.L., Green, K., and Schermer, T. (1982). On the distinction between prosodic and semantic factors in word identification. Journal of the Acoustical Society of America, 71, UU6.
29. Nakatani, L.H., & Dukes, K.D. (1977). Locus of segmental cues for word juncture. Journal of the Acoustical Society of America, 62, 714- 719.
30. Nakatani, L.H., & O'Connor-Dukes, K.D. (1980). Phonetic parsing cues for word perception. Unpublished manuscript. Murray Hill, N.J.: Bell Laboratories.
31. Nakatani, L.H., & Schaffer, J.A. (1978). Hearing "words" without words: Prosodic cues for word perception. Journal of the Acoustical Society of America, 63, 234-245.
32. Nakatani, L.H., O'Connor, K.D., & Aston, C.H. (1981). Prosodic aspects of American English speech rhythm. Phonetica, 38, 84-106.
33. Oden, G.C., & Massaro, D.W. (1978). Integration of featural information in speech perception. Psychological Review, 85, 172-191.
34. Oller, D.K. (1973). The effect of position in utterance on speech segment duration in English. Journal of the Acoustical Society of America, 54, 1235-1247.
35. Oshika, B.T., Zue, V.W., Weeks, R.V., Neu, H., and Aurbach, J. (1975). The role of phonological rules in speech understanding research. IEEE Transactions on Acoustics, Speech, and Signal Processing, ASSP-23, 104-112.
36. Pierrehumbert, J. (1979). The perception of fundamental frequency declination. Journal of the Acoustical Society of America, 66, 363- 369.
37. Pisoni, D.B., Carrell, T.D., and Gans, S.J. (1983). Perception of the duration of rapid spectrum changes in speech and nonspeech signals. Perception and Psychophysics, 34, 314-322.
38. Pollack, I. (1952). The information of elementary auditory displays. Journal of the Acoustical Society of America, 24, 745-749.

39. Port, R.F., and Dalby, J. (1982). Consonant/vowel ratio as a cue for voicing in English. Perception and Psychophysics, 32, 141-152.
40. Remez, R.E., Rubin, P.E., Pisoni, D.B., and Carrell, T.C. (1981). Speech perception without traditional speech cues. Science, 212, 947- 950.
41. Repp, B.H. (1982). Phonetic trading relations and context effects: New Experimental evidence for a speech mode of perception. Psychological Bulletin, 92, 81-110.
42. Repp, B.H. (1983). Trading relations among acoustic cues in speech perception: Speech-specific but not special. Haskins Laboratories Status Report on Speech Research SR-76. New Haven: Haskins Laboratories, 129-132.
43. Repp, B.H., Liberman, A.M., Eccardt, T., and Psestsky, D. (1978). Perceptual integration of acoustic cues for stop, fricative, and affricate manner. Journal of Experimental Psychology: Human Perception and Performance, 4, 621-637.
44. Schwab, E.C. (1981). Auditory and phonetic processing for tone analogs of speech Unpublished doctoral dissertation. Buffalo, N.Y.: State University of New York at Buffalo.
45. Stevens, K.N. (1980). Acoustic correlates of some phonetic categories. Journal of the Acoustical Society of America, 68, 836-842.
46. Streeter, L.A. (1978). Acoustic determinants of phrase boundary perception. Journal of the Acoustical Society of America, 64, 1582- 1592.
47. Summerfield, Q. (1974). Towards a detailed model for the perception of voicing contrasts. In Speech Perception (No. 3). Belfast: Department of Psychology, Queen's University of Belfast.
48. Summerfield, Q. (1975 a). Cues, contexts, and complications in the perception of voicing contrasts. In Speech Perception (No. 4). Belfast: Department of Psychology, Queen's University of Belfast.
49. Summerfield, Q. (1975 b). Information processing analysis of perceptual adjustments to source and context variables in speech. Unpublished doctoral dissertation. Belfast: Department od Psychology, Queen's University of Belfast.
50. Summerfield, Q., and Haggard, M.P. (1972). Speech rate effects in the perception of voicing. In Speech Synthesis and Perception (No. 6). Cambridge: Psychology Laboratory, University of Cambridge.

PERCEPTUAL INTEGRATION OF RISE TIME AND SILENCE IN AFFRICATE/FRICATIVE AND PLUCK/BOW CONTINUA

Peter Howell and Stuart Rosen
University College London

The voiceless affricate/fricative distinction has figured prominently in current theorizing about speech perception. Two classes of perceptual theory have drawn support from it. First are the natural auditory sensitivities theories in which phonemic categories are thought to be based on simple acoustic properties - the rise time of the frication noise in the case of voiceless affricate/fricative (Cutting & Rosner, 1974; Stevens, 1981). Second are the articulatory referential theories which propose that information about how sounds are produced is used during perception (Dorman, Raphael & Liberman 1976).

Evidence for the view that special auditory sensitivities exist for the perception of rise time comes from studies of categorical perception. Cutting and Rosner (1974) reported that a voiceless affricate/fricative contrast varying in frication rise time and duration was categorically perceived with a category boundary at 40 ms. More importantly, a non-speech continuum consisting of sawtooth stimuli varying essentially in rise time alone was also categorically perceived and the boundary occurred at about the same value of rise time as it did with the voiceless affricate/fricative continuum. These results were though to show that rise times of about 40 ms served as a natural, auditorily-determined boundary which was used in speech to achieve a separation of affricates from fricatives (Stevens, 1981). In a series of studies, we have shown that this interpretation is not tenable. For one thing, a boundary of 40 ms does not distinguish affricates from fricatives in real speech (Howell & Rosen, 1983a). Also, neither the non-speech (Rosen & Howell, 1981, 1983) nor speech (Howell & Rosen, 1984) continua are perceived categorically. There is no evidence for a natural boundary in either case.

An explanation of the perception of the affricate/fricative distinction in terms of a natural sensitivity for rise time implies that distinguishing such sounds is a relatively simple auditory process. Others have demonstrated that a combination of several cues is necessary and that rise time alone will not suffice. Dorman, Raphael, and Isenberg (1980), for example, reported that perception of the affricate/fricative contrast depends on the vocalic portion of the utterance, the duration of the closure interval, the presence or absence of a release burst, and the duration of the fricative noise, in addition to the rise time of frication. These results demonstrate that telling affricates from fricatives is more complex than the natural sensitivities accounts allow. Given that there are a number of cues which interact, there are several hypotheses concerning the way this happens. Delgutte (1982) has argued that two of these cues may

interact at the auditory level. Others propose that cues are combined by the listener's use of articulatory knowledge.

An important piece of support for articulatory-referential explanations of the perception of the affricate/fricative distinction is the finding of Dorman et al. (1976), replicated by Repp, Liberman, Eccardt, and Pesetsky (1978). Both studies showed that a longer period of silence before a given duration of noise was needed for an affricate to be reported when the test item was preceded by an utterance spoken at a fast rate than when the precursive phrase was spoken slowly. These data are odd in that it might be supposed that as a sentence is spoken faster, both silence and frication would decrease in direct proportion. Test items preceded by speech spoken at a fast rate ought then to require less silence (and, in fact, noise) for an affricate to be reported.

Repp et al. note that Gay (1978) showed that the duration of all intervals is not reduced equally as speech is spoken at faster rates. The duration of the silence in plosives was affected less than the duration of the surrounding vocalic intervals. Repp et al. argue, following a suggestion in Dorman et al. (1976), that this might explain the anomalous result. They assumed that, as speaking rate increased, the duration of the silent gap associated with the affricate would reduce less than the duration of the fricative noise, as did the silent gap relative to the vowel in Gay's (1978) study. For each of the continua that Repp et al. tested, noise duration was constant and so the noise would take up proportionately more of the sentence as speech rate was increased. The effectively longer noise duration would bias the listener to hearing a fricative, since longer noise durations normally signal fricatives (Gerstman, 1957). Thus, subjects would need proportionately more silence (occurring only with affricates) to offset this bias. In other words, an effective increase in noise duration because of increased speaking rate must be offset by an even bigger increase in the duration of silence that precedes the noise.

Repp et al. consider that such an explanation supports articulatory-referential perception, since: "On that assumption, the boundaries would be set not by the number, diversity or temporal distribution of the cues but by a decision that they do (or do not) plausibly specify an articulatory act appropriate for the production of a single phonetic segment." (p. 622). Though Repp et al. explicitly require articulatory "plausibility" in the stimuli, certain aspects of them are implausible. For instance, it is likely that rise time varies with speech rate, so the constant rise time they used across stimulus sets would make their stimuli implausible as tokens of speech spoken at different rates. Also, frication duration would not stay constant as speech rate varied as in their stimuli, but would probably reduce in duration. Finally, no burst was included in their stimuli, though these almost always occur at the release of an affricate (Howell & Rosen, 1983a; Isenberg, 1978).

There is also a logical problem in the way they have applied their explanation. Though the articulatory-referential account has been presented as an explanation of what a mechanism would make of plausible stimuli, the research on cue combination has been worked the other way round - i.e., what would a mechanism that uses articulatory

knowledge make of a stimulus that a vocal tract is unlikely to produce.

The principal problem for Repp et al.'s explanation is that they had no data that indicated how the duration of silence and frication changed in affricates spoken at different rates. Their account requires that silence is reduced less than frication in duration as speech rate increases. In the absence of appropriate data, Repp et al. relied on Gay's data on plosives, as noted above. However, since their argument involves the duration of intervals within affricates (silence and frication), a more sensible comparison would have been between acoustic intervals within the plosives (say, silence and transitions), rather than between the vowel and part of the preceding plosive. Gay does not report measurements on the transitions in detail, but he does indicate that the relationship Repp et al. would require does not hold. He reports that "... transition time was reduced during fast speech, to about the same degree as that for stop consonant closure, some 5-10 ms" (Gay, 1978, p.225).

It is still, of course, possible that silence in affricates is reduced less than other intervals and, consequently, takes up a bigger proportion of the affricate, as Repp et al., suppose. Presumably it was the wish to get more pertinent data that motivated Isenberg (1978) to measure temporal factors in naturally spoken affricates and fricatives. The corpus consisted of the words "ditch" and "dish" spoken in sentence frames at different speech rates. Isenberg measured the duration of the preceding plosive and vowel, the silent interval (when it occurred), and frication duration. These measurements were then expressed as a proportion of the overall sentence duration, a poor measure since rate can change within a sentence: speakers can speak a sentence fast overall, with local parts spoken slowly and vice versa. Ignoring this point, Isenberg would need to show an increase in the proportion of the sentence taken up by silence relative to frication to support Repp et al.'s argument that the perceptual result is explicable in terms of articulation. An interaction should occur between these intervals and speech rate in an analysis of variance. No such interaction occurred. Though Isenberg claimed support for Repp et al.'s argument based on regression lines fitted to individual subject data, the lack of an interaction in the analysis of variance nullifies this conclusion.

Also, there are other data that flatly contradict the relation required by Repp et al., and sought by Isenberg. Maddieson (1980) measured silence and frication duration of intervocalic voiceless affricates and fricatives in Spanish, English and Italian. He reported that the proportion and duration of silence decreased as speaking rate increased.

To summarize, Isenberg's data supports the articulatory plausibility explanation whilst Maddieson's data indicates the need for an alternative explanation. In order to resolve the discrepancy between these two sets of results, we undertook to measure the duration of silence, rise time and overall duration of affricates and fricatives spoken at different rates (Howell & Rosen, 1983b). Measurements were made on the sentences "I saw a chip/ship in the water" spoken by four speakers, three times each at three different rates (slow, medium, and fast).

The critical data to assess Repp et al.'s explanation derive from the affricates. For these sounds, silence, rise time, and overall duration decreased as rate increased. However, at a fast rate silence takes up a smaller proportion relative to frication, whereas Repp et al.'s explanation requires that silence is a bigger proportion. These results are in line with Maddieson's findings and contrary to those of Isenberg. Isenberg's discrepant results may have been obtained because of the sentence contexts he employed: These were "I meant to say talk ditch/dish fast". In these sentences, the final affricate or fricative is followed by a second fricative. The affricate and fricative or the two fricatives would show considerable coarticulation and could easily lead to errors in measuring the duration of frication, which might be rate-dependent.

These measurements demand a reassessment of Repp et al.'s perceptual experiment: they had argued that longer periods of silence are needed for affricates to be perceived when speech rate increases, because this relationship occurs in the articulation of the sounds, and speech sounds are perceived by reference to articulation. Since the relationship between perception and articulation does not hold, some other explanation must apply.

First, however, an experiment equivalent in all crucial respects to that of Repp et al. was conducted in order to confirm their findings. There are essentially two parts of Repp et al.'s experiment: how the perception of burst of noise as affricate or fricative is affected by the duration of the noise and the duration of a preceding period of silence. Second, how perception of these same stimuli is affected when they occur in sentence frames spoken at different rates.

The sentences "Why don't we say chop/shop again?" spoken by one male speaker were recorded at two different speaking rates. Measurements of the fricative noises were made on the "chop" (affricate) and "shop" (fricative) sentences. The "shop" stimulus spoken at a slow rate was employed as the stimulus to be edited to produce the experimental material. A neutral noise duration (average across affricates and fricatives and across the two speech rates) was calculated. This duration (131.3 ms) was imposed on the slow "shop" by excising a medial portion of the noise. Two other stimulus durations were specified and imposed on the stimulus - 20 ms greater and 20 ms less than this. These three stimuli were inserted into both the sentence frames the "shop"s had occurred in (slow and fast). Eleven different stimuli were produced at each rate and for each duration of frication by inserting a period of silence varying between 0 and 100 ms in 10 ms steps.

The sounds were presented in three blocks in random order to eight listeners. Within each block all sounds had the same frication duration. At each frication duration, the eleven sounds with different amounts of preceding silence at both speech rates were presented ten times each. The listener were asked to indicate whether the test item sounded more like "chop" or "shop".

The results are shown in Figure 1. At the top are the data from the sentences spoken at a slow rate and at the bottom at a fast rate. The ordinates are the percentage of "chop" responses and the abscissae the silent gap duration. The points connected together derive from judgments made about stimuli with the same frication duration (short, medium, long). Each curve from each section of the figure shows that affricate report increases as the duration of the silent gap increases. Moreover, for both sentence rates, affricates are more readily reported to have occurred when the frication duration is short.

The final feature to note is that affricates are more readily perceived for all frication durations at shorter silent gaps when the sentence is spoken slowly than when spoken fast (the ogives in the top part of the figure are shifted to the left in comparison with those at the bottom). Z scores computed from the maximum likelihood estimates of the phoneme boundaries and the corresponding standard

errors, however, show that the difference between slow and fast phrases is only significant when the frication duration is short. Analyses corresponding to these are not reported by Repp et al., though it appears, from the phoneme boundaries reported, that the reverse of this occurred - i.e. there was a bigger effect when frication duration was long. With the exception of this aspect, the results substantiate the findings of Repp et al.

If these relationships are due to processes occurring at an auditory level, then decisions about non-speech analogues might show corresponding differences in the way they are perceived. To check this prediction, the following experiment was conducted.

A non-speech sound (a sawtooth waveform of 100 Hz fundamental frequency) with the same envelope as the sound with the medium duration used in the preceding experiment was substituted for the speech sound. This non-speech sound was inserted into the sentences spoken at two rates and portions of silence varying between 0 and 100 ms (in 10 ms steps) were introduced. These sounds were presented ten times each in random order to eight listeners. The

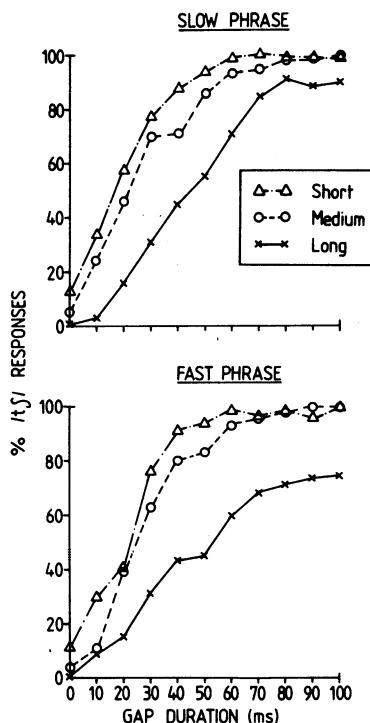


FIGURE 1.

listeners were asked to indicate whether the items sounded like a plucked or bowed string.

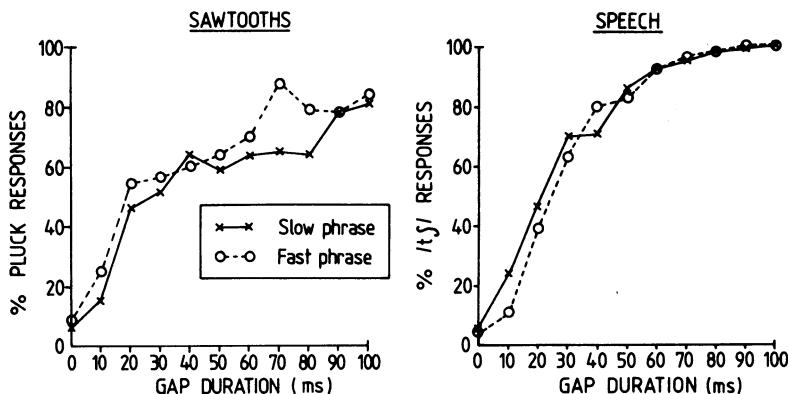


FIGURE 2.

The percentage of pluck responses as a function of silent gap duration are shown in Figure 2 (alongside are shown the corresponding speech data redrawn from Figure 1). The points which are connected together come from conditions which had the sentence at the same rate (the rate can be identified from the symbols and the caption in the inset). It can be seen that at both rates the percentage of pluck responses increases as gap duration increases. The only difference between the speech and non-speech data is that the nonspeech curves are in different order to those with speech, with fewer plucks reported in the slow phrase than the fast. In this case, and unlike the situation with speech, the difference was significant (i.e., Z scores like those computed earlier showed that the category boundaries occurred at a significantly smaller gap duration in the fast phrase [two-tailed]). It is unfortunate that the nonspeech data is only available for one frication duration (this was chosen as a compromise between our own results showing bigger differences at slower rates and those of Repp et al. showing the reverse). Until further data becomes available, it may cautiously be concluded that there is a difference between speech and nonspeech in the rate effect. Clearly, though, perception of pluck/bow and affricate/fricative are both affected similarly by a preceding portion of silence.

The latter data raise as many questions as they answer: first, against the articulatory-referential theory, they show that the effect of gap duration applies to non-speech as well as speech (a similar conclusion follows from the claims of Delgutte, 1982). An auditory explanation of the gap effect might seem appropriate - it is possible that more plucks/affricates are reported for a given rise time when preceded by a silent gap because of the fast-adaptation properties of the auditory nerve (Delgutte, 1982). It would, however, not be possible to account for all of the findings with speech which have been reported by Repp et al., and replicated here on the basis of auditory processes unless additional assumptions are made. In order to account for the rate effects on speech, some forward masking from the diphthongal vowel on the preceding word ("say") would have to be

hypothesized. Moreover, the properties of the vowel would have to change with speech rate, so that more masking occurs when the speech is spoken at faster rates, and so that a longer silent period is needed to offset it.

The two obvious candidates for this are vowel fall time and spectral shape - more abrupt falls would produce more masking than gradual ones and energy closer to the frequency region of the following frication should be more influential in masking. Indeed, rough measurements on our sentence frames supports the view that the fall time of the vowel in "say" is more rapid when the vowel is spoken fast than slow. For the sentence frames employed in the test, the fall times of the vowels were 42 and 85 ms measured from oscillograms (employing a similar procedure to Howell and Rosen, 1983a). The same phenomenon can be seen in van Heuven's (1983) intensity displays (his Figure 6). No noticeable difference in spectral shape of the two vowels preceding the affricate/fricative occurred. Thus, it may be that the fall time of the vowel influences judgments about the following sound.

Though this explanation is appealing, it must be qualified, as the vowel and frication noises fall in such different frequency regions and, therefore, little masking would be expected. This qualification does not apply to the sawtooths where, because of the greater spectral similarity between them and the preceding vowel, masking could potentially occur. Yet, in the nonspeech case at the medium frication duration, the results go in the opposite direction to that predicted by the masking explanation (in other words, the results do not confirm the prediction in the condition where the explanation should apply best). Even so, this seems the most informative direction to progress: we plan to measure fall times of a bigger sample of vowels preceding affricates and fricatives at different rates and set up psychoacoustic tests to see whether the envelope of a preceding sound affects rise time perception.

In summary, the data reported here show that the trading of silence and frication at different rates in affricates and fricatives is inconsistent with Repp et al.'s articulatory account. The influence of a preceding gap on perception of a following sound is similar whether the following sound is a burst of frication noise or a sawtooth with the same envelope. This can be accounted for by an auditory explanation: the puzzle that remains is whether and why the functions relating gap duration and pluck/affricate report differ with speech rate.

REFERENCES

1. Cutting, J.E. and Rosner, B.S. (1974). Categories and boundaries in speech and music. *Perception & Psychophysics*, 16, 564-570.
2. Delgutte, B. (1982). Some correlates of phonetic distinctions at the level of the auditory nerve. In: R. Carlson and B. Granström (Eds.) *The representation of Speech in the Peripheral Auditory System*. Amsterdam North Holland.
3. Dorman, M.F., Raphael, L.J., and Liberman, A.M. (1976). Further observations on the role of silence in the perception of stop consonants. *Haskins Laboratories Status Report*, SR-48, 197-207.

4. Dorman, M.F., Raphael, L.J., and Isenberg, D. (1980). Acoustic cues for a fricative-affricate contrast in word-final position. *Journal of Phonetics*, 8, 397-405.
5. Fitch, H.L., Halwes, T., Erickson, D.M., and Liberman, A.M. (1980). Perceptual equivalence for two acoustic cues for stop consonant manner. *Perception & Psychophysics*, 27, 343-350.
6. Gay, T. (1978). Effect of speaking rate on vowel formant movement. *Journal of the Acoustical Society of America*, 63, 223-230.
7. Gerstman, L.J. (1957). Perceptual dimensions for the friction portions of certain speech sounds. Unpublished doctoral dissertation. New York University.
8. Heuven van, V.J. (1983). Rise time and duration of frication noise as perceptual cues in the affricate-fricative contrast in English. In: M. van den Broecke, V. van Heuven, and W. Zonneveld (Eds.), Sound Structures, 141-157. Foris publications, Dordrecht.
9. Howell, P. and Rosen, S. (1983a). Production and perception of rise time in the voiceless affricate/fricative distinction. *Journal of the Acoustical Society of America*, 73, 976-984.
10. Howell, P. and Rosen, S. (1983b). Closure and frication measurements and perceptual integration of temporal cues for the voiceless affricate/fricative contrast. In: Speech, hearing and language: Work in Progress VCL, 1, 109-117.
11. Howell, P. and Rosen, S. (1984). Natural auditory sensitivities as universal determiners of phonemic contrasts. In: B. Butterworth, B. Comrie, and O. Dahl (Eds.), Explanations of Linguistic Universals, 205-235. The Hague: Mouton.
12. Isenberg, D. (1978). Effect of speaking rate on the relative duration of stop closure and fricative noise. *Haskins Laboratories Status Report*, SR-55/56, 63-79.
13. Maddieson, I. (1980). Palato-alveolar affricates in several languages. *UCLA Working Papers in Phonetics*, 51, 120-126.
14. Repp, B.H., Liberman, A.M., Eccardt, T., and Pesetsky, D. (1978). Perceptual integration of acoustic cues for stop, fricative and affricate manner. *Journal of Experimental Psychology: Human Perception and Performance*, 4, 621-637.
15. Rosen, S. and Howell, P. (1981). Plucks and bows are not categorically perceived. *Perception and Psychophysics*, 30, 156-168.
16. Rosen, S. and Howell, P. (1983). Sinusoidal plucks and bows are not categorically perceived, either. *Perception and Psychophysics*, 36, 233-236.
17. Stevens, K.N. (1981). Constraints imposed by the auditory system on the properties used to classify speech sounds: Data from phonology, acoustics and psychoacoustics. In: T.F. Myers, J. Laver, and J. Anderson (Eds.), The Cognitive Representation of Speech. Amsterdam: North Holland.

REVERSAL OF THE RISE-TIME CUE IN THE AFFRICATE-FRICATIVE CONTRAST: AN EXPERIMENT ON THE SILENCE OF SOUND*

Vincent J. van Heuven

Dept. of Linguistics/Phonetics Laboratory, Leyden University, P.O. Box 9515, 2300 RA Leiden, The Netherlands

1. INTRODUCTION

1.1. Cues underlying the affricate-fricative contrast

Traditionally, linguists distinguish true consonants (or: obstruents) along a three term manner of articulation dimension: stop, fricative, and affricate. In this paper we shall be concerned with the contrast between two of these categories: affricate versus fricative, as in the word pair chop - shop. This contrast has multiple acoustic cues, which are listed in table I:

Table I, acoustic cues involved in the affricate-fricative contrast

- (1) Duration of the preconsonantal vowel (Isenberg, 1978)
- (2) Decay rate of the preconsonantal vowel amplitude (Debrock, 1977)
- (3) Duration of the pre-burst silent interval (e.g., Kuipers, 1955; Truby, 1955)
- (4) Formant transitions of the preconsonantal vowel (Isenberg, 1978; Dorman, Raphael & Isenberg, 1980)
- (5) Rise time of the friction noise amplitude (e.g., Gerstman, 1957)
- (6) Duration of the friction noise (e.g., Gerstman, 1957)
- (7) Rise time of the post-consonantal vowel amplitude (Debrock, 1977)
- (8) Presence/absence of a release burst (Dorman et al., 1980)

Perceptual relevance has not been established for all of these acoustic correlates, let alone in a single experiment, but over the years the research has focused on three of them, and their trading relations: noise amplitude rise time (henceforth: rise time), noise duration (henceforth: duration), and the duration of the silent interval separating the preceding vowel and the friction noise (henceforth: interval). So far the following effects have emerged:

*I thank Peter Vroege for running the experiment described in this paper.

The waveform editing programme for the DEC Micro PDP-11/23 was developed in our laboratory by Ing. J.J.A. Pacilly and Drs. A.F.E. van der Horst.

Table II, Effects of cues involved in the affricate-fricative contrast

	fricative	affricate
rise time	gradual	abrupt
duration	relatively long	relatively short
interval	short (absent)	long

The research on trade-offs between these cues has had a long history, beginning with Gerstman (1957). A re-analysis of his data (van Heuven, 1979) has shown that duration and rise time can be traded only in the very narrow range between 90 and 130 ms noise duration, indicating that duration is the primary cue to the contrast, at least for stimuli with the contrast in initial position.

Tradings involving the interval can only be examined in a context with a segment (typically a vowel) preceding the contrast. Though there are several studies manipulating interval as a single parameter (e.g., Kuipers, 1955; Truby, 1955), trading research has been quite limited. Two-parameter studies involving the interval (against duration) are described by Repp, Liberman, Eccardt & Pesetsky (1978), and by Dorman et al. (1980): interval versus release burst, duration, and formant transitions. Regular trading relations were established for the interval parameter in all of these studies: the effect of a longer interval cueing affricate could be offset by a more fricative-like value for the competing parameter.

Tradings involving the rise time parameter in context have been investigated least of all. Yet, in one study rise time, too, was found to trade regularly with interval (Dorman, Raphael & Liberman, 1979; experiment also described in Dorman et al., 1980). Here the affricate-fricative contrast was examined in word-final position (ditch - dish): the effect of slower rise time cueing fricative could be counteracted by a longer interval (37 vs. 57 ms interval for cross-overs at 0 and 35 ms rise time, respectively).

Van Heuven (1983) studied the effects of rise time and duration in word-initial position (chop - shop) for isolated stimuli, and for the same tokens embedded in a carrier Why don't you say ... again? For isolated words I obtained results that were essentially similar to Gerstman's (1957), with rise time and duration trading in the regular fashion. However, when in try-outs listeners heard the same tokens in context, they reported exclusively affricates for silent intervals exceeding 20 ms. When, in the final experiment, the interval was fixed at an even shorter value of 15 ms, convincing cross-overs were obtained through manipulating rise time and/or duration, but as the results revealed, the contribution of the rise time had been reversed relative to its function in isolated tokens: in this context longer rise time contributed to affricate.

1.2. Possible causes for the reversal

In my (1983) paper I explained this curious reversal of the rise time cue in context as an effect of forward masking. This seemed a reasonable hypothesis, since the pre-burst vowel had been recorded in an utterance preceding a chop token, i.e., with a relatively abrupt intensity off-ramp as is characteristic of pre-stop vowels (Debrock, 1977). Plomp (1964), among others, has shown that the human ear is relatively insensitive to auditory stimulation shortly after the abrupt termination of a high intensity acoustic event. A subsequent low intensity sound will not be heard during this period of masking, or will at least appear weaker than when presented in isolation. The masking period may extend for as long as 250 ms, though the effect rapidly decays over time.

Due to masking, then, our listeners might have heard a brief interval of silence (or reduced energy), suppressing and replacing the low intensity noise onset of the friction sound. The smoother the noise onset, the longer the noise would remain below the masking threshold, creating a longer perceived gap as well as a shorter perceived noise duration, which two illusions then conspired to cue affricate.

However, in spite of the prima facie attractiveness of this account, there are complications that may force us to reconsider. For masking to occur it is necessary that the frequency distribution of the masker (here: vowel) coincides or overlaps with that of the probe (here: friction noise). Thus, in Plomp (1964) pure tones were masked by white noise (see further, Resnick, Weiss & Heinz, 1979). In my vowel-noise sequences it seems unlikely that the masker and probe frequency distributions were sufficiently similar to cause strong masking effects: the vowel /ei/ has most of its energy below 2500 Hz, whereas the /sh/ noise has its energy concentrated above this value (Heinz & Stevens, 1961).

As an alternative explanation for the reversal of the rise time cue in context I now propose the following. Let us assume that it is a necessary condition for listeners to perceive an affricate that the stimulus contain a brief interval of silence (or reduced energy) immediately preceding the friction noise, so as to reflect the presence of an articulatory closure (cf. Dorman et al., 1979). If such a silent interval is absent from the physical stimulus, the listener may have to reinterpret the acoustic signal trying to satisfy the condition for a silent interval. It is conceivable, then, that he will consider the low energy portion in the smooth noise onset to be the silent interval.

Notice that this account is fundamentally different from the masking hypothesis. In the latter case it is assumed that the gradual noise onset is obscured due to a peripheral mechanism whereby the abrupt termination of acoustic energy cannot be resolved by the human ear. In the alternative view the ear is perfectly capable of resolving the intensity envelope of the vowel-friction sequence, but reinterprets the available cues at a more central level. The purpose of the present paper is to choose between these two competing explanations.

1.3. Approach

If the cue reversal is indeed a matter of reinterpreting the available cues so as to perceive a silent interval, it should not matter to the listener whether the low intensity portion of the stimulus is located in the noise onset, or in the offset of the preceding vowel, as long as the energy dip occurs at the VC-boundary, i.e., at a point in time where a stop-closure can be located.

Therefore, changing the vowel offset from abrupt to gradual should increase the number of affricate judgments when the vowel is closely followed by the friction sound; moreover, gradual vowel offset and noise onset should reinforce one another, since both manipulations can be reinterpreted as the silent interval. Masking, on the other hand, should disappear when the vowel intensity decays gradually, and no effect of masking should obtain at all when the vowel off-ramp is longer than some 70 ms (cf. van den Broecke & van Heuven, 1983). As a result, the noise rise time cue should reverse after an abrupt vowel termination, but function normally (i.e., as in stimulus initial position) when the vowel decay is smooth. Thus, varying noise rise time versus vowel decay time in a two-parameter study will allow us to choose between the competing explanations.

2. METHOD

A male native speaker of R.P.-English recorded the sentence Why don't you say shop again observing normal intonation and timing. The audiosignal was digitised (10 kHz, 12 bits, 4.5 kHz LP cut-off) and stored in computer memory (DEC Micro PDP-11/23). Using a digital tape editing program a new utterance was created by concatenating parts of the digital record, as follows.

The utterance was truncated after the word say at a positive going zero crossing at the end of the glottal period whose intensity was no less than 90 percent of the peak value reached throughout the vowel. As it happened, this eliminated the final three glottal periods from the vowel. To this abrupt vowel termination were appended a 10 ms silent interval and a 60 ms stretch of steady state [sh]-noise, gated out from the centre portion of the original fricative. This in turn was followed by the final 20 ms of friction before the CV-boundary in shop and the remainder of the original utterance. As a result the friction noise had an abrupt onset, and lasted for 80 ms, including the transition into the following vowel. The particular temporal organisation was chosen so as to create a stimulus that could be interpreted as either shop or chop.

The resulting record was DA converted and passed through analog gates (Grason-Stadler 1284B) which were modified so as to allow continuous adjustment of rise/fall times (van den Broecke & van den Broek, 1978) with a precision of .1 ms. Twenty-five stimulus types were manufactured differing in the vowel offset in say and the friction onset in the ambiguous word shop/chop: five vowel decay times, ranging from smooth (80 ms linear decay) to abrupt (0 ms decay) in steps of 20 ms were generated, and orthogonally combined with 5 noise rise times (same steps). These were recorded onto audio tape in

8 different random orders, preceded by 10 practice items, and separated by 2 s interstimulus intervals (offset to onset).

The entire tape was played in a quiet room over headphones to 5 audiometrically normal adult native English listeners. They were instructed to decide for each stimulus whether they perceived shop or chop with binary forced choice.

3. RESULTS AND CONCLUSION

Percent fricative judgements was determined for each of the 25 stimulus types ($N = 40$ judgments per stimulus type). Figure 1 plots these percentages as a joint function of vowel offset (horizontal dimension) and noise onset (vertical dimension).

Let us concentrate, first of all, on the results obtained for stimuli with abruptly terminating vowels. Here we notice that abrupt noise onset is associated with fricative, and gradual onset with affricate. This effect runs counter to what is usually claimed in the literature (where gradual noise onset is a cue for fricative), but replicates our earlier results for this parameter with shop/chop tokens presented in a spoken context. Apparently, both the abrupt noise onset and the 10 ms of physical silence are completely masked by the preceding vowel. Both to ourselves and to our subjects the transition of vowel into consonant sounded perfectly smooth.

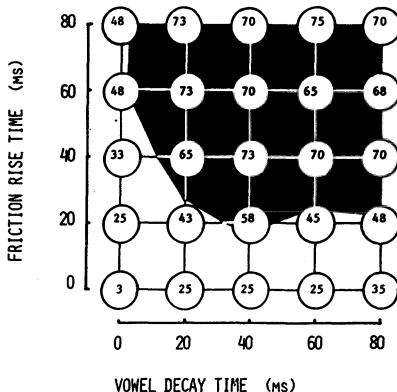


FIGURE 1. Percent chop responses as a function of the rise time of the friction noise and decay time of the preconsonantal vowel. The phoneme boundary separating affricate (shaded) from fricative (open) areas is drawn through the 50% cross-over points, which were determined by linear interpolation between stimuli straddling the boundary.

So far the results are in line with either explanation. However, the reversal of the rise time cue should disappear when the occurrence of forward masking is prevented. To this effect the preconsonantal vowel offset was systematically varied between abrupt (strong forward masking) and gradual (no or very weak masking). Clearly, the results falsify the prediction based on masking: the reversal of the noise onset cue is maintained regardless of the vowel offset characteristic, and more gradual vowel offsets are associated with affricate. The effects of vowel offset and noise onset are roughly additive, at least if we ignore the fact that the affricate scores plateau at 70-75%. These results unequivocally support the alternative hypothesis based on a central reinterpretation of low intensity sound as silence.

4. DISCUSSION

Though the masking hypothesis has been convincingly falsified, it should be noted that some measure of masking still persists. It was observed that the 10 ms silent gap separating vowel and friction sound was inaudible, which effect has to be ascribed to masking. Apart from this, temporal resolution of the intensity envelope was excellent throughout the experiment, since even the insertion of a mere 20 ms noise on-ramp or vowel off-ramp brought about a 22% increase of affricate judgments (cf. figure 1). Thus it would seem that the effects of forward masking in the present speech context are extremely limited, which finding is in line with e.g. Slis & van Nierop (1970) and Resnick et al. (1979) who showed that vowel onto consonant masking is quite small. Even at masker-probe intervals as short as 25 ms the amount of masking never exceeded -18 dB, while the intensity difference between adjacent vowels and consonants in normal speech is always less than this value.

Secondly, our results bear out that a silent interval is indeed a necessary condition for the perception of the affricate (or stop) manner. When such an interval is physically absent, low intensity portions of the stimulus flanking the VC-boundary are reinterpreted as silence. This behaviour seems to support the view that during speech perception the acoustic cues are evaluated in the light of what the listener knows about articulation. One wonders if reinterpretation of cues could be used in a more principled way to examine the relative importance of multiple cues in phonetic contrasts. Clearly, if one cue can be reinterpreted as an other, but not vice versa, the non-negotiable cue is the stronger of the two. In this light our results indicate once more that rise time is a manner cue of limited importance, especially for contrasts occurring in connected speech.

Thirdly, "sound" reinterpreted as "silence" provides a less powerful cue to affricate than physical silence does. We may observe that affricate judgments plateau at 70-75%, which means that the affricate end of the stimulus space was not highly convincing. Given the results of other studies much better exemplars of affricates can be generated if a proper period of silence is inserted between vowel and friction burst.

Finally, the reinterpretation hypothesis assumes that the perceived length of silence is much larger than the time interval that energy drops below threshold. If only the below-threshold portion of the energy dip contributed to the perceived silent interval, affricate judgments should not have plateaued at 75%; instead, each smoother noise onset or vowel offset should have boosted the percentage of affricates. Therefore it might be worthwhile exploring in rather more detail the perceptual equivalence of true silence and sound reinterpreted as silence. The adjustment paradigm seems particularly suited to this purpose. Subjects can be asked to adjust the duration of a true silent interval (with sharply defined boundaries) so as to be perceptually equal to a stimulus with smooth energy dips (symmetrically or asymmetrically) distributed on both sides of the VC-boundary. When such experiments are done with non-speech stimuli (e.g., sawtooth-white noise sequences), the adjusted silent interval should equal the time the energy dip remains below threshold. In more speech-like stimuli (e.g., vowel-affricate), subjects will tend to

exaggerate the perceived length of silence so as to make the stimulus fit a stop percept.

REFERENCES

1. Broecke, M.P.R. van den and Broek, D. van der (1978). Matching slopes with electronic switches. Progress Report of the Institute of Phonetics Utrecht 3.2, 24-26.
2. Broecke, M.P.R. van den and Heuven, V.J. van (1983). Effect and artifact in the auditory perception of rise and decay time: speech and non-speech. Perception and Psychophysics 33, 305-313.
3. Debrock, M. (1977). An acoustic correlate of the force of articulation. Journal of Phonetics 5, 61-80.
4. Dorman, M.F., Raphael and L.J., Liberman, A.M. (1979). Some experiments on the sound of silence in phonetic perception, Journal of the Acoustical Society of America 65, 1518-1532.
5. Dorman, M.F., Raphael, L.J. and Isenberg, D. (1980). Acoustic cues for a fricative-affricate contrast in word-final position, Journal of Phonetics 8, 397-405.
6. Gerstman, L.J. (1957). Perceptual dimensions for the friction portions of certain speech sounds. Unpublished Ph.D. dissertation, New York University.
7. Heinz, J.M. and Stevens, K.N. (1961). On the properties of voiceless fricative consonants. Journal of the Acoustical Society of America 33, 589-596.
8. Heuven, V.J. van (1979). The relative contribution of rise time, steady time, and overall duration of noise bursts to the affricate-fricative distinction in English. In J.J. Wolf and D.H. Klatt (eds.), ASA*50 Speech Communication Papers. The Acoustical Society of America, New York, 307-311.
9. Heuven, V.J. van (1983). Rise time and duration of friction noise as perceptual cues in the affricate-fricative contrast in English. In M.P.R. van den Broecke, V.J. van Heuven and W. Zonneveld (eds.), Sound structures, studies for Antonie Cohen. Foris, Dordrecht, 141-157.
10. Isenberg, D. (1978). Effect of speaking rate on the relative duration of stop closure and fricative noise. Haskins Laboratories Status Report on Speech Research SR-55/56, 63-79.
11. Kuipers, A.H. (1955). Affricates in intervocalic position, Haskins Laboratories Quarterly Progress Report 15, appendix 6.
12. Plomp, R. (1964). Rate of decay of auditory sensation. Journal of the Acoustical Society of America 36, 277-284.
13. Repp, B.H., Liberman, A.L., Eccardt, T., and Pesetsky, D. (1978). Perceptual integration of acoustic cues for stop, affricate, and fricative manner. Journal of Experimental Psychology: Human Perception and Performance 4, 621-637.
14. Resnick, S.B., Weiss, M.S. and Heinz, J.M. (1979). Masking of filtered noise bursts by synthetic vowels. Journal of the Acoustical Society of America 66, 674-677.
15. Slis, I.H. and Nierop, D.J.P.J. van (1970). On the forward masking threshold of vowels in VC-combinations. IPO Annual Progress Report 5, 68-72.
16. Truby, H. (1955). Affricates. Haskins Laboratories Quarterly Progress Report 11, 7-8.

POSSIBLE ACOUSTIC BASES FOR THE PERCEPTION OF VOICING CONTRASTS*

Richard E. Pastore
Department of Psychology, State University of New York at Binghamton,
New York 13901, United States of America.

Fifteen years ago the main questions for speech research, at least from a psychoacoustic perspective, concerned exactly how speech perception is special. At that time there was no strong motivation to study the perception of acoustic analogs to speech cues. Beginning about twelve years ago a few papers demonstrated for simple stimuli some of the phonemes which had been claimed to be unique to speech. In the late 1970s psychoacoustic research on speech perception tended primarily to challenge some of the strong claims about unique speech phenomena; this research also served to allow investigators to begin to learn about concepts, procedures, and potential acoustic cues. Recent psychoacoustic research has begun to map perceptual properties of simple possible analogs to components and cues in speech stimuli. This psychoacoustic research is guided by the extensive literature which has identified potential speech cues, and the perceptual interactions among such cues. The primary goal for this current psychoacoustic research is to gain an understanding of the nature of the perception of such stimuli. In achieving this goal, we will provide a better understanding of the front-end of the system which results in speech perception. This psychophysical mapping of acoustic analogs to speech cues might identify natural discontinuities which provide the basis for speech contrasts, but the research is not predicated on the assumed existence of such natural discontinuities.

In this paper I will attempt to summarize investigations of possible psychoacoustic bases for the voiced/voiceless distinction for initial position English stop consonants. This contrast has been the focus of much of my laboratory's research in recent years. In now comparing the findings for speech cues with the findings for various non-speech analogs to these cues, I am led to two conclusions. First, there are only few similarities in the perception of simple, relatively static acoustic stimuli. Second, there are a number of similarities in the perception of more complex, more dynamic acoustic analogs to speech stimuli.

*This research was supported in part by National Science Foundation grant 8302873 to the author and BRSG grant 07RR07149-12, Biomedical Research Support Grant Program, National Institutes of Health. The author acknowledges Jody Kaplan Layer, Robert J. Logan, Crystle Morris, Rosemary Szczesuil, and Virginia Wielgus for their significant contributions to the research described in this paper, and Charlotte MacLatchy for her critical reading of this paper.

Table I, Synthetic speech voicing (VOT) boundaries in initial position English stop consonants

GENERAL

VOT: 20 - 50 ms with voicing lag

PLACE of ARTICULATION: ±Cued by Burst and F2 transitionij

<u>Labial</u>	<u>Apical</u>	<u>Velar</u>	ms	Lisker & Abramson, 1970
25	35	42	ms	Miller, 1977
24.6	27.8	28.3	ms	Soli, 1983
24.2		35.8	ms	

VOWEL DURATION:

	<u>Short</u>	<u>Long</u>	<u>Diff.</u>	ms	
/bi-pi/	28.1	29.5	1.4	ms	Summerfield, 1981
/bis-pis/	23.5	26.4	2.9	ms	Summerfield, 1981
/biz-piz/	22.8	26.5	3.7	ms	Summerfield, 1981
/bi-pi/ (natural tokens)			1.9	ms	Jongman, 1986
/beace-peace/	210 20.	431 27.	ms	Diff. 7. ms	Miller, 1986
/beef-peef/	27.	32.	5.	ms	Miller, 1986

VOWEL F1 FREQUENCY:

F1:	200	300	400	500	Hz
Exp 1:	49.5	41.1	31.9	29.9	Summerfield, 1982
Exp 2:	40.0	36.4	29.5	28.1	Summerfield, 1982

F1 TRANSITION DURATION:

/da-ta/	25 21.	55 32.	70 41.	85	100 47. ms	Lisker, et al., 1977
/da-ta/	23.			39.	ms	Miller & Eimas, 1981

F1 ONSET FREQUENCY:

±Increased F1 Onset Frequencyij	Stevens & Klatt, 1974
results in	Lisker, 1975
±Shorter VOTij	Lisker, et al. 1977

Stevens & Klatt, 1974
Lisker, 1975
Lisker, et al. 1977
Summerfield & Haggard, 1977.

VOWEL F2 FREQUENCY:

No VOT dependency	Summerfield & Haggard, 1977.
-------------------	------------------------------

(Table I, continued)

F2 TRANSITION DURATION:

Very little effect on VOT

Lisker et al., 1977

RELATIVE ASPIRATION INTENSITY:

Asp. Ampl.:	- 6	0	+ 6	dB
VOT:	31.8	28.1	27.3	ms

Repp, 1979

Table I summarizes the relationship between various parameters of synthetic initial stop consonant continua and the locations of the voicing boundary defined in terms of the physical stimulus characteristics correlated with Voice Onset Time (VOT). This summary, while extensive, is not necessarily exhaustive either in terms of the stimulus properties which influence the voicing boundary or the published studies investigating the properties listed. Rather than simply asserting the dependency of VOT boundary on a specific parameter, I have summarized the actual VOT boundaries to provide a basis for evaluating the magnitude and reliability of the given dependency. When an author has not provided mean boundary locations, I have estimated the boundary locations from published figures.

We first will examine the findings for synthetic stop consonants summarized in Table 1. In English, the voicing boundary for a VOT continuum occurs only with a voicing lag of approximately 20 to 50 ms; we note that other languages exhibit an additional boundary at a voicing lead (Lisker and Abramson, 1970). Altering the cues for place of articulation from a front to a back consonant results in increased VOT boundaries, although the magnitude of this change in VOT ranges from a large 17 ms to less than 4 ms. Increasing vowel duration results in a systematic, but relatively small 1.4 to 7 ms increase in the voicing boundary. While these duration effects all are all statistically significant, most are probably considerably smaller than the JND. On this basis, Jongman (1986) questions whether duration effects are phonetically relevant.

Increasing the vowel F1 frequency results in a decrease in the voicing boundary. Increasing the duration of the F1 transition results in increased voicing boundaries, while increasing F1 Onset Frequency results in shorter voicing boundaries. There is little or no dependency of voicing boundary on F2 characteristics. Finally, increasing the relative amplitude of the aspiration noise, with voicing amplitude held constant, results in shorter voicing boundaries. These findings define two criteria which will be used to evaluate the possible contribution of acoustic cues to the perception of the voiced/voiceless distinction: the magnitude of the onset characteristics and the dependency of those onset characteristics on specific stimulus parameters. In the following Tables, the "+" and "-" symbols before the reference indicate that the findings are, or are not, consistent with each of these criteria.

Table II, Possible acoustic cues for VOT

LABELING TONE ANALOGS TO VOICED COMPONENTS

ISOLATED F1 WITH TRANSITIONSteady-State Frequency:

Frequency:	<u>400</u>	<u>600</u>	<u>800</u> Hz
Delay:	<u>26.5</u>	<u>24.5</u>	<u>20.5</u> ms

(+,+) Szczesuil et al., 1983

Transition Duration:

Duration:	<u>55</u>	<u>70</u>	<u>100</u> ms
Delay:	<u>20.5</u>	<u>31.9</u>	<u>52.0</u> ms

(+,+) Pastore et al., 1984a

Steady-State Duration:

Duration:	<u>150</u>	<u>300</u>	<u>450</u> ms
Delay:	<u>34.9</u>	<u>33.6</u>	<u>32.0</u> ms

(+-) Pastore et al., 1984a

F1 WITH F2 AND TRANSITIONSF1 Steady-State Frequency:

Frequency:	<u>400</u>	<u>600</u>	<u>800</u> Hz
Delay:	<u>47.9</u>	<u>37.4</u>	<u>29.7</u> ms

(+,+) Pastore et al., 1984b

Steady-State Duration:

Subjects	<u>F2 Type</u>	<u>120</u>	<u>300</u> ms
Practiced(4):	Velar	<u>47.7</u>	<u>42.8</u> ms

(+-) Pastore et al., 1985

Practiced(4):	Labial	<u>51.7</u>	<u>48.5</u> ms
---------------	--------	-------------	----------------

(-,-) Pastore et al., 1985

Naive (18):	Labial	<u>32.4</u>	<u>35.5</u> ms
-------------	--------	-------------	----------------

(+,+) Pastore et al., 1986a

NOISE ONSET TIMEGENERAL:

Categorical Perception: 15.1 ms (-,+) Miller et al., 1976

F1 Frequency:

Frequency:	<u>200</u>	<u>300</u>	<u>400</u>	<u>500</u> Hz
Exp. 1:	<u>27.2</u>	<u>20.0</u>	<u>16.9</u>	<u>19.2</u> ms

(-,-) Summerfield, 1982

Exp. 2:	<u>21.1</u>	<u>20.9</u>	<u>20.6</u>	<u>19.9</u> ms
---------	-------------	-------------	-------------	----------------

(-,-) Summerfield, 1982

Table II summarizes the findings for two possible acoustic cues. The upper portion of this table summarizes our recent research on the perception of continua representing the delayed onset of tone analogs to the center frequencies of the voiced portion of CV syllables. In all conditions subjects were asked to divide the ordered

continuum into two response categories, and then to use these categories to label random sequences of the stimuli. Unless otherwise noted, results are based upon 80-100 trials per data point for 3 or 4 psychophysically experienced subjects. All results are statistically significant ($p \leq 0.05$ or better), based upon within-subject comparisons.

As previously indicated, the first "+/-" symbol before the reference citation indicates whether or not the boundary locations are within the approximate 20 to 50 ms range normally reported for synthetic speech continua. The second "+/-" symbol indicates whether or not there is a significant parameter dependency, independent of the range of boundaries, which is in the same direction as that reported for voicing contrasts.

Nearly all labeling boundaries are within the approximate 20 to 50 ms VOT range, hence the initial "+" symbols. The labeling boundary for isolated F1 transition analogs is appropriately dependent upon the final steady-state frequency and the transition duration. The labeling boundary is also dependent upon steady-state duration, but in the direction opposite to that found for synthetic speech continua.

Adding an F2 transition and an F2 steady-state analog to the F1 stimulus maintains the findings for F1 steady-state frequency. However, the manipulation of steady-state duration provides some interesting results. With experienced subjects the boundary locations for the velar F2 analogs fall within the range on expected onset delays for velar stimuli, but decrease, rather than increase, with increased duration. The boundaries for the labial F2 analogs are at delays larger, rather than smaller, than for the velar stimuli and also exhibit an unexpected, inverse dependency on duration. Naive subjects with the same labial conditions yielded much shorter onset delay boundaries (though possibly more characteristic of apical than labial stimuli), and exhibited the expected direct dependency on duration. The naive and experienced subjects probably were attending to different perceptual characteristics of the stimulus continuum.

Tone analogs to the voiced component CV continua thus exhibit reasonable labeling boundaries with appropriate dependencies on steady-state frequency, on transition duration, and, at least under some conditions, on steady-state duration. Pols and Schouten (this volume) provide an additional mapping of the perception of such frequency transitions.

The bottom of this table summarizes research with the noise-buzz stimuli employed by Miller, Wier, Pastore, Kelly, and Dooling (1976) and Summerfield (1982). Miller et al. demonstrated categorical perception, but at a relatively short boundary, while Summerfield found relatively short boundaries, but failed to exhibit a consistent dependency on frequency.

Table III

Temporal order of onset				
<u>Simultaneous/Successive</u>				
TONE PAIRS				
Threshold:	2 ms			(-, -) Hirsh, 1959
Labeling (2 resp; lag):	11-23 ms			(-, -) Summerfield, 1982
Labeling (2 resp, lead/lag):	14 ms			(-, -) Pisoni, 1977
Labeling (3 resp, lead/lag):	-21 ms, +24 ms			(+, ?) Pisoni, 1977
Order Identification				
TONE PAIRS:				
<u>Relative Amplitude:</u>				
No effect: +/- 20 Phon (Rise-time \geq 7 ms)				(-, -) Hirsh, 1959
No effect: +/- 12 dB				(-, -) Pisoni, 1977
<u>Frequency:</u>				
$\pm 17\text{-}24$ ms (independent: freq. & lead/lag)				(-, -) Hirsh, 1959
Low Freq.	$\frac{200}{20.6}$	$\frac{300}{14.1}$	$\frac{400}{13.3}$	$\frac{500}{11.7}$
Exp. 1:				(-, +) Summerfield, 1982
Exp. 2:	21.4	24.4	18.8	22.4
				(-, -) Summerfield, 1982
<u>Stimulus Duration:</u>				
Duration:	$\frac{10}{4.7}$	$\frac{30}{9.4}$	$\frac{100}{10.0}$	$\frac{300}{12.2}$ ms
Threshold				(-, +) Pastore et al., 1982
Labeling	7.2	12.8	13.9	18.9
				(-, +) Pastore et al., 1982
<u>Rise Time:</u>				
Rise Time: $\frac{10}{100 \text{ ms Dur}}$	$\frac{30}{9.1}$	$\frac{50}{10.9}$	$\frac{100}{13.3}$	$\frac{18.2}{ms}$
				(-, +) Pastore et al., 1982
300 ms Dur: 12.1	13.9	17.9	23.6	ms
				(-, +) Pastore et al., 1982
<u>TONE/NOISE BURST:</u>				
17-24 ms independent of lead/lag				(?, -) Hirsh, 1959
TONES WITH INITIAL TRANSITIONS:				
<u>F1 Transition Duration:</u>				
Duration:	$\frac{25}{18}$	$\frac{50}{24}$	$\frac{75}{29}$ ms	
TOT:				(+, +) Hillenbrand, 1984
<u>F1 Frequency:</u>				
Frequency:	$\frac{250}{26}$	$\frac{450}{18}$	$\frac{750}{22}$ Hz	
Discrim.:				(+, -) Hillenbrand, 1984
Labeling:	26	21	25	(+, -) Hillenbrand, 1984

(Table III, continued)

TONE/NOISE WITH INITIAL TRANSITIONS:

Noise/Tone lead thresh.:	25 ms	Noise lead	(+,-) Pastore et al., 1986b
Noise lead threshold:	35 ms		(+,-) Pastore et al., 1986b
Tone lead threshold:	(greater than 56 ms)	(+,-)	Pastore et al., 1986b

Table III summarizes the findings for the judgment of temporal order of onset (TOT or Temporal Onset Time). Simultaneous/sequential threshold for the tone pairs is a very brief 2 ms, while labeling results in considerably longer boundaries which, with three categories, are approximately equal for high tone lead and lag. Most temporal order studies involve some form of identification of onset order. Many early TOT studies employed tone pairs, where the threshold or labeling boundary seems to be at relatively brief onset differences. These boundaries were independent of amplitude, yielded mixed results for the frequency of the lower component, but did exhibit the expected pattern of results for stimulus duration and initial rise time. Rosen and Howell (this volume) have re-analyzed these data and provide a compelling argument that TOT for such simple tonal stimuli cannot be an acoustic basis for the perception of voicing contrasts.

Hirsh (1959) investigated the order identification threshold for the pairing of a tone with a noise burst, again finding independence of stimulus characteristics and equal thresholds for tone and noise leading conditions. Hillenbrand (1984) investigated the order perception of tone pairs with onset transitions. These stimuli exhibited boundaries within our criterion range and an expected dependency on F1 transition duration, but not on F1 frequency.

A tangential observation is that most of the TOT labelling boundaries are at longer onset differences than the thresholds on which the labeling category boundaries theoretically are based. This difference in threshold and labeling boundary locations would seem to imply that labeling behavior may be based upon the use of conservative response criteria.

The final condition is one we have almost completed investigating. The two components of the temporal order task are: third octave noise following an F2 contour, and a mixture of two tones following F1 and F2 contours. The contours are defined in terms of both amplitude and frequency changes in time. The results of the first experiment are shown in Figure 1. As with Hirsh (1959), subjects were required to indicate whether the tone or noise had an earlier onset. The continuum of asynchronous onsets ranged from noise onset leading by 56 ms to tone onset leading by 56 ms. For stimuli with a velar F2 frequency contour, subjects exhibited only a single threshold located at a noise lead of from 10 to 30 ms, which is too brief to provide an acoustic basis for voicing contrast. The labial analog stimuli also exhibited only a single threshold, but this was at a significantly longer noise lead. This threshold asymmetry is not typical of TOT for stationary stimuli, but is equivalent to the VOT voicing asymmetry found for English stop consonants.

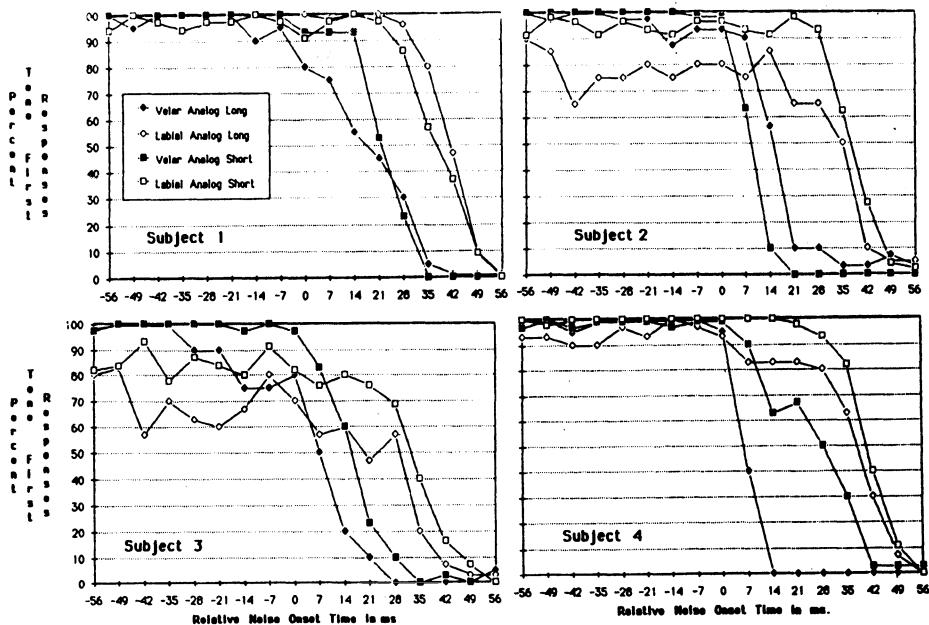
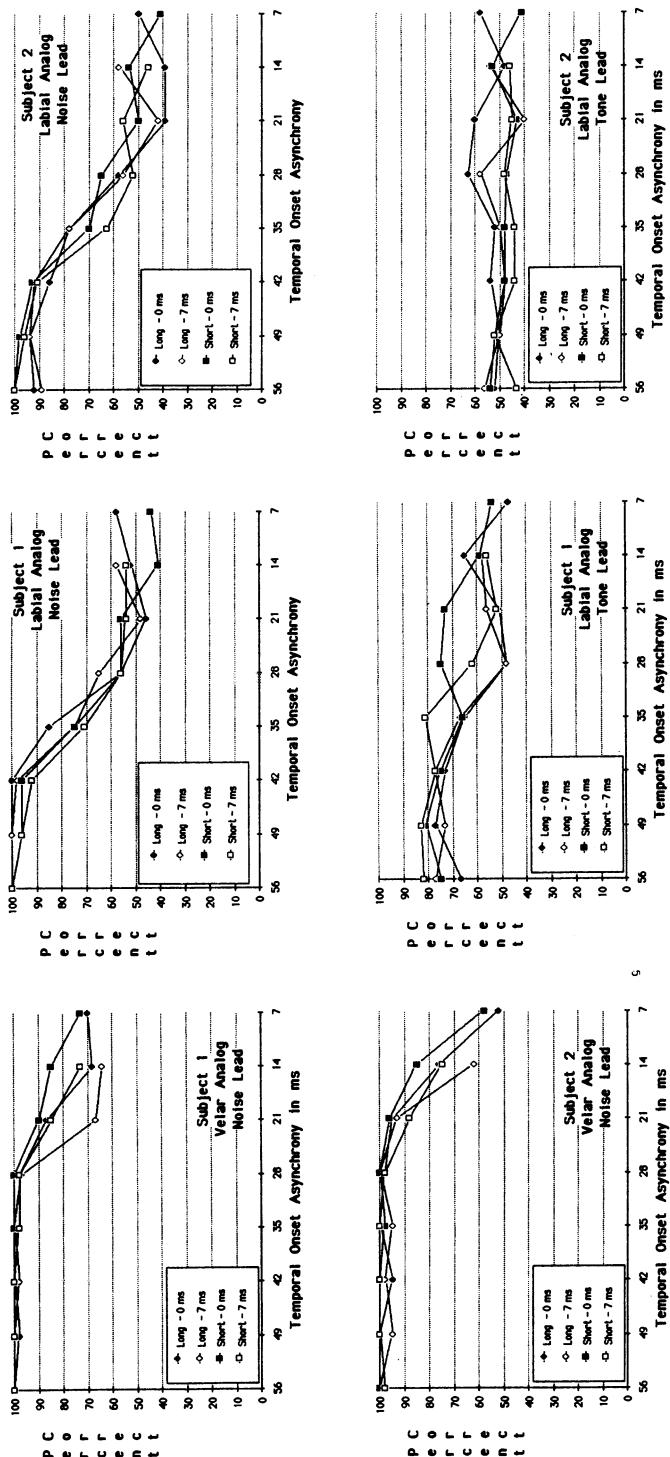


FIGURE 1.

We next used a 2-IFC discrimination task to measure the 75% thresholds for identifying longer onset differences. We ran separate conditions for noise and tone leading, each with 0 and 7 ms onset standards and with comparisons ranging up to 56 ms. The remaining figures are psychometric functions for Subjects 1 and 2, which are similar both to the results from Experiment 1, and to the results for the other two subjects. Figure 2 shows the functions for the Noise Lead condition with velar analog stimuli. We again find brief 10 to 25 ms thresholds. Across subjects, the 0 and 7 ms standards for these velar analog stimuli produced an average 8 ms difference in threshold, which was statistically significant [$F(1,3) = 11.7$, $p < .05$] and indicates relative discrimination performance. Figure 3 shows the functions for the labial analog stimuli for Subject 1. Here we find 30 to 45 ms thresholds for noise lead, and a difficult task for tone lead. Figure 4 shows the analogous results for Subject 2. Subjects 3 and 4 also produced similar thresholds which were equivalent the 0 and 7 ms standards for these Labial analog stimuli, indicating a common basis for discrimination performance.

The research I have summarized leads to the two conclusions cited at the beginning of this paper. First, the perception of synthetic speech stimuli and of simple, relatively static acoustic stimuli exhibit many differences and only few similarities. The second, and more important, conclusion is that there are many similarities in the perception of synthetic speech continua contrasted in voicing and the perception of moderately complex, dynamic acoustic analogs to such speech stimuli. These findings provide a reasonable basis for the hypothesis that some aspects of the perception of speech are the result of perceptual characteristics common to simpler acoustic stimuli, but certainly do not prove the validity of that hypothesis.



REFERENCES

1. Hillenbrand, J. (1984). Perception of sine-wave analogs of voice onset time stimuli. *Journal of the Acoustical Society of America*, *75*, 231-240.
2. Hirsh, I.J. (1959). Auditory perception of temporal order. *Journal of the Acoustical Society of America*, *31*, 759-767.
3. Jongman, A. (1986). Effects of speaking rate on the perception of syllable-initial stop consonants. *Journal of the Acoustical Society of America*, *75*, S9 (Abstract).
4. Lisker, L. (1975). Is it VOT or a first-formant transition detector? *Journal of the Acoustical Society of America*, *57*, 1547-1551.
5. Lisker, L. and Abramson, A.S. (1964). A cross-language study of voicing in initial stops: Acoustic measurements. *Word*, *20*, 384-427.
6. Lisker, L. and Abramson, A.S. (1970). The voicing dimension: Some experiments in comparative phonetics. In: *Proceedings of the 6th International Congress of Phonetic Sciences*, Prague, 1967, (Academia, Prague), 563-567.
7. Lisker, L., Liberman, A.M., Erickson, D.M., Dechovitz, D., and Mandler, (1977). On pushing the voice-onset-time (VOT) boundary about. *Lang. Speech*, *20*, 209-216.
8. Miller, J.D., Wier, C.C., Pastore, R.E., Kelly, W.J., and Dooling, R.J. (1976). Discrimination and labeling of noise-buzz sequences with varying noise-lead times: An example of categorical perception. *Journal of the Acoustical Society of America*, *60*, 410-417.
9. Miller, J.L. (1977). Nonindependence of feature processing in initial consonants. *Journal of Speech and Hearing Research*, *20*, 519-528.
10. Miller, J.L. (1986). Rate-dependent processing in speech perception. In: *Progress in the Psychology of Language*, vol.III, edited by A. Ellis (Erlbaum, Hillsdale, N.J.) (in press).
11. Miller, J.L. and Eimas, P.D. (1981). Contextual perception of voicing by infants. Paper presented to the biennial meeting of the Soc. Res. Child. Dev., 2-5 April 1981, Boston, MA [cited in Summerfield, 1981].
12. Pastore, R.E. (1981). Possible psychoacoustic factors in speech perception. In: *Perspectives in the Study of Speech*, edited by P.D. Eimas and J.L. Miller (Erlbaum, Hillsdale, N.J.), Chap. 5.
13. Pastore, R.E., Harris, L.B., and Kaplan, J.K. (1982). Temporal order identification: Some parameter dependencies. *Journal of the Acoustical Society of America*, *71*, 430-436.
14. Pastore, R.E., Morris, C., Logan, R., and Layer, J.K. (1986). Duration effects revisited: Labeling of tone analogs to voicing contrasts. *Journal of the Acoustical Society of America*, *79*, (Suppl. 1, S9 (Abstract).
15. Pastore, R.E., Logan, R., Morris, C. and Layer, J.K. (1986). Perceptual learning effects in the perception of speech-like stimuli. Meeting of the American Psychological Association, Washington D.C., August, 1986.
16. Pastore, R.E., Morris, C. and Layer, J.K. (1985). Duration effects on labeling of tone analogs to F1-cutback continua. *Journal of the Acoustical Society of America*, *78*, S69 (Abstract).

17. Pastore, R.E., Szczesuil, R., Nowikas, K. and Logan, R. (1984). The role of F1 cutback in the perception of voicing contrasts. Journal of the Acoustical Society of America, 75, S65 (Abstract).
18. Pastore, R.E., Wielgus, V.G. and Szczesuil, R. (1984). F1-cutback interactions in the perception of voicing contrasts. Journal of the Acoustical Society of America, 76, S28 (Abstract).
19. Pisoni, D.B. (1977). Identification and discrimination of the relative onset time of two-component tones: Implications for voicing perception in stops. Journal of the Acoustical Society of America, 61, 1352-1361.
20. Repp, B.H. (1970). Relative amplitude of aspiration noise as a voicing cue for syllable-initial stop consonants. Language and Speech, 22, 173-189.
21. Soli, S. (1983). The role of spectral cues in discrimination of voice onset time differences. Journal of the Acoustical Society of America, 73, 2150-2165.
22. Stevens, K.N. and Klatt, D.H. (1974). Role of formant transitions in the voiced-voiceless distinction for stops. Journal of the Acoustical Society of America, 55, 653-659.
23. Summerfield, Q. (1981). On articulatory rate and perceptual constancy in phonetic perception. Journal of Experimental Psychology: Human Perception & Performance, 7, 1074-1095.
24. Summerfield, Q. (1982). Differences between spectral dependencies in auditory and phonetic temporal processing: Relevance to the perception of voicing in initial stops. Journal of the Acoustical Society of America, 72, 51-61.
25. Summerfield, Q. and Haggard, M.P. (1977). On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants. Journal of the Acoustical Society of America, 62, 435- 448.
26. Szczesuil, R, Pastore, R.E., and Rosenblum, L. (1983). Perception of sine-wave analogs to individual formants of CV syllables. Journal of the Acoustical Society of America, 74, S66 (Abstract).

IS THERE A NATURAL SENSITIVITY AT 20 MS IN RELATIVE
TONE-ONSET-TIME CONTINUA? A REANALYSIS OF HIRSH'S (1959)
DATA*

Stuart Rosen
Department of Phonetics and
Linguistics
University College London
4 Stephenson Way,
London, NW1 2HE, U.K.

Peter Howell
Department of Psychology
University College London
Gower St.
London, WC1E 6BT, U.K.

Over the last dozen years, there has been much speculation concerning the possibility of auditory underpinnings to the structure of phonemic categories. Support for this hypothesis is usually drawn from studies of categorical perception, a mode of perception in which it was originally supposed that sounds can only be discriminated from one another to the extent to which they are labelled differently. This is meant to contrast with the more common situation, often called "continuous perception", in which the ability to discriminate far outstrips the ability to label differentially (Miller, 1956). Practically speaking, a continuum is said to be categorically perceived when there is: (1) a sharp categorization function, (2) a peak in the discrimination function at the category boundary and (3) near-chance discrimination performance within categories (Studdert-Kennedy, Liberman, Harris, and Cooper, 1970). Categorical perception was initially thought to be confined to speech sounds, and to arise from a reference to some aspect of the articulatory process (Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967). Since then, the phenomena of categorical perception have been obtained with a number of nonspeech auditory continua, and even with visual stimuli (e.g., Pastore, Ahroon, Baffuto, Friedman, Puleo, and Fink, 1977). These studies gave rise to the idea that "categorical perception" need not rely on the use of categories at all, but could arise simply from a non-uniform discriminability across the stimulus continuum. Few would argue against such a possibility. More controversial are the notions, put forward most concisely by Stevens (1981), that these auditory sensitivities are responsible for the categorical perception of speech sounds, and are the basis for all phonemic categories.

There is relatively little evidence supporting this point of view. Restricting ourselves only to studies with adult human listeners, all of these concern the voicing distinction in initial homorganic plosive stops¹. Both Miller, Wier, Pastore, Kelly, and Dooling (1976) and Pisoni (1977) have demonstrated categorical perception of nonspeech continua that have acoustic characteristics reminiscent of those which can signal the voicing distinction in plosives. Pisoni's work has attracted rather more scrutiny because he claimed that the acoustic contrast incorporated in his continuum is the primary acoustic contrast employed by listeners in perceiving the voice-onset-time (VOT)

*This work has been supported by the Medical Research Council of the United Kingdom.

contrast. We will not address that issue here. Our current concern will be Pisoni's finding of a natural auditory sensitivity at a, perhaps, not intuitively predictable point. Even more oddly, it is widely supposed that this result is consonant with what Hirsh (1959) found in a set of similar, although more wide-ranging experiments. In fact, Hirsh's results are quite different from Pisoni's in a number of ways. We shall see that Hirsh's study has been misinterpreted since its first appearance and, ironically, has been marshalled in support of both sides of a controversy concerning the adequacy of auditory processes as a basis for the categorical perception of VOT continua.

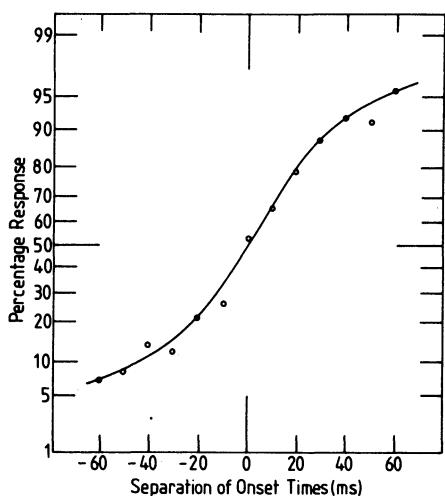


FIGURE 1. Mean results from a task requiring the judgement of the relative order of onset of two tones. Points to the right of 0 on the abscissa indicate that the higher tone led the lower. Points to the left of 0 indicate that the lower tone led the higher. The ordinate indicates the percentage of times the subject reported that the high tone led the low. The smooth curve is fit by eye to the original data points. Redrawn from Figure 4 of Hirsh (1959) including only those points that deal with two-tone stimulus complexes.

Hirsh's (1959) primary interest was in the ability of listeners to determine the order of occurrence of two sounds as a function of the temporal disparity between their onsets. A variety of different sounds were used, but we need only concern ourselves with the results with two pure tones, as these are most comparable to Pisoni's (1977) stimuli. The experimental paradigm was simple. A particular stimulus consisted of two sinusoids of different frequencies, with a range of relative onset times varying from -60 to +60 ms (i.e., the low frequency tone starting 60 ms before the high frequency tone to the high starting before the low by 60 ms). The two sinusoids terminated simultaneously and were about 500 ms long. Five subjects listened to the stimulus complex played repetitively until they could decide which of the two tones came first, the lower or the higher. Figure 1 shows the results, averaged over the five listeners and the five possible frequency pairs.

The accuracy of the subjects' responses improves smoothly with increasing separation of onset times. Hirsh, as is commonly done, summarized this result by choosing the point at which the subjects were performing at a level of 75% correct, concluding that a little less than 20 ms was needed to resolve the order of events. Note also that

there is no constant error: when the onsets of the two stimuli are simultaneous, subjects say the high tone leads the low as often as they say the low leads the high.

The first suggestion that the discriminability of Hirsh's stimuli might be non-uniform was advanced by Miller et al. (1976) who supposed that distinct percepts occur as one increases the separation of the two onsets: "Thus, as the amount by which the high tone precedes the low is increased, perceptual boundaries or thresholds are crossed corresponding to the perceptual effects of nonsimultaneity, Gestalt sequence with obvious ordering, and ordered onsets of two distinct percepts... one would expect to find perturbations in the Weber fraction at the perceptual boundaries." Hirsh did not perform a discrimination experiment but such perturbations should evidence themselves in the labelling function. As noted above, however, Hirsh's data shows a smooth increase in subject performance with increasing onset separation, and no evidence of discontinuities.

Pisoni (1977) took Miller et al.'s suggestion and performed both labelling and discrimination experiments with a continuum closely modeled on Hirsh's stimuli, with some slight differences. Pisoni used a single two-tone complex with components at 500 and 1500 Hz. Their relative onset times varied from -50 to +50 ms. For the labelling experiment, subjects were first trained with feedback to respond appropriately to the endpoint stimuli, -50 and +50 ms. After 320 trials in a random order, the -30 and +30 ms stimuli were introduced for a further 160 trials with feedback. During the identification experiment proper, all 11 stimuli were presented in a random order (15 presentations per stimulus) without feedback. Pisoni presented the results for each of the eight subjects individually. For six of the eight, the category boundary occurred not at 0 ms, as it did for Hirsh's subjects, but in the region where the high tone led the low by 15 to 25 ms. In order to compare the results from the two studies in more detail, we took the average results from both Hirsh (1959) and Pisoni (1977) and fitted cumulative normal curves to them using a maximum-likelihood technique (Bock and Jones, 1968, also known as probit analysis). The estimated category boundary was 0 ms for Hirsh's study and 13.6 for Pisoni's. Also, the slope of the curve fit to Hirsh's results was about 1.7 times shallower than the slope fit to Pisoni's indicating that Hirsh's subjects were rather less sensitive to changes in relative tone-onset-time than Pisoni's were. This is also reflected in the fact that Hirsh's subjects never did better than a 95% correct labelling of the order of the tones, even when there was 60 ms between the two onsets, while Pisoni's subjects averaged about 99% correct for onset asynchronies of +50 ms.

There are a number of differences between Hirsh's and Pisoni's studies which may be responsible for the discrepancy in the obtained category boundary. The three minor differences in the stimuli do not seem to account for the differences in results. Firstly, although Pisoni used other frequencies for his two tones than Hirsh did, Hirsh used a variety of frequency pairs (250-300 Hz, 250-1200 Hz, 250-4800 Hz, 1000-2000 Hz and 1000-4800 Hz) and obtained nearly the same results in all cases. Secondly, Hirsh's tones were typically at nearly the same intensity level², whereas Pisoni's 1500 Hz tone was 12 dB lower in intensity than the 500 Hz tone. Pisoni, however, varied the upper frequency tone over a 24 dB range (-12 dB to +12 dB re the level of

the lower tone) and found no differences in the category boundaries obtained in labelling. Finally, Hirsh's stimuli had a base duration of 500 ms, whereas Pisoni's were 230 ms. Perhaps category boundaries on tone-onset-time continua shorten with increasing base duration. Pastore, Harris, and Kaplan (1982), however, found category boundaries to increase with increasing stimulus duration.

Having ruled out stimulus differences, then, the most likely cause of the discrepancy is the task set the subjects. Differences in instructions to the subjects may have caused them to attend to different aspects of the stimulus complex. Hirsh asked his to identify which of the two tones came first - instructions that favour a category boundary at simultaneity. Pisoni trained his subjects to respond differentially to exemplars of each category on the continuum and gave no verbal labels to the important stimulus characteristics. One other factor may be important. Pisoni gave feedback in initial training with stimuli at relatively extreme positions on the continuum. Hirsh (1959) makes no explicit statement about whether feedback was used or not, but informs us (Hirsh, personal communication) that feedback was never given. Had feedback been given on each trial, this might have encouraged the subjects to place their boundary at a value near 0 ms in order to maximize their performance.

A few other studies have investigated the labelling of tone-onset-time continua. Pisoni (1980) used the same continuum as previously in an adaptation study. Only the baseline results (i.e., an ordinary labelling experiment) need concern us here. Four such experiments were run. The obtained boundary (averaged over subjects) varied from 8.3 to 17.9 ms, for an overall mean of 13 ms, nearly identical to the value we calculated previously for the 1977 study.

Summerfield (1982) also explored the perception of tone-onset-time continua. Four frequency pairs were possible with 2500 Hz always as the upper frequency tone. The lower frequency tone was at 200, 300, 400, or 500 Hz. The relative levels of the tone were set to match the relative levels of the formants in a set of analogous voice-onset-time stimuli, and so varied with the frequency of the lower tone in the pair. The level of the upper tone was therefore -8, -5, 0 and, +4 dB relative to the level of the lower tone for the frequencies of 200 to 500 Hz, respectively. In Experiment 1, Summerfield used an adaptive technique to estimate the category boundary and found the four-subject mean to decrease monotonically from 20.6 ms to 11.7 ms with increases in the frequency of the lower tone. In a more typical labelling experiment, in which all the stimuli occurred equally often, there was no significant change of the category boundary with frequency of the lower tone. The mean boundary (over subjects and conditions) was 21.5 ms. This is quite different from the 13 or so ms Pisoni has consistently found, and very different from Hirsh's boundary of 0 ms. In this latter experiment however, no negative tone-onset-times occurred, only values between 0 and +60 ms. As is well known (Parducci 1965), the range of values used in labelling experiments is an important determinant of the category boundary obtained for both nonspeech and speech (see Howell and Rosen, 1984, for a review) so it is perhaps not surprising that Summerfield found a longer category boundary in this case. If range effects were important, his own adaptive studies, which used a continuum extending down to values of -20 ms, would be expected to

show shorter category boundaries. Consistent with this prediction, the mean category boundary was 14.9 ms (ignoring the influence of the frequency of the lower tone). It also seems likely that Summerfield's instructions to label the stimuli as to whether the onsets of the two component tones were "simultaneous" or "successive" would encourage subjects to place their category boundary at a positive relative onset-time.

It is important to resolve these serious differences, especially in the light of common misinterpretations of what Hirsh's data actually show. There are many comments in the literature (one of them ours) that imply or state that Hirsh's data show some evidence for a discontinuity in discriminability near 20 ms. This basic misinterpretation has to do with the nature of a psychometric function. Hirsh found performance in his experiments to vary smoothly with increasing onset separations, and, as we pointed out above, followed tradition in choosing the 75% correct point (17 ms) to summarize his data. He could just as well have picked 60%, giving about 5 ms as the crucial time, or, perhaps more appropriately for linguistic use (to ensure more reliable reception), 90% giving about 30 ms. None of these choices is preordained.

The fallacy in reasoning may be more easily appreciated when applied to a simple psychophysical continuum. Suppose subjects were presented with a continuum of sounds varying from 60 to 70 dB SPL in 1 dB steps and asked to label them as "loud" or "soft". We might find they had a category boundary at 65 dB and, by interpolation, that a stimulus of 66.5 dB was necessary for the label "loud" to occur 75% of the time. If we consider all stimuli above 65 dB as "loud", we might also say that presenting a sound at 66.5 dB led to 75% correct performance. So far so good. But would we then want to argue that the pair of stimuli 65 and 67 dB were inherently more discriminable than the pair 68 and 70 dB because they straddled this (arbitrarily defined) 75% point?

By the same token, there is no reason to suppose from Hirsh's labelling function alone that discriminability between tone-onset-time stimuli will be better around 17 ms. On the other hand, we do not preclude such a possibility. This requires a somewhat more complicated analysis.

We can predict a discrimination function from Hirsh's labelling data under some simple and fairly reasonable assumptions. Using Thurstonian Case V analysis (assuming a uni-dimensional psychological continuum where stimulus densities are Gaussian distributed with equal variance; see Torgerson, 1958, for details) we take the normal deviates of the proportion of "high precedes low" judgements (Figure 1) as the scale values for each of the stimuli. In order to predict performance in a discrimination task for, say, two-step comparisons (usually the most informative with stimulus spacings commonly used, as three-step comparisons often show ceiling effects near perfect performance and one-step comparisons often show floor effects at chance), the d' values between the appropriate stimuli are computed by taking the difference between the scale values³. We can then convert these values into the proportion correct that might be expected in an ABX task using a method based on signal detection theory developed by Macmillan, Kaplan, and Creelman (1977) and the tables of Kaplan,

Macmillan and Creelman (1978). Since this transformation increases monotonically, it will preserve any peaks in the d' function, so for that purpose it is not crucial whether we examine d' or proportion correct: we use the latter for a later comparison to Pisoni's (1977) results. Figure 2 shows the final outcome.

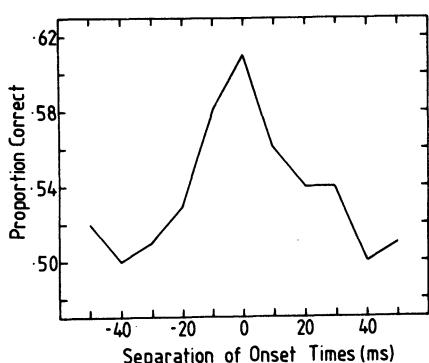


FIGURE 2, ABX discrimination results for stimuli varying in relative tone-onset-time predicted from Hirsh's (1959) data of Figure 1.

As we argued from the labelling function, there is no increased sensitivity around tone-onset-times of 20 ms. On the other hand, discriminability is not uniform across the continuum. The discrimination function is significantly peaked, with best discrimination near simultaneity, 0 ms. Under the assumptions we have made, performance is simply a monotonic transformation of the derivative of the labelling curve. Therefore, the peak is a reflection of the fact that the labelling function is steepest at its centre and flattens towards its edges, as is clearly seen in Figure 1.

This result might be expected if Weber's Law described the discrimination of relative onset time. Remember that we have made predictions for discrimination based on a fixed 20 ms difference between stimuli; if the difference limen for different values of relative onset time was proportional to the magnitude of the relative onset time (as would happen if Weber's Law applied), then the constant 20 ms change should lead to best performance at the shortest relative onset times (i.e. simultaneity), which decreased monotonically with increasing relative onset time, just as we have found. This also implies that the original labelling function of Figure 1 should be more linear under a logarithmic transformation of the relative-onset-time axis. We leave out of consideration 0 ms relative onset time as this value cannot be log-transformed. Taking advantage of the symmetry of Hirsh's results around 0 ms, we average the absolute values of the normal deviates corresponding to performance at each of the six onset times from 10 to 60 ms. (In other words, we average the performance at -10 ms with that of 10 ms, -20 ms with that of 20 ms, and so on.) These values are plotted as a function of linear and log relative onset time in Figure 3.

With the linear scaling on the left of Figure 3, we see the curvature of the function already displayed in Figure 1: a decreasing slope with increasing onset time. When the stimuli are plotted on a logarithmic scale, however, the slope of the curve seems to remain constant out to the longest relative onset times measured. This is further support for the notion that Weber's Law holds for these stimuli, and that no special acuity exists for relative onset times near 20 ms.

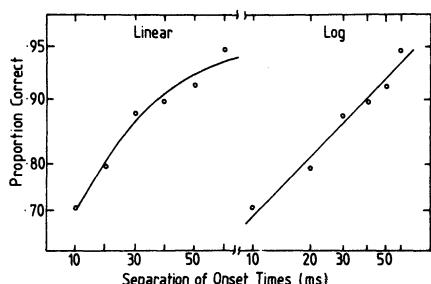


FIGURE 3. Linear and logarithmic scaling of Hirsh's (1959) data (seen in Figure 1), averaged over positive and negative relative onset times, as a function of the absolute value of the relative onset times. The value for a separation of 0 ms is not used because it cannot be logarithmically scaled. The smooth curve on the left is fit by eye to the data points; that on the right is a least-squares straight-line fit.

Pisoni (1977) found quite a different result when he presented his stimuli in a standard ABX two-step (20 ms) paradigm. Although he also found a primary peak in the discrimination function, it occurred at a value of 10 to 20 ms. This correspondence between the peak of the discrimination function and the category boundary obtained in a labelling experiment was taken as evidence that the tone-onset-time continuum was categorically perceived⁴. Pisoni's data is quite convincing on this point, and it was therefore argued that both the labelling boundary and the discrimination peak were due to auditory processes. This claim was strengthened by the demonstration that the subjects' discrimination performance was not a result of their using the previously-learned labels;

subjects with no previous experience of identifying the stimuli showed the same pattern of discrimination. Contrary to many claims, however, this result is not compatible with Hirsh's (1959) findings.

Hirsh's results have, however, a long history of misinterpretation by workers in speech perception. They were first cited, soon after they appeared, in a study by Liberman, Harris, Kinney and Lane (1961) who considered Hirsh's stimuli to incorporate a similar acoustic contrast to their own VOT stimuli. Before we can address that issue, however, we will need to describe that study in detail.

Liberman et al. (1961) demonstrated categorical perception for a /də/ to /tə/ continuum in which the onset of the first formant (F1) relative to the upper formants was progressively delayed from 0 to 60 ms in 10 ms steps. A nonspeech control condition was also included in which the Pattern Playback schematic spectrograms were turned upside down and further modified before use. These were not heard as speech and resulted in a continuum in which the third and highest formant had its onset delayed relative to the two lower formants. Strictly speaking, it was not possible to say anything about the categoricity of the perception of the nonspeech stimuli as they were not presented for labelling, only in the ABX discrimination paradigm. Discrimination performance for these stimuli was much inferior to that obtained for the speech stimuli, however, and it was concluded that (assuming the control to be fair) the performance obtained with speech could be interpreted as an example of distinctiveness acquired through learning.

Liberman et al. also compared their results to those obtained by Hirsh, whose stimuli they considered another nonspeech control for the F1-cutback stimuli. Two claims were made: first, that the overall performance exhibited by Hirsh's subjects judging the nonspeech continuum was inferior to Liberman et al.'s subjects judging the speech continuum, and second, that the speech discrimination functions were peaked, unlike the discrimination functions from nonspeech.

Liberman et al.'s primary error is a misinterpretation of the nature of Hirsh's experiment. They seem to consider Hirsh's results as a discrimination function, instead of the labelling function it is. Thus, Hirsh's finding of 75% correct performance with relative onset times of about 17 ms is compared to their own finding of 75% correct performance in an ABX task with less than 12 ms difference in time of onset on the F1-cutback continuum. Only the speech labelling functions of Liberman et al. can be directly compared to Hirsh's results. This comparison, when made, leads to much stronger support, in fact, for Liberman et al.'s assertion of acquired distinctiveness for the speech sounds. They state ".... in every case a change of 10 msec in the first-formant cutback is sufficient to shift the responses from 75% /d/ to 75% /t/..." while Hirsh's data show that a change of about 35 ms in relative onset time is necessary to shift the responses from 75% "low leads high" to 75% "high leads low".

Our analysis summarized in Figure 2 is needed to address Liberman et al.'s second claim: "One finds in Hirsh's results no indication of the sharp peaks so clearly evident in the discrimination functions of the present experiment". This is equivalent to looking for sharp peaks in the speech labelling functions! The discrimination function we have predicted from Hirsh's data does, in fact, show a clear peak.

On the other hand, here we can confirm again Liberman et al.'s assertion that performance with the speech continuum is far superior to that obtained with nonspeech. For a 20 ms difference in relative onset time, we predict that Hirsh's subjects would obtain, at best, 61% correct in an ABX discrimination task. Liberman et al.'s subjects, as noted above, did about 75% correct with differences of slightly less than 12 ms.

This reanalysis of Hirsh's data in no way impugns the substance of what Liberman et al. (1961) were saying. As we have noted, performance with the speech stimuli is much better than that obtained for the nonspeech stimuli. Furthermore, even though Hirsh's data may, contrary to Liberman et al.'s assertions, contain discrimination peaks, they are not in the same place as the peaks for the speech continuum (about 20 to 30 ms).

If however, we take Pisoni's (1977) results as a nonspeech control condition, few of these points hold. Taking only the data from the 5 best subjects of 8 (Liberman et al. used the best 11 of 20), the labelling functions seem to be as sharp as, and their discrimination abilities at least roughly equivalent to Liberman et al.'s subjects. Also, there is a significant peak in the discrimination function at a value not too far from the one obtained with speech.

Many questions remain. First, there is the problem of the discrepancy between Hirsh's (1959) findings and the more recent ones of Pisoni (1977, 1980) and Summerfield (1982). Second, even if these later results are upheld, their theoretical underpinnings are shaky. Hirsh's data seemed explicable on the basis of a limitation in the ability to identify the order of events since the labelling function he obtained was symmetric. Pisoni and Summerfield's asymmetric labelling functions imply there is more to it than that, although Pisoni argues explicitly that it is such a limitation that underlies both the position of the category boundary and the peak in the discrimination function. If this were the case, one would expect similar discrimination peaks in the continuum when the low tone led the high. Pisoni's (1977) discrimination data from untrained subjects (Experiment II), when averaged, do show evidence of a secondary discrimination peak in this region, although it is considerably smaller than the peak in the region where the high tone leads the low. Furthermore, of the 11 subjects who performed at levels above chance (one subject did not), only 6 show convincing discrimination peaks in the "low leads high" region, while all show peaks in the "high leads low" region.

Third, there is the problem of the variability of the obtained category boundaries even in the modern studies. If the labelling function is determined by a natural auditory property, the obtained boundaries should be invariant over manipulations in the range of stimuli presented. Our interpretation of Summerfield's (1982) data suggests this is not the case.

Only empirical studies can resolve these issues. Primary among these would be a replication of Hirsh's (1959) study with a particular emphasis on the effect of varying instructions to the subject. Secondly, more standard psychophysical investigations are desirable in order to assess the discriminability of changes in tone-onset-time across the continuum, and how these changes are affected by the frequency and amplitude relationships between the two component tones of the complex. Initial forays in this direction have been made by Pastore et al. (1982).

These studies would be well worthwhile even if temporal order identification does not form the basis for the perception of voice onset time. As Hirsh originally pointed out, determining the order of auditory events is still an important aspect of auditory perception in general, with wide-ranging implications for understanding the perception of music and speech.

FOOTNOTES

1. Cutting and Rosner's (1974) study of the categorical perception of nonspeech "pluck"/"bow" and voiceless affricate/fricative continua, all of which were based on variations in rise time, was once considered influential in this regard. Unfortunately, their results have not withstood replication. See Rosen and Howell (in press) for a review.

2. Hirsh (1959), at least for the condition where the stimulus complex contained tones at 250 and 1200 Hz, used the phon scale to set the tones to be equal in loudness at a level of 80 phons. Assuming the tones in the other four two-tone conditions to be set at the same level, the intensity difference between the two tones would not have

been greater (and often smaller) than 2 dB (Fletcher and Munson 1933). In addition Hirsh varied the level of the 250 Hz tone, setting it to 60 and 70 phons, and found results "... not very different from the equal-loudness case." Note though, that in Pisoni's main experiment, the lower frequency tone was the more intense.

3. This may seem a little odd in that we are making the assumption that discrimination is predictable from identification or, in Macmillan, Kaplan, and Creelman's (1977) definition, that the continuum is categorically perceived! It is legitimate here because Pynn, Braida, and Durlach (1972) show that in experiments where the signals span a small range, discrimination distances inferred from identification experiments are close to those estimated directly. That Hirsh's (1959) stimuli do indeed only span a small range is attested to by the fact that the subjects never do better than about 95% correct at the extremes of the range, even though only two categories are involved.

4. Pisoni shows that the peak in the obtained discrimination function is well predicted from the labelling function using the so-called "Haskins formula". This formula can also be applied to Hirsh's labelling data, which then predicts a discrimination function peaked at 0 ms, just as our previous analysis did.

REFERENCES

1. Bock, R.D. and Jones, L.V. (1968). The Measurement and Prediction of Judgment and Choice. San Francisco: Holden-Day.
2. Cutting, J.E. and Rosner, B.S. (1974). Categories and boundaries in speech and music. Perception & Psychophysics 16, 564-570.
3. Fletcher, H. and Munson, W.A. (1933). Loudness, its definition, measurement, and calculation. Journal of the Acoustical Society of America, 5, 82-108. Also in: (1969) Forty Germinal Papers in Human Hearing, ed. J.D. Harris, pp. 175-184. Groton, Connecticut: The Journal of Auditory Research.
4. Hirsh, I. (1959). Auditory perception of temporal order. Journal of the Acoustical Society of America, 31, 759-767.
5. Howell, P. and Rosen, S. (1984) Natural auditory sensitivities as universal determiners of phonemic contrasts. In: Explanations for Language Universals, ed. B. Butterworth, B. Comrie, and O. Dahl. Berlin: Mouton. Also: Linguistics 21, 205-235.
6. Kaplan, H.L., Macmillan, N.A., and Creelman, C.D. (1978). Tables of d' for variable-standard discrimination paradigms. Behaviour Research Methods & Instrumentation, 10, 796-813.
7. Liberman, A.M., Cooper, F.S., Shankweiler, D.P., and Studdert-Kennedy, M. (1967). Perception of the speech code. Psychological Review, 74, 431-461.
8. Liberman, A.M., Harris, K.S., Kinney, J.A. and Lane, H. (1961). The discrimination of relative onset time of the components of certain speech and nonspeech patterns. Journal of Experimental Psychology, 61, 379-388.
9. Macmillan, N.A., Kaplan, H.L., and Creelman, C.D. (1977). The psychophysics of categorical perception. Psychological Review, 84, 452-471.
10. Miller, G.A. (1956). The magical number seven plus or minus two, or, some limits on our capacity for processing information. Psychological Review, 63, 81-96.

11. Miller, J.D., Wier, C.C., Pastore, R.E., Kelly, W.J., and Dooling, R.J. (1976). Discrimination and labeling of noise-buzz sequences with varying noise-lead times: An example of categorical perception. *Journal of the Acoustical Society of America*, 60, 410-417.
12. Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review*, 72, 407-418.
13. Pastore, R.E., Ahroon, W.A., Baffuto, K.J., Friedman, C., Puleo, J.S., and Fink, E.A. (1977). Common-factor model of categorical perception. *Journal of Experimental Psychology: Human Perception and Performance*, 3, 686-696.
14. Pastore, R.E., Harris, L.B., and Kaplan, J.K. (1982). Temporal order identification: Some parameter dependencies. *Journal of the Acoustical Society of America*, 71, 430-436.
15. Pisoni, D.B. (1977). Identification and discrimination of the relative onset time of two-component tones: Implications for voicing perception in stops. *Journal of the Acoustical Society of America*, 61, 1352-1361.
16. Pisoni, D.B. (1980). Adaptation of the relative onset time of two-component tones. *Perception & Psychophysics*, 28, 337-346.
17. Pynn, C.T., Braida, L.D. and Durlach, N.I. (1972). Intensity perception: III. Resolution in small-range identification. *Journal of the Acoustical Society of America*, 51, 559-566.
18. Rosen, S. and Howell, P. (in press). Auditory, articulatory and learning factors in categorical perception. In: Categorical Perception, ed. S. Harnad. Cambridge University Press.
19. Stevens, K.N. (1981). Constraints imposed by the auditory system on the properties used to classify speech sounds: Data from phonology, acoustics and psycho-acoustics. In: The Cognitive Representation of Speech, ed. T.F. Myers, J. Laver and J. Anderson. Amsterdam: North Holland.
20. Studdert-Kennedy, M., Liberman, A.M., Harris, K.S., and Cooper, F.S. (1970). The motor theory of speech perception: A reply to Lane's critical review. *Psychological Review*, 77, 234-249.
21. Summerfield, Q. (1982). Differences between spectral dependencies in auditory and phonetic temporal processing: Relevance to the perception of voicing in initial stops. *Journal of the Acoustical Society of America*, 72, 51-61.
22. Torgerson, W.S. (1985). Theory and methods of scaling. New York: John Wiley & Sons, Inc.

AUDITORY CONSTRAINTS ON SPEECH PERCEPTION*

Randy L. Diehl
University of Texas at Austin

A dominant issue dividing theorists in virtually every branch of cognitive science these days concerns the topic of explanatory generality. One school of thought, exemplified by Allen Newell and Herbert Simon (1972) in artificial intelligence and by John Anderson (1983) in psychology, likes to view the mind as a single general-purpose device that can be applied indifferently across a wide range of stimulus and problem-solving domains. A quite different viewpoint, recently articulated by philosopher Jerry Fodor (1983) and by the late David Marr (1982), holds that the only psychological mechanisms that are scientifically tractable and hence worth investigating are those that are modular. Modules, in Fodor's sense, are information-processing devices that are "domain specific" i.e., they apply only to a restricted range of stimuli. In addition, modules are taken to be more or less automatic in their operation and to have a fixed neural architecture that is presumably innately organized.

As everyone in this audience will recognize, the debate between generalists and modularists has a fairly long history within the field of speech perception. The modularity perspective has been best represented by Alvin Liberman and his colleagues, who have claimed that the perception of speech sounds requires a specialized mode of phonetic processing above and beyond the general auditory and cognitive capabilities used in detecting, discriminating, and recognizing nonspeech acoustic patterns (Liberman, 1982; Liberman, Cooper, Shankweiler and Studdert-Kennedy, 1967; Liberman & Mattingly, 1985; Repp, 1982). On the other side of the issue have been investigators such as James Miller and Patricia Kuhl (Miller et al., 1976; Kuhl & Miller, 1975; 1978), Richard Pastore (1981), David Pisoni (1977, Pisoni, Carrel & Gans, 1983) and Bert Schouten (1980), among others. Despite their varying theoretical positions, the critics of the speech-mode hypothesis typically share the view that many interesting phenomena of speech perception result from general auditory processes and constraints and need not be viewed as speech specific.

Over the years, the debate over the speech-mode hypothesis has focussed on a variety of empirical phenomena, including categorical perception, laterality effects, and selective adaptation effects. At the moment, much of the debate centers on the interpretation of a set of phenomena known as phonetic trading relations (Repp, 1982). For any given phonetic segment there are, in general, many relevant acoustic cues. A phonetic trading relation is

*Work supported by N.I.H. Grant HD/8060

said to exist when a change in the value of one cue can be offset by an opposing change in another cue so that phonetic quality is preserved.

Bruno Repp (1982) has argued that most phonetic trading relations are not readily explained in terms of general auditory psychophysics, but instead seem to require the notion of a speech mode of perception. For example, Mann and Repp (1980) found an effect of the following vowel on the identification of a noise segment that was intermediate in frequency characteristics between [s] and [ʃ]. This segment was more likely to be labeled "s" before [u] than before [a]. Such an effect can presumably be explained in terms of the listener's tacit knowledge of fricative-vowel coarticulation and its acoustic effects. Lip rounding appropriate for [u] occurs during the preceding fricative segment, producing a lower-frequency fricative noise. To compensate perceptually for this, listeners may more readily accept a lower-frequency noise as [s] (rather than [ʃ]) in front of a rounded vowel.

Over the past two years, my co-workers Ellen Parker, Keith Kluender, Beverly Wright, and Margaret Walsh and I have undertaken a series of studies to determine the extent to which certain phonetic trading relations have a basis in general audition. Our strategy has been first to demonstrate a trading relation between acoustic dimensions that signal a phonetic contrast and then to determine whether the same type of trading relation exists for analogous acoustic dimensions when they signal perceptual distinctions among nonspeech stimuli. (For a similar approach, see Hillenbrand, 1984; Pisoni, 1977; Pisoni, Carrell, and Gans, 1983; Summerfield, 1982). To date, all of our experiments have yielded a clear parallel between speech and nonspeech results. Accordingly, we propose that many of the important trading relations in speech do, in fact, derive from general auditory principles and constraints rather than from a speech mode of perception.

I will now review some of our main findings, starting with Ellen Parker's (1985) recent dissertation work.

Parker chose to investigate a well-known trading relation in the perception of the voiced-voiceless distinction in initial stop consonants. Both Lisker (1975) and Summerfield and Haggard (1977) have shown that the location of the voiced-voiceless perceptual boundary along the voice-onset-time (VOT) dimension depends on the onset frequency of the first formant. For lower F1 onset frequencies, the voiced-voiceless boundary occurs at greater values of VOT.

Following Pisoni (1977) and Summerfield (1982), Parker created nonspeech analogues of VOT stimuli. Each consisted of two co-terminous sinusoids with the onset of the lower-frequency tone lagging that of the higher-frequency tone by durations from 0 ms to 120 ms. Two stimulus series were generated, one having a lower-tone frequency of 250 Hz, the other with a lower-tone frequency of 750 Hz. For both series, the frequency of the higher component was 1250 Hz.

In these stimuli, the two tones are taken to be abstractly analogous to the first and second formants of speech stimuli, and the

tone-onset-time (or TOT) variation is supposed to correspond to voicing delays in VOT stimuli. Pisoni (1977) has shown that listeners discriminate and label TOT stimuli in a manner roughly comparable to VOT stimuli and has argued that performance in both cases is determined by a general auditory limit on the temporal resolution of onset differences. (See also Miller et al., 1976.)

In Parker's study, subjects were asked to judge whether the onsets of two tones were simultaneous or not and to indicate their answers by pressing one of two buttons. Before identifying the test items, subjects received training trials with feedback on the endpoint stimuli of both series (40 trials per stimulus). Following training, the identification session consisted of 10 presentations of each stimulus from both series.

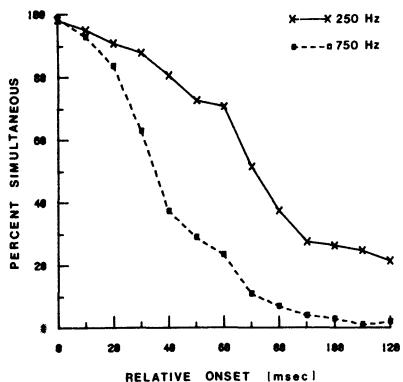


FIGURE 1, Percentage of stimuli whose two tonal components were judged as having simultaneous onsets. The Abscissa represents the actual delay of the lower-frequency component (either 250 Hz or 750 Hz) relative to the higher-frequency component (1250Hz). From Parker (1985).

Parker repeated the experiment, using a variety of stimulus and procedural modifications. She tested a broader range of lower- and higher-tone frequencies; she presented the stimulus conditions in a blocked rather than a mixed design; and she varied the level of subject training. In each of six additional experiments, her original results were replicated. Parker concluded that the nonspeech trading relation was indeed a reliable effect.

In our view, Parker's work provides strong support for a general auditory account of the speech trading relation between VOT

The results of this experiment are shown in Fig. 1, which plots the percentage of judgments of simultaneous tone onset as a function of the actual delay of the lower tone. It can be seen that when the lower-frequency tone is 250 Hz (rather than 750 Hz), considerably longer onset delays are required for subjects to consistently label the onsets as nonsimultaneous. This pattern of results essentially duplicates the trading relation discussed earlier between VOT and first-formant onset frequency. Recall that as the onset frequency of F1 is lowered, the location of the perceived voiced-voiceless boundary is shifted toward greater VOT values.

Parker was frankly surprised by the magnitude of this nonspeech trading relation because in an earlier experiment along similar lines Summerfield (1982) had found only a small and unreliable effect of lower-tone frequency on judgments of onset simultaneity. Accordingly,

and F1 onset frequency. Such an account would explain why listeners require longer VOT values to perceive a stop as voiceless in front of high vowels (with their low F1 frequencies) than in front of low vowels (Diehl, Lang & Parker, 1980). It would also explain why both humans and chinchillas show the same systematic variation in the VOT labeling boundary as a function of place-of-articulation (Kuhl & Miller, 1978; Lisker & Abramson, 1970). (For any given VOT value, velar stops, with their longer F1 transition, tend to have a lower F1 onset frequency than alveolar and labial stops. This was definitely true of the Lisker and Abramson stimuli used in both the human and chinchilla studies).

Let me turn now to the case of the voiced-voiceless distinction in word-medial stop consonants (e.g., rabid vs. rapid). As has been well documented by Lisker (1977; Lisker et al. 1969) and others, the differences between the voiced sound [b] and the voiceless sound [p] in medial position include the following: [b] is produced with a shorter lip closure duration, the [b] closure is preceded by a longer vowel, and the [b] closure interval contains glottal pulsing. Earlier perceptual studies of medial voicing have demonstrated robust trading relations between closure duration and presence or absence of glottal pulsing on the one hand (Lisker, 1978) and between closure duration and preceding vowel duration on the other (Raphael, 1972).

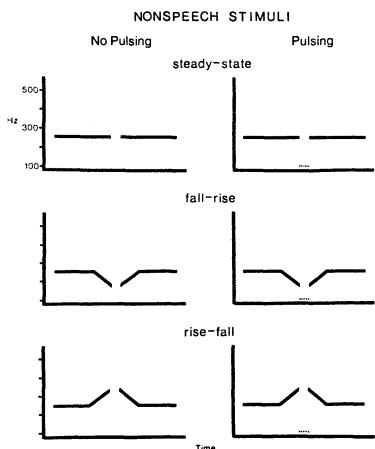


FIGURE 2, Fundamental frequency (F0) contours of the steady-state, fall-rise, and rise-fall square-wave stimuli used by Parker, Diehl, and Kluender (1986). The stimuli on the right contain glottal pulsing in the medial gap, whereas those on the left do not.

We decided to try to replicate these speech effects and then to test whether parallel perceptual effects could be obtained with nonspeech analogue stimuli. In one study (Parker, Diehl & Kluender, 1986), two speech stimulus series, both ranging perceptually from [aba] to [apa], were created by varying the closure interval of the medial stop. The two series differed only with respect to the presence or absence of glottal pulsing during the closure interval. Both stimulus sets were generated by digitally editing a token of [apa] produced by a male talker. For the pulsing stimulus series, the closure intervals contained a glottal buzz from a naturally produced [aba], whereas for the no-pulsing stimulus series, the closure interval contained only silence.

We also prepared several sets of nonspeech stimuli that mimicked the temporal and peak amplitude properties of the speech pulsing and no-pulsing stimuli. Figure 2 displays the fundamental frequency contours of these nonspeech stimulus sets. In the first set,

every item consisted of two steady-state square-wave segments equal in duration to the pre- and post-closure segments, respectively, of the speech stimuli. One nonspeech stimulus series in this set, depicted at the upper left, had silent intervals of varying duration separating the square-wave segments. These intervals corresponded exactly in duration to the silent closure intervals of the speech no-pulsing items. The other series, represented on the top right, was identical, except that the intervals separating the square-wave portions contained the same segments of glottal buzz that were used in the speech pulsing stimuli.

The second set of square-wave stimuli was identical to the first, except that fundamental frequency decreased linearly from 256 Hz to 175 Hz over the last 40 ms of the initial square-wave segment and then rose again over the initial 40 ms of the second square-wave segment. We call these the fall-rise stimuli. Finally, in the third square-wave set, the direction of fundamental frequency change was reversed in the vicinity of the medial gap, from 256 Hz to 337 Hz and then back again. These we refer to as the rise-fall stimuli.

In a given stimulus condition, subjects first were trained with feedback to press one of two buttons corresponding to each of the series-endpoint stimuli. Then during the experimental session, subjects listened to the entire series and were asked to label each item by pressing the button corresponding to the series-endpoint stimulus most similar to that item.

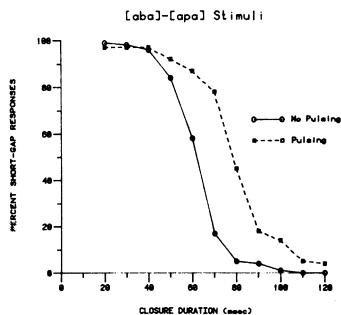


FIGURE 3, Mean percentage of short-gap responses to the /aba/-/apa/ stimuli used by Parker, Diehl, and Kluender (1986).

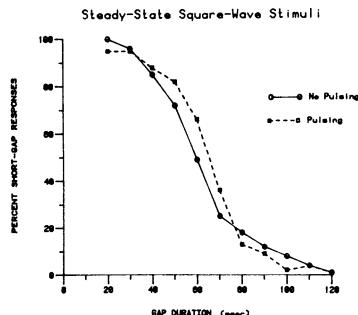


FIGURE 4, Mean percentage of short-gap responses to the steady-state square-wave stimuli used by Parker, Diehl, and Kluender (1986).

Figure 3 shows the labeling results for the speech stimuli. Here the percentage of responses corresponding to the short-medial-gap endpoint stimulus is plotted as a function of closure duration. Notice that the presence of glottal pulsing during the closure interval produces a shift in the labeling boundary toward more short-gap (or [b]) responses. This is, of course, what we would expect from previous studies of this trading relation.

Now, let us consider the results for the three nonspeech stimulus sets, starting with the steady-state square-wave stimuli,

presented in Figure 4. The shift in the labeling boundary is in the same direction as that for the speech stimuli, but the effect is small and not statistically significant. Figure 5 displays the results for the fall-rise square-wave stimuli. Again, there is a boundary shift, with pulsing contributing to a greater percentage of short-gap responses, and this time the effect is significant ($p < .05$). Finally, Figure 6 shows the results for the rise-fall square-wave stimuli. Here there was no reliable boundary between the pulsing and no-pulsing conditions.

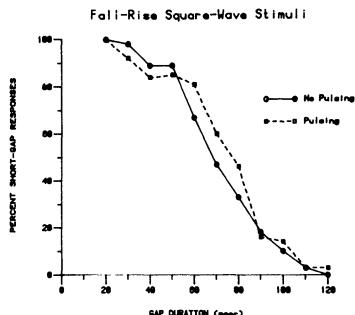


FIGURE 5, Mean percentage of short-gap responses to the fall-rise square-wave stimuli used by Parker, Diehl, and Kluender (1986).

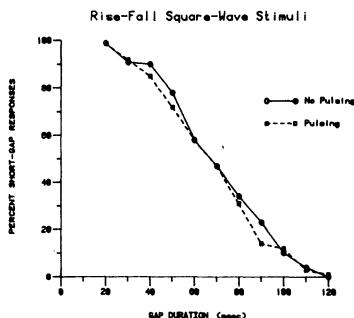


FIGURE 6, Mean percentage of short-gap responses to the rise-fall square-wave stimuli used by Parker, Diehl, and Kluender (1986).

Thus, the trading relation between closure duration and closure pulsing observed for speech sounds such as [aba] and [apa] can be at least partially duplicated with nonspeech analogue stimuli that are not phonetically categorizable. And the parallel between speech and nonspeech goes further than this. Note that the effect of pulsing occurred for the fall-rise stimuli but not for the rise-fall stimuli. In natural utterances, voiced stop consonants are usually produced with a rising FO contour following the closure release, whereas voiceless stops are often produced with a falling FO after the release (House & Fairbanks, 1953; Klatt, 1975; Lehiste & Peterson, 1961). Moreover, these FO differences serve as perceptual cues to the voiced-voiceless distinction in stops (Fujimura, 1971; Haggard, Ambler & Callow, 1970; Haggard, Summerfield & Roberts, 1981). We suggest that closure pulsing and a low FO value after release have mutually reinforcing auditory effects, perhaps because they both contribute to a sustained period of low-frequency periodic energy (Stevens, Keyser & Kawasaki, 1986). This would help to explain the pattern of our nonspeech results and would also account for the cue value of FO contours in natural speech. (Our view here is consistent with some recent work of Kingston, 1986, who showed that FO variation is apparently not just an unavoidable physical byproduct of the voiced-voiceless contrast. Instead, he argued that FO is deliberately regulated for its cue value in languages where voicing is distinctive rather than allophonic.)

What is the likely auditory basis of the trading relation between closure duration and closure pulsing? Our guess is that the presence of glottal pulsing during the initial portion of the closure

interval may effectively shorten the perceived duration of the closure, biasing the listener toward short-gap (i.e., voiced) responses.

Recently, we conducted a similar study of the perception of medial voicing, this time looking at the trading relation between closure duration and preceding vowel duration (Kluender, Diehl & Wright, 1985). First, two series of speech stimuli ranging from [aba] to [apa] were created from a naturally produced [apa] by varying the silent closure interval from 20 to 110 ms. For the long vowel series, we iterated pitch periods during the first vowel to produce a 245 ms initial syllable. For the short vowel series, the same pitch periods were excised, leaving an initial syllable duration of 188 ms. We also created square-wave analogues of these speech stimuli, along the lines of the previous experiment. The square-wave segment durations were equal to the corresponding syllable durations, and the speech closure intervals were precisely duplicated as silent gaps in the square-wave stimuli. The experimental procedure was very similar to that of the last experiment.

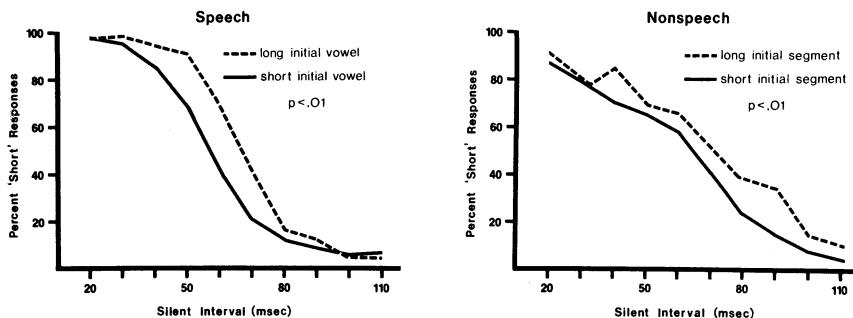


FIGURE 7, Percentage of short-gap responses to the [aba]-[apa] stimuli (left) and the square-wave stimuli (right) used by Kluender, Diehl, and Wright (1985).

In Figure 7, we see the labeling results for both the speech and the nonspeech stimuli. The percentage of responses corresponding to the short-medial-gap endpoint stimulus is plotted as a function of the medial silent interval. Although the variance in the labeling functions is greater for the nonspeech stimuli, the size of the boundary shift is comparable in the two cases.

A reasonable auditory hypothesis is that the medial gap duration is judged relative to the initial segment duration in both speech and nonspeech. A longer initial segment makes a given gap duration seem shorter by contrast. In the case of speech, this means that a longer initial syllable will make a medial stop appear more voiced.

The use of vowel length as a correlate of voicing in medial and final stops is, of course, not unique to English; it has very nearly the status of a phonetic universal (Chen, 1970). Over the years, a variety of articulatory or phonatory explanations have been offered to account

for this universal, but as Lisker (1974) has argued, none of these production-based explanations survives careful analysis. Our results suggest that the universal may, in fact, derive from auditory factors.

Let me try to make this point in a more general way. Those who favor some version of the speech-mode hypothesis tend to view perceptual trading relations as being rooted in the physics and physiology of speech production. The various acoustic correlates of, say, medial voicing are seen as necessary byproducts of the way speech sounds are produced. Either through learning or evolutionary adaptation, human listeners are able to assess the cue value of these acoustic correlates, and trading relations arise accordingly.

An alternative perspective, supported by the kinds of results described here, is that many important trading relations in speech perception are based on properties of the general auditory system. Certain cue combinations are selected by a speech community because they have mutually enhancing auditory effects. (For a similar view, see Stevens, Keyser & Kawasaki, 1983). By this account, voiced medial stops (to take one example) are produced as they are not simply out of physical or physiological necessity, but rather because a short closure appears even shorter (i.e. less voiceless) if preceded by a relatively long vowel and if partially filled with glottal pulsing.

In conclusion, we do not deny that speech categorization requires tacit knowledge of a rather considerable number of speech-specific facts. However, to the extent that a speech trading relation can be directly explained by properties of the auditory transfer function, it need not be assigned to the listener's store of tacit knowledge, and the overall perceptual model is thereby simplified. An appeal to speech-specific knowledge should always be, in other words, an explanation of last resort.

REFERENCES

1. Anderson, J.R. (1983). The architecture of cognition. Cambridge, MA: Harvard University Press.
2. Chen, M. (1970). Vowel length variation as a function of the voicing of the consonant environment. Phonetica, 22, 129-159.
3. Diehl, R.L., Lang, M., and Parker, E.M. (1980). A further parallel between selective adaptation and contrast. Journal of Experimental Psychology: Human Perception and Performance, 6, 24-44.
4. Fant, G. (1960). Acoustic theory of speech production. The Hague: Mouton.
5. Fodor, J.A. (1983). The modularity of mind. Cambridge, MA: The MIT Press.
6. Fujimura, O. (1971). Remarks on stop consonants: Synthesis experiments and acoustic cues. In: Form and substance: Phonetic and linguistic papers presented to Eli Fischer-Jorgenson, 11th February, 1971, 221-232. Akademisk Forlag
7. Haggard, M.P., Ambler, S., and Callow, M. (1970). Pitch as a voicing cue. Journal of the Acoustical Society of America, 47, 613-617.
8. Haggard, M.P., Summerfield, A.Q., and Roberts, M. (1981). Psychoacoustical and cultural determinants of phoneme boundaries:

- Evidence from trading FO cues in the voiced-voiceless distinction. Journal of Phonetics, 9, 49-62.
9. Hillenbrand, J. (1984). Perception of sine-wave analogs of voice onset time stimuli. Journal of the Acoustical Society of America, 75, 231-240.
10. House, A.S. and Fairbanks, G. (1953). The influence of consonant environment upon the secondary acoustical characteristics of vowels. Journal of the Acoustical Society of America, 25, 105-113.
11. Kingston, J. (1986). Are FO differences after stops accidental or deliberate? Journal of the Acoustical Society of America, 79, S27 (Abstract).
12. Klatt, D.H. (1975). Voice onset time, frication, and aspiration in word-initial consonant clusters. Journal of Speech and Hearing Research, 18, 686-706.
13. Kluender, K.R., Diehl, R.L., and Wright, B.A. (1985). Perception of duration of medial silent intervals in speech and nonspeech signals. Journal of the Acoustical Society of America, 77, S27 (Abstract).
14. Kuhl, P.K. and Miller, J.D. (1975). Speech perception by the chinchilla: The voiced-voiceless distinction in alveolar plosive consonants. Science, 190, 69-72.
15. Kuhl, P.K. and Miller, J.D. (1978). Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. Journal of the Acoustical Society of America, 63, 905-917.
16. Lehiste, I. and Peterson, G.E. (1961). Some basic considerations in the analysis of intonation. Journal of the Acoustical Society of America, 33, 419-425.
17. Liberman, A.M. (1982). On the finding that speech is special. American Psychologist, 37, 148-167.
18. Liberman, A.M., Cooper, F.S., Shankweiler, D.P., and Studdert-Kennedy, M. (1967). Perception of the speech code. Psychological Review, 74, 431-461.
19. Liberman, A.M. and Mattingly, I.G. (1985). The motor theory of speech perception revised. Cognition, 21, 1-36.
20. Lisker, L. (1974). On "explaining" vowel duration variation. Glossa, 8, 233-246.
21. Lisker, L. (1975). Is it VOT or a first-formant transition detector? Journal of the Acoustical Society of America, 57, 1547-1551.
22. Lisker, L. (1977). Rapid vs. ravid: A catalogue of acoustic features that may cue the distinction. Paper presented at the 94th meeting of the Acoustical Society of America, Miami Beach.
23. Lisker, L. (1978). On buzzing the English /b/. Paper presented at the 95th meeting of the Acoustical Society of America, Providence, R.I.
24. Lisker, L. and Abramson, A.S. (1970). The voicing dimension: Some experiments on comparative phonetics. In: Proceedings of the 6th International Congress of Phonetic Sciences, Prague, 1967. (pp. 563-567). Prague, Czechoslovakia: Academia.
25. Lisker, L., Abramson, A.S., Cooper, F.S. and Schvey, M.H. (1969). Transillumination of the larynx in running speech. Journal of the Acoustical Society of America, 45, 1544-1546.
26. Mann, V.A. and Repp, B.H. (1980). Influence of vocalic context on perception of the /ʃ/-/s/ distinction. Perception & Psychophysics, 28, 213-228.
27. Marr, D. (1982). Vision. San Francisco, CA: W.H. Freeman and Company.

28. Miller, J.D., Wier, C.C., Pastore, R.E., Kelly, W.J., and Dooling, R.J. (1976). Discrimination and labeling of noise-buzz sequences with varying noise-lead times: An example of categorical perception. *Journal of the Acoustical Society of America*, **60**, 410-417.
29. Newell, A. and Simon, H. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.
30. Parker, E.M. (1985). Auditory constraints on phonetic categorization: Trading relations in speech and nonspeech. Unpublished doctoral dissertation, University of Texas, Austin, Texas.
31. Parker, E.M., Diehl, R.L., and Kluender, K.R. (1986). Trading relations in speech and nonspeech. *Perception & Psychophysics*, **39**, 129-142.
32. Pastore, R.E. (1981). Possible psychoacoustic factors in speech perception. In: P. Eimas and J. Miller (Eds.), *Perspectives on the Study of speech*, 165-205. Hillsdale, NJ: Erlbaum.
33. Pisoni, D.B. (1977). Identification and discrimination of the relative onset time of two component tones: Implications for voicing perception in stops. *Journal of the Acoustical Society of America*, **61**, 1352-1361.
34. Pisoni, D.B., Carrell, T.D., and Gans, S.J. (1983). Perception of the duration of rapid spectrum changes in speech and nonspeech signals. *Perception & Psychophysics*, **34**, 314-322.
35. Raphael, L.J. (1972). Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English. *Journal of the Acoustical Society of America*, **51**, 1296-1303.
36. Repp, B.H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin*, **92**, 81-110.
37. Schouten, M.E.H. (1980). The case against a speech mode of perception. *Acta Psychologica*, **44**, 71-98.
38. Stevens, K.N., Keyser, S.J., and Kawashima, H. (1986). Toward a phonetic and phonological theory of redundant features. In: J.S. Perkell and D.H. Klatt (Eds.), *Invariance and variability in speech processes*, 426-449. Hillsdale, NJ: Erlbaum.
39. Summerfield, A.Q. (1982). Differences between spectral dependencies in auditory and phonetic temporal processing: Relevance to the perception of voicing in initial stops. *Journal of the Acoustical Society of America*, **72**, 51-61.
40. Summerfield, A.Q. and Haggard, M. (1977). On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants. *Journal of the Acoustical Society of America*, **62**, 435-448.

STUDIES OF POSSIBLE PSYCHOACOUSTIC FACTORS UNDERLYING SPEECH PERCEPTION*

Donald G. Jamieson

Speech and Audition Laboratory, Department of Psychology,
University of Calgary, 2500 University Dr. N.W.
Calgary, Alberta T2N 1N4, Canada

As with all sounds which are presented to the ear, speech sounds are subject to processing by the auditory system. Thus, whatever their importance for the listener, whatever the message conveyed, and however the listener may use linguistic knowledge, context, or other non-acoustic cues to assist in decoding the message from a speaker, the speech sounds which are presented to the listener's ear are subject to the same "obligatory processing" as are nonspeech sounds. In some situations, we might expect that such obligatory processing would result in the diminution of the perceptual salience of an acoustic cue--for example, as a consequence of masking between two portions of a signal. In other circumstances, we might expect that the auditory system would enhance the perceptual salience of an acoustic cue--for example, as a consequence of spectral-temporal integration. Of course the extent to which both perceptual factors and higher-order variables influence the responses which are actually observed in an experiment will reflect the demands of the particular task which is set for the listener (cf., the paper by MacMillan, Braida, Goldberg, and Khazatsky, in this volume, for further discussion of this point). In part, speech psychophysics attempts to isolate sensory and higher-order factors to characterize the influence of such "obligatory processing" on acoustic signals, with a view to understanding how such processing influences speech perception.

Of course we already know that, in some respects, speech sounds are well-matched to the capabilities of the human auditory system. For example, we know that the sounds of speech occur within the region of greatest spectral sensitivity (e.g., Robinson and Dadson, 1956). Moreover, the notion of natural auditory sensitivities appeals for several reasons: such natural sensitivities complement the constraints associated with speech production mechanisms in helping to explain why different languages tend to use similar acoustic landmarks

*The research reported here was supported by grants from the Natural Sciences and Engineering Research Council, Health and Welfare Canada, and the Alberta Heritage Foundation for Medical Research. I am grateful to Fred Wightman and Terry Dolan for their hospitality while I was a Visiting Fellow at the Waisman Center, University of Wisconsin, Madison, and to Peter Assman, Meg Cheesman, Vince Dilollo, Blas Espinoza-Varas, Tom Johnson, Linda McEvoy, Terry Nearey, Curtis Ponton, Mike Procter, Anton Rozsypal, and Susan Rvachew for their advice and assistance at various stages of the project. Special thanks are due to Elzbieta Slawinska with whom much of this work was undertaken.

(e.g., Lisker and Abramson, 1970), and they offer a mechanism by which human infants could classify speech sounds in ways that are similar to those of the skilled talker (e.g., Eimas and Miller, 1980; Jusczyk, Pisoni, Walley, and Murray, 1980; Morse, 1972).

Indeed, it has been suggested that several phonetic distinctions may depend on specific auditory constraints (e.g. Pastore, 1981; Schouten, 1980; Stevens, 1981, 1984; but see Howell and Rosen, this volume, and Rosen and Howell, this volume, for an alternative view). For the most intensively studied distinction, several authors have demonstrated that tone-onset time discrimination (TOT; the discrimination of the relative onset times of two overlapping tones) shows certain parallels with voice-onset time (VOT) discrimination: both are non-monotonic functions of time, with maximum discriminability around 30-50 ms, and in both cases, discriminability seems to depend on frequency (cf. Diehl, this volume; Hillenbrand, 1984; Jusczyk, et al., 1980; Pastore, this volume; Pastore, Harris and Kaplan, 1981; Pisoni, 1977).

However, with the exception of the TOT research, the details of the possible influence of auditory system processing on speech perception have remained sketchy, as a consequence both of the paucity of research on human auditory sensitivity to complex, speech-like sounds, and of our still-limited understanding of what acoustic cues we use when we listen to speech. This chapter summarizes the results of a series of studies which have examined the possible influence of auditory system processing on sensitivity to frequency transition cues.

For consonant-vowel syllables, it is well established that both the direction of formant frequency transitions and their duration provide relevant linguistic information (cf., Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967; Dorman, Cutting, and Raphael, 1975). First, transition duration reflects the speed with which the vocal tract changes shape, and thus provides an important cue to the manner of speech sound production. For example, formant transitions between the appropriate frequencies will cue /wa/ if they take 60 ms or more, while they will cue /ba/, if they take less time. Second, the direction of rapid (40-60 ms) formant transitions provides an important cue to the place of articulation of a sound, distinguishing the stop consonants /da/, /ga/, and /ba/. The alveolar stop /da/ is cued by an initial, falling second formant (F2) transition, together with a rising first formant (F1) transition, while the velar stop /ga/ is cued by a more steeply-falling F2, with the same rising F1, and the bilabial stop /ba/ is cued by a rising F2 and a rising F1 (cf., Potter, Kopp, and Kopp, 1966; Liberman, Ingemann, Lisker, Delattre, and Cooper, 1959; Holmes, Mattingly, and Shearme, 1964).

Is it possible that the way the auditory system processes speech-like signals causes sounds with formant transitions in the range 40-60 ms to have special importance for speech? In fact, unless the auditory system does influence perception in this manner, one would expect faster transitions to be more important in speech. First, by analogy to Weber's Law, the ability to discriminate between transitions should be better when the transitions occur over a shorter interval of time. Thus, a fixed, 10 ms difference in transition duration should be more discriminable with short transition durations (such as 20 vs.

30 ms) than with longer transition durations (such as 40 vs. 50 ms). Thus, performance should decrease monotonically as transition duration increases, and in fact it does when listeners discriminate transitions which are presented in isolation from their following steady state (Jamieson and Slawinska, 1983). Second, while formant transitions contain more information per unit time than other voiced speech, faster transitions would permit still higher rates of information transfer (cf., Shannon and Weaver, 1949). Third, the motor system is capable of faster transitions, and a variety of formant transition rates occur for different consonant-vowel syllables. Moreover, the 40-60 ms interval required for voiced syllable-initial stop consonants tends to resist compression when the overall speaking rate is increased (Port, 1981).

For these reasons, one might postulate that the manner in which the human auditory system processes signals which are comprised of a transition followed by a steady state tends to decrease the perceptual salience of very rapid transitions, while increasing the perceptual salience of intermediate (perhaps 40-60 ms) transitions. In turn, such changes in perceptual salience would influence which formant transitions tended to be used in speech--i.e., those which could best be discriminated by the listener would be favored in the evolution of language.

In fact, available psychoacoustic evidence does suggest that certain frequency transition rates may be favored: Nabelek and Hirsh (1969) found that 30 ms glides were better discriminated than either 10 or 100 ms glides, when the final frequency of the glide was continued as a steady state pure tone. In their substantial experiment, Nabelek and Hirsh allowed the duration of the steady state to vary

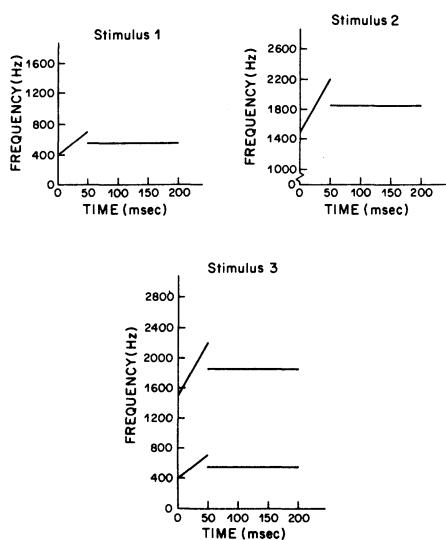


FIGURE 1, Samples of the stimuli used by Jamieson and Slawinska (1983). The stimuli in Series 1 consist of linear tone-sweeps beginning at 400 Hz and ending at 700 Hz, followed by a steady state pure-tone at the midpoint of the transition (i.e., 550 Hz). The stimuli in Series 2 consist of linear tone-sweeps beginning at 1500 Hz and ending at 2000 Hz, followed by a steady state pure-tone at the transition midpoint. The stimuli in Series 3 were a combination of both of the above complexes. Within each series, transition duration varied in 10 ms steps from 10 ms to 90 ms, while steady state duration covaried from 190 to 110 ms, to maintain overall stimulus duration at a constant 200 ms.

with the duration of the glide (the ratio of steady state to glide was fixed at 3:1 in all cases). Since they did not sample the range between 30 ms and 100 ms, it is entirely possible that the actual maximum sensitivity lies beyond 30 ms. To investigate this possibility, we undertook further examination of the possible special relationship between discriminability and the duration of frequency transitions.

More than a dozen experiments have now been conducted to determine whether auditory processing might produce a natural auditory sensitivity for speech formant transitions which occur over 40-60 ms, and if so, how such a sensitivity might arise. Elzbieta Slawinska and I have conducted experiments using nonspeech stimuli in order to examine the possibility that humans are especially prepared to perceive certain types of frequency transitions. Our first experiments (Jamieson and Slawinska, 1983) used signals beginning with a brief sinusoidal frequency glide followed by a longer, steady-state pure tone (cf., Figure 1). We fixed the steady state tone at the transition-center frequency, rather than at the transition-final frequency to minimize the possibility that our pure tone stimuli might be heard as speech (Remez, Rubin, Pisoni, and Carell, 1981), while otherwise mirroring a normal CV speech sound quite closely. Glides varied in duration from 10 ms to 90 ms, in 10 ms steps, while the duration of the steady state covaried from 190 to 110 ms to maintain an overall stimulus duration of 200 ms. Subjects listened to pairs of glide-tone stimuli, and indicated whether or not they heard a difference between the sounds. Our listeners were University staff and students, aged 25-41. All had normal hearing as determined by air-conduction pure-tone audiometry, and were tested individually, in a double-walled, IAC sound attenuating chamber. Each trial consisted of: a 1000 ms intertrial interval preceding the first stimulus; the presentation of the first stimulus; a silent 400 ms interstimulus interval; the presentation of the second stimulus, which was the same as the first on 50% of the trials and which differed from the first by a constant amount (10 or 20 ms, in different experiments) on the remaining trials; and a 4000 ms response interval during which the observer indicated whether the stimuli just presented were the same or different. Stimuli were presented at a comfortable listening level, adjusted separately for each subject at the beginning of the experiment. These levels varied from 68 to 83 dBA.

One experiment examined the discriminability of pairs of glide-tone stimuli when the glide covered the frequency range appropriate for the F1 center frequency of the syllable /ba/. In a second experiment, the glides covered the range appropriate for the F2 center frequency of /ba/. In a third experiment, the sounds were a mixture of the stimuli of the previous experiments, which approximated the combined variations in F1 and F2 for /ba/.

All our listeners displayed a non-monotonic discrimination function in this task, with a single maximum in the 40-60 ms range (see Figures 2 and 3). These results hold for both 10 and 20 ms step sizes, for both the F1 and F2 frequency ranges, and for the combination (F1 and F2) stimulus. They also held for a subsequent series of experiments (Jamieson and Slawinska, 1984) in which the steady state pure-tone was fixed at the final frequency of the transition. Thus, the results are not specific to the difficulty of the discrimination, nor to the rate of change of the transition, nor to the

range of frequencies covered by the transition. The effect is not changed by extensive practice with accuracy feedback.

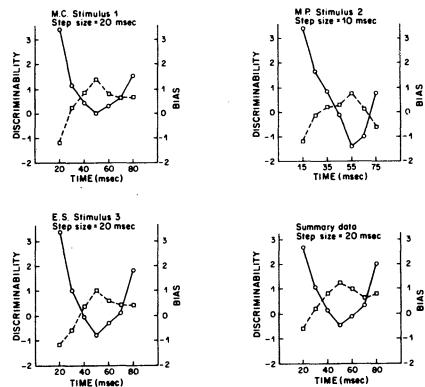


FIGURE 2, Examples of functions relating discriminability (dashed lines: $-\ln \eta$; cf., Luce, 1963) to transition duration and of the bias functions (solid lines; $\ln \beta$; cf., Luce, 1983) for data from Jamieson and Slawinska (1983). Discrimination performance is typically best when the glide takes place over 40-60 ms, independent of stimulus condition. This effect is found over a wide range of variations in the frequency excursion of the transitions and in the relations of the steady state frequency to the frequency of the transition (cf., Jamieson and Slawinska, 1983). Subsequent experiments (Jamieson and Slawinska, 1984) have shown that the bias function does not always show a minimum near the peak of the discrimination function.

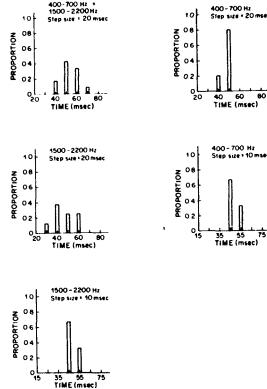


FIGURE 3, Examples of histograms showing the distributions of discrimination maxima for data from Jamieson and Slawinska (1983). Discrimination performance is typically best when the transition takes place over 40-60 ms, independent of stimulus condition.

The middle panel displays a histogram showing the distribution across subjects of the transition duration values which yield maximum discriminability ($-\ln \eta$) when the transition is followed by a contiguous steady state pure-tone, and overall stimulus duration remains fixed at 200 ms.

The bottom panel displays a histogram showing the distribution across subjects of the transition duration values which yield maximum discriminability ($-\ln \eta$) when the transition alone is presented. In this case, the discrimination function typically follows a Weber-like function.

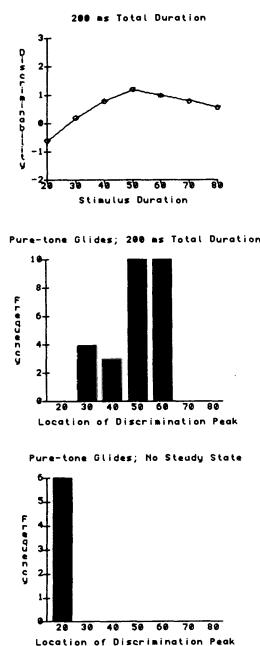


FIGURE 4. The top panel displays a plot relating discriminability ($-\ln n$; cf., Luce, 1963) to transition duration, when the overall stimulus duration was fixed at 200 ms and the steady state duration was equal to the transition duration.

The middle panel displays a histogram showing the distribution of discrimination function maxima, when overall stimulus duration was fixed at 200 ms, and the steady state duration was equal to the transition duration. Performance is typically best when the transition takes place over 40-60 ms, independent of stimulus condition.

The bottom panel displays a histogram showing the distribution of discrimination function maxima when there is no steady state after the transition.

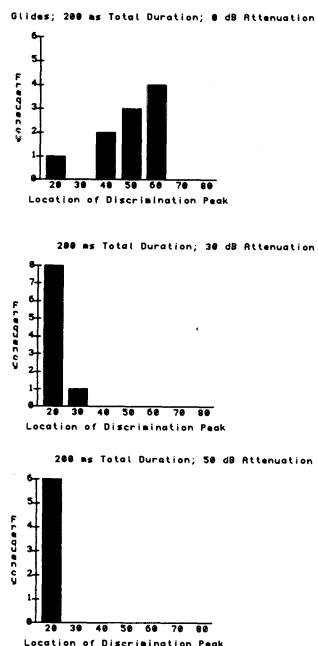


FIGURE 5. The three histograms summarize the distribution, across individual subjects, of the peak of the discrimination functions for various transition-steady state continua, as the amplitude of the steady state is attenuated, relative to transition. The full amplitude condition is displayed in the top panel, the 30 dB attenuation condition is displayed in the middle panel, and the 50 dB attenuation condition is displayed in the bottom panel. The leftward shift in the peak of the discrimination function with increasing attenuation of the steady state amplitude, is clearly visible in this sequence of panels. Data are from Jamieson and Slawinska (1984).

(Caption Figure 4, continued)

Performance typically follows a Weber-like function, being best when the transition takes place over 10 ms and declining monotonically as transition duration increases. Situations intermediate between the middle and bottom panels tend to produce discrimination functions with maxima between 10 and 50 ms.

In what way are these non-monotonic discrimination functions related to speech perception? First, the discrimination peak when transitions take, say, 50 ms corresponds to a tendency to place speech sounds with formant transitions which are shorter than 50 ms in one category (e.g., /ba/) and transitions which are longer than 50 ms in another category (e.g., /wa/); such categorization maximizes the between-class perceptual differences between phones, while maximizing the within-class perceptual similarity. Second, the enhanced discriminability for frequency transitions which take 40-60 ms may help to explain why the transitions cueing place of articulation distinctions occur within this interval: the enhanced perceptual salience for small changes in transitions which occur over 40-60 ms may improve sensitivity to transition direction in this region.

Above, it was postulated that the discrimination functions displayed in Figure 2 represented the manner in which the human auditory system processes signals which are comprised of a transition followed by a steady state: auditory processing tends to decrease the perceptual salience of slower (40-60 ms duration) transitions. Perhaps the easiest way to understand these changes in the perceptual salience of transitions, in "conventional psychoacoustic terms", is to postulate some form of masking of the transition part of the stimulus by the steady state portion. This notion is consistent with the results that discrimination functions show a peak sensitivity at 40-60 ms when a steady state follows the transition, but that the functions generally follow Weber's Law when the transition portion of the stimulus is presented alone, without the steady state. This explanation also predicts that parametric variations in the steady state portion of the stimulus will shift the peak of the discrimination function. For example, reducing the duration, or the amplitude, of the steady state should decrease the masking of the transition by the steady state; with less masking, the discrimination peak should shift to a shorter transition duration. Figure 4, taken from Jamieson and Slawinska (1984), shows that precisely this result occurs when the duration of the steady state is reduced parametrically. Figure 5, also taken from Jamieson and Slawinska (1984), shows that a parallel result occurs when the amplitude of the steady state is reduced parametrically.

The effect of shortening the steady state (vowel) duration on listeners' perception of synthetic speech CVs varying along a "manner" continuum from /ba/ to /wa/, is well-known: listeners accept faster transition stimuli as exemplars of /wa/ (Miller and Liberman, 1979), as the vowel duration is decreased. This leftward shift in the category boundary as the steady state is shortened parallels the leftward shift in the peak of the discrimination function for our stimuli, when the steady state is shortened. However, the mechanisms to which these effects are attributed are very different: Miller and Liberman interpreted their result as due to a "rate normalization" process in which the listener infers a more rapid rate of speech because of the shortened vowel; whereas at a slower speech rate, a given transition

duration (say 40 ms) would be taken to represent a /ba/ sound, at this faster (inferred) speech rate, the same transition duration is taken to represent /wa/. However, based on our results with nonspeech sounds, at least some portion of the "rate normalization" effect is attributable to a more basic, auditory processing--perhaps a type of masking.

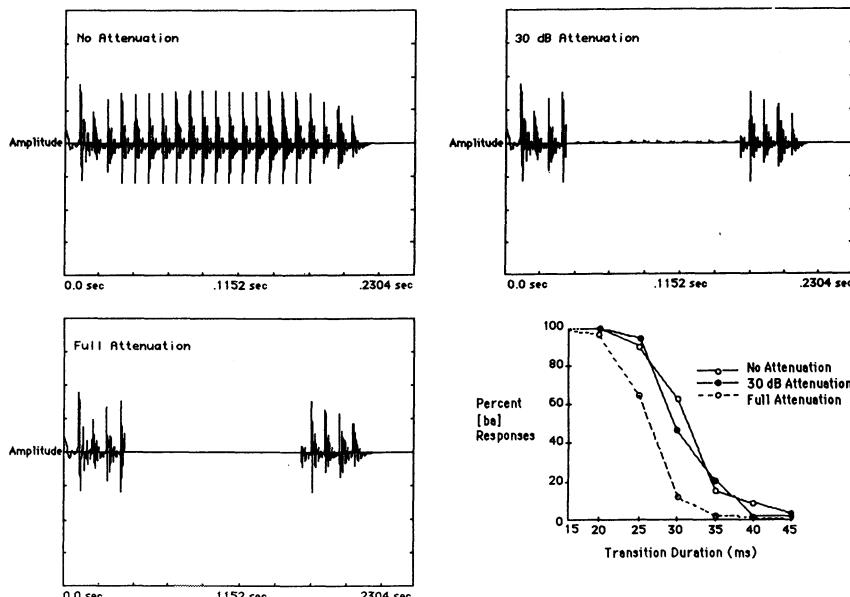


FIGURE 6. Three panels display examples of stimuli from midpoint /bad/-/wad/ continua, with vowel amplitude attenuated as indicated in the legend. The panel in the lower right hand corner displays identification functions for these three attenuation conditions. There, the shift in the category boundary to the left, indicating that listeners accept faster transitions as /wad/ as vowel amplitude is reduced, is visible. Data are from Jamieson, Johnson, and Rvachew (1986).

To further investigate this latter possibility, we wanted to study the effects of parametrically reducing the amplitude of the steady state, relative to the transition portion of the signal, on the identification of speech sounds. Attenuating the vowel in CV by even 20 dB destroys most of the vowel quality. However, in a CVC environment, the vowel can be completely attenuated, while preserving its identity (Strange, Jenkins, and Johnson, 1983). Jamieson, Johnson, and Rvachew (1986) used this technique to synthesize a typical, full-amplitude continuum varying from /bad/ to /wad/ in seven steps. This continuum was then digitally altered to produce a series of continua in which only the initial 10 ms of the steady state was at full amplitude. After 10 ms, the steady state was attenuated by a specified amount. Leaving the vowel amplitude fixed for the first 10 ms of the steady state ensures that the vowel is readily heard, and it also works against our hypothesis, since the effects of the subsequent reduction in "masking", due to attenuation of the remaining portion of the vowel, would be minimized. However, identification functions do change, and in the predicted direction, as a consequence of vowel attenuation. Three of the conditions are shown in Figure 6: no attenuation; 30 dB attenuation; and full attenuation. As can be seen

from the figure, identifications of /wad/ required shorter duration transitions as the amplitude of the steady state was reduced. Thus, just as Miller and Liberman demonstrated that reducing the duration of the steady state shifts the category boundary to shorter transition durations, this experiment shows that reducing the amplitude of the steady state shifts the boundary of identification functions in a similar fashion. However, in natural language, steady state (vowel) amplitude does not undergo a natural parametric variation relative to transition amplitude. Thus, unlike the "rate-normalization" interpretation of the duration manipulation, there is no obvious articulatory referent for the listener in the case of our amplitude manipulation. An implication is that some or all of the "rate-normalization" effect which Miller and Liberman induced by varying vowel duration reflects a type of masking by the steady state on the transition portion of the syllable. This masking is reduced when either the duration or the amplitude of the steady state is reduced. The effect is similar whether the transition/steady state complex takes the form of a CV syllable or of the pure-tone sequences used in the experiments described earlier in the paper.

What implications do these results have for speech perception? First, they suggest that a form of intra-speech masking occurs in speech, as a consequence of the contiguity of transition/steady state sequences. This masking alters the perception of transitions, inducing one or more types of natural auditory sensitivity to "emerge". The first type of natural auditory sensitivity can be termed "warping", since the listener's sensitivity to the acoustic continuum of stimuli is "warped" to reduce sensitivity to some stimuli (for example with our stimuli, for very fast transitions, 10-30 ms in duration), relative to others (with our stimuli, for intermediate transitions, 40-60 ms); this warping produces a maximum in the discrimination function for transition/steady state stimuli at around 50 ms, and encourages listeners to place stimuli with faster transitions in one category and those with slower transitions in a second category. In turn, this perceptual sensitivity would encourage languages to use distinctions between sounds which are based on the duration of transitions.

A second type of "natural auditory sensitivity" which may be induced by the interaction of the steady state with the transition, could be termed "magnification". Here, the region of enhanced sensitivity increases listeners' sensitivity to a different type of cue. For example, the fact that listeners are most sensitive to transitions which take about 40-60 ms, would encourage languages to have distinctions which are based on the slope or direction of frequency transitions taking about 40-60 ms; listeners would be less sensitive to such cues if the transitions took slightly more or slightly less time.

The details of the interaction between the steady state and the transition portion of the stimulus also account for much or all of the "rate-normalization" effect which Miller and Liberman induced by varying vowel duration. Changes in vowel amplitude produce effects which parallel "rate normalization". However, unlike duration, vowel amplitude does not vary parametrically relative to transition amplitude in natural language. Thus, there is no obvious articulatory referent for the listener in the case of our amplitude manipulation. What is in common between the two situations is that the auditory system performs its normal, obligatory processing on both speech and non-

speech signals, enhancing some apparent acoustic cues and attenuating others; as a consequence, at least part of the "rate normalization" effect seems to be attributable to reductions in some form of complex "masking" by the steady state on the preceding consonant. It seems likely that an improved understanding of such auditory processes as those described above will play a major role in improving our understanding of speech perception.

REFERENCES

1. Dorman, M., Cutting, J., and Raphael, L. (1975). Perception of temporal order in vowel sequences with and without formant transitions. Journal of Experimental Psychology: Human Perception and Performance, 104, 121-129.
2. Eimas, P.D. and Miller, J.L. (1980). Contextual effects in infant speech perception. Science, 209, 1140-1141.
3. Hillenbrand, J. (1984). Perception of sine-wave analogs of voice onset time stimuli. Journal of the Acoustical Society of America, 75, 231-240.
4. Hirsh, I.J. (1959). Auditory perception of temporal order. Journal of the Acoustical Society of America, 75, 231-240.
5. Holmes, J.N., Mattingly, I., and Shearme, J. (1964). Speech synthesis by rule. Language and Speech, 7, 127-143.
6. Howell, P. and Rosen, S. (1983). Natural auditory sensitivities as universal determiners of phonemic contrasts. Linguistics, 21, 205-235.
7. Jamieson, D.G., Johson, T., and Rvachew, S. (1986). A role for intra-speech masking in "rate-normalization" on a stop-semivowel continuum. Alberta Conference on Language, Banff.
8. Jamieson, D.G. and Slawinska, E.B. (1983). Sensitivity to rate-of-change of frequency transition. Journal of the Acoustical Society of America, Suppl. 1 74, S67.
9. Jamieson, D.G. and Slawinska, E.B. (1984). The discriminability of transition duration: Effects of the amplitude and duration of following steady state. Journal of the Acoustical Society of America, Suppl. 1 76, S29.
10. Jusczyk, P., Pisoni, D., Walley, A., and Murray, J. (1980). Discrimination of relative onset time of two-component tones by infants. Journal of the Acoustical Society of America, 67, 262-270.
11. Liberman, A., Cooper, F., Shankweiler, D., and Studdert-Kennedy, M. (1967). Perception of the speech code. Psychological Review, 74, 431-461.
12. Liberman, A.M., Ingemann, F., Lisker, L., Delattre, P., and Cooper, F. (1959). Minimal rules for synthesizing speech. Journal of the Acoustical Society of America, 31, 1490-1499.
13. Lisker, L. and Abramson, A.S. (1970). The voicing dimension: Some experiments in comparative phonetics. Proceedings of the sixth international congress of phonetic sciences. Prague: Academia, 563-567.
14. Luce, R.D. (1963). Discrimination. In: R. Luce and E. Galanter (Eds.), Handbook of Mathematical Psychology, Wiley: New York.
15. Miller, J.L. and Liberman, A.M. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. Perception & Psychophysics, 25, 457-465.
16. Morse, P. (1972). The discrimination of speech and nonspeech in early infancy. Journal of Experimental Child Psychology, 14, 477.

17. Nabelek, A. and Hirsh, J.I. (1969). On the discrimination of frequency transitions. Journal of the Acoustical Society of America, 45, 1510-1519.
18. Pastore, R. (1981). Possible psychoacoustic factors in speech perception. In: P. Eimas and J. Miller (Eds.), Perspectives on the Study of Speech, Lawrence Erlbaum: Hillsdale, N.J.
19. Pastore, R.E., Harris, L.B., and Kaplan, J.K. (1981). Temporal order identification: Some parameter dependencies. Journal of the Acoustical Society of America, 71, 430-436.
20. Pisoni, D. (1977). Identification and discrimination of the relative onset time of two component tones. Implications for voicing perception in stops. Journal of the Acoustical Society of America, 61, 1352-1361.
21. Port, R. (1981). Linguistic timing factors in combination. Journal of the Acoustical Society of America, 69, 262-274.
22. Potter, R., Kopp, G., and Kopp, H. (1966). Visible Speech. New York: Dover Publications.
23. Remez, R., Rubin, P., Pisoni, D., and Carell, T. (1981). Speech perception without traditional speech cues. Science, 212, 947-950.
24. Robinson, D. and Dadson, R. (1956). A re-determination of the equal-loudness relation for pure tones. British Journal of Applied Physics, 7, 166-181.
25. Rosen, S. and Howell, P. (1986). Auditory, articulatory and learning factors in categorical perception. In: S. Harnad (Ed.). Categorical perception, Cambridge: Cambridge University Press (in press).
26. Schouten, M.E.H. (1980). The case against a speech mode of perception. Acta Psychologica, 44, 71-98.
27. Shannon, C. and Weaver, W. (1949). The Mathematical Theory of Communication, Urbana, Illinois: University of Illinois Press.
28. Stevens, K. (1981). Constraints imposed by the auditory system on the properties used to classify speech sounds: Data from phonology, acoustics and psychoacoustics. In: T. Myers, J. Laver, and J. Anderson (Eds.), The Cognitive Representation of Speech. Amsterdam: North Holland.
29. Stevens, K.N. (1984). Evidence for the role of acoustic boundaries in the perception of speech sounds. Speech Communication Group: Working Papers Volume IV, Cambridge, M.A.: Research Laboratory of Electronics, MIT.
30. Strange, W., Jenkins, J., and Johnson, T. (1983). Dynamic specification of coarticulated vowels. Journal of the Acoustical Society of America, 74, 695-705.

PERCEPTION OF TONE, BAND, AND FORMANT SWEEPS

Louis C.W. Pols and M.E.H. Schouten

Institute of Phonetic Sciences, University of Amsterdam
and Institute of Phonetics, University of Utrecht

1. INTRODUCTION

Natural speech can be described as an ongoing sequence of dynamic events with a certain linguistic meaning. Whatever speech perception model one prefers, undoubtedly these dynamic events have to be analyzed and interpreted by the listener. There is a substantial literature about how specific dynamic events, such as formant transitions, are labelled, identified, discriminated, memorized, masked, matched, or adapted to, in the context of speech-like stimuli. However, surprisingly little has been published about the basic properties of our hearing for analyzing and perceiving such dynamic events in the form of psychophysical stimuli. What is the just-noticeable-difference, what is the detection threshold, what is the internal dynamic spectrum? It almost seems that the use of dynamic (speech) stimuli is more common practice in electro-physiological studies (e.g. Proceedings edited by Carlson and Granström, 1982) than in psychophysics. The locus theory of plosive perception, for instance, requires some form of extrapolation of formant transitions, whether or not via an innate language module which refers to articulatory gestures (Liberman and Mattingly, 1985). The model proposed by Lindblom and Studdert-Kennedy (1967) about perceptual compensation for formant frequency undershoot, implies similar perceptual capabilities. What is needed for this at the peripheral level is that the listener is able to perceive and interpret the short transitions occurring in CV- and VC-type syllables.

This paper reports about an ongoing series of experiments on identification and discrimination of various types of dynamic stimuli ranging from single-tone and two-tone sweeps, via single- and multiple-band or formant sweeps, to natural speech segments eventually. We will also report about preliminary experiments which have been done on the matching of synthetic formant transitions.

In another series of experiments we studied the information present in natural band sweeps for consonant identification. These natural band sweeps (vocalic transitions) were segmented from CV- and VC-parts in isolated words or free conversation, and were then presented to subjects for identification of the intended consonant (Klaassen-Don, 1983; Klaassen-Don and Pols, 1984; Pols, 1979; Pols and Schouten, 1978, 1981, 1982; Schouten and Pols, 1983, 1984). A subset of these stimuli was also presented in its original sentence context for two-alternative forced-choice plosive identification (Pols, 1986; Pols and Schouten, 1985; Schouten and Pols, 1984).

Finally, we refer to experiments about the identification of vowel segments taken from natural speech to which more and more of the vocalic transition towards the surrounding consonants is added (Buiting et al., 1983; Koopmans-van Beinum et al., 1984).

2. IDENTIFICATION AND DISCRIMINATION OF TONE SWEEPS

We considered the upward or downward sweep of a single pure tone the simplest and most basic dynamic signal to be studied. We asked subjects to identify such short tone sweeps as either going up or down (Schouten, 1985), and later added a third response alternative, namely "level" (Schouten, 1986). A tone sweep is completely specified by its centre frequency, form of transition function, sweep rate, and duration. In order to prevent onset and offset clicks, a time window is

Table I, Various parameters for the different identification experiments.

Reference	duration (ms)	sweep rate (oct/s)	F_C (Hz)	sweep type	response categories	number of subjects
Expt. 1a Schouten (1985) "tone sweep"	20	0	400	(1) up	(a)	12
	30	5	1300	(2) down		
	40	10	2700			
	50	20		noise-		
		40		sweep-		
		60		noise		
Expt. 1b Schouten (1985) "two-tone sweep"	20	0	400+	(1) up	(b)	12
	30	5	1300	(2) down		
	40	10		(3)		
		20	1300+	noise-		
		30	2700	sweep-		
		40		noise		
Schouten (1986) "tone sweep 3-way choice"	20	0	1300	(1) up	(c)	15
	30	5		(2) down		
	40	10		noise- level		
	50	20		sweep-		
		40		noise		
		60				
Expt. I Schouten and Pols (1986) "band sweep"	15	0	1300	(1) up	(d)	15
	20	5		(2) down		
	25	10	BW=			
	30	20	200 Hz	3-ms		
	35	30		cosine		
	40	40		window		
Pols and Schouten (1986) "band sweep"	15	0	to or	(1-2) up	(e)	plan
	20	5	from	(3-4) down		
	25	10	1300			
		20		3-ms		
		30		cosine		
		40	200 Hz	window		

often applied. However, because of the sometimes very short duration of the stimuli, we preferred to add 100 ms of low-level white noise at both ends of the tone sweeps (Pols and Schouten, 1981). The form of the transition function which we used is exponential, which in fact is a linear function if expressed in terms of octaves per second. In Table I the values of the various parameters are given for the different identification experiments.

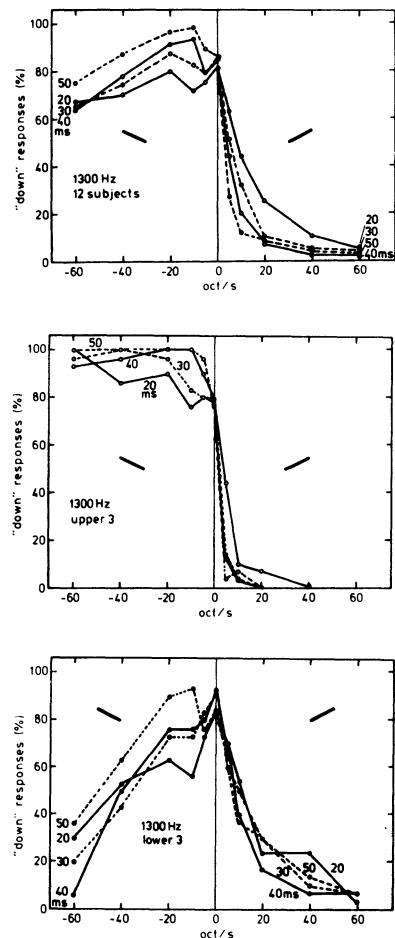


FIGURE 1, Percentages of "down" responses averaged over all 12 subjects and over the upper and lower quartiles. The direction of the tone sweep is indicated in each panel. The centre frequency was 1300 Hz.

In Fig. 1 the percentage of "down" responses, averaged over all 12 subjects, as well as averaged over the upper and lower groups of three subjects, are presented as a function of sweep rate, with sweep duration as a parameter. Each left-hand panel represents the falling stimuli, whereas the rising stimuli are on the right-hand side. Only the results for a centre frequency of 1300 Hz are presented here; for more details see Schouten (1985). The percentage of "down" responses is displayed instead of the percentage of "correct" responses, because in this way it is more apparent that, especially for the lower-quartile-subjects, zero- or low-sweep rates are judged to move "down", irrespective of the actual physical direction. There is a great preference for level stimuli to be judged as going down. Especially the poorer listeners have a fairly wide range of sweep rates which they apparently are unable to distinguish from steady tones and which, as a result, receive a majority of "down" responses. This implies that in a discrimination experiment we must expect that rising and steady tones will be better discriminated than falling and steady tones, or that falling sweeps and steady tones have more in common perceptually than rising sweeps and steady tones.

Before going on to the discrimination results, we will briefly present the identification results for the two-tone sweeps. For a specification of the parameters, see Table Ib, for the percentage of down responses, see Fig. 2. Here again we see the same strong bias: low sweep rates are interpreted as steady-state stimuli and get "down" responses; the greater the sweep rate (either up or down), the more the average listener is inclined to regard it as going up.

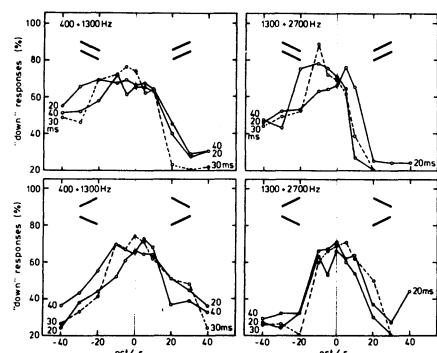


FIGURE 2, Percentages of "down" responses averaged over all 12 subjects for the double-tone sweeps, the directions and the centre frequencies of which are indicated in the panels.

Table II, Various parameters for the different discrimination experiments.

Reference	duration (ms)	sweep rate (oct/s)	F_c (Hz)	sweep type	response type	number of subjects
Expt. 2 Schouten (1985)	10 15 20 25 30 "tone sweep"	0 5 10 20 30 40	1300	(1) \— (2) \/\— (3) —/\— noise- sweep- noise	difference in first or second pair @	12
Expt. II Schouten and Pols (1986) "band sweep"	15 20 25 30 35 40	0 5 10 20 30 40	1300 BW = 200 Hz 3-ms cosine window	(1) /— (2) \— (3) \/\— same 3-ms cosine window	same @ plan	15
Pols and Schouten (1986) "band sweep"	15 20 25 30 35 40	0 5 10 20 30 40	to or from 1300 3-ms cosine window	(1) \/\— (2) ---\— same 3-ms cosine window	plan @	plan

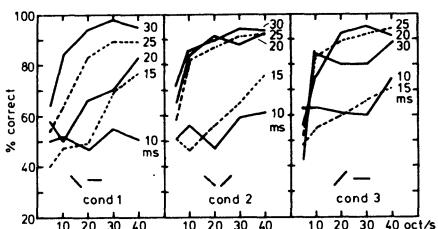


FIGURE 3, Correct sweep discrimination scores averaged over all 12 subjects for the three indicated pairs of tone sweeps.

Since, apparently, most subjects have great difficulty in specifying whether a tone sweep goes up or down, it is a more natural and more realistic question (with respect to its usefulness in speech perception) to ask whether subjects are able to discriminate between rising, falling, and steady tones and formants. For a specification of the experimental parameter, see Table IIa. We used a so-called four-interval forced-choice paradigm: two pairs of stimuli, of which three stimuli are identical whereas one is different, are presented, and the subject only has to indicate whether the first or the second pair contains the different stimulus.

The percentage correct scores, averaged over the 12 subjects, are given in Fig. 3 for the three conditions: falling vs. steady, rising vs. steady and falling vs. rising (in this last condition only equal, positive, and negative sweep rates were compared). The scores for the 10-ms stimuli are at chance level, and we see a substantial improvement towards 30-ms stimuli. If we take 75% as a criterion, we get minimal combinations of sweep rate and sweep duration as indicated in Table III. Discrimination of rising sweeps requires sweep rates of at least 10 oct/s and durations of at least 20 ms; discrimination of falling sweeps requires longer durations and/or higher sweep rates. Thus we can probably conclude for tones (and perhaps for formant transition as well?) that they have to rise at a rate of at least 10 oct/s for at least 20 ms, or they will be "processed" as falling transitions.

Table III, Critical discriminability values (75% correct) for the indicated pairs of tone and band sweeps.

	down vs. level	down vs. up	up vs. level
single-tone sweep	20 ms, 40 oct/s 25 ms, 20 oct/s	20 ms, 10 oct/s	20 ms, 10 oct/s
single-band sweep	20 ms, 30 oct/s 25 ms, 20 oct/s 35 ms, 10 oct/s	20 ms, 20 oct/s 25 ms, 10 oct/s 30 ms, 5 oct/s	20 ms, 40 oct/s 25 ms, 20 oct/s 30 ms, 10 oct/s

Schouten (1986) showed that the tendency to perceive level stimuli as falling sweeps does not disappear entirely if subjects are given a three-way choice (up-down-level); see Table Ic. Especially the poorer subjects are still guided strongly by this down-preference.

3. IDENTIFICATION AND DISCRIMINATION OF BAND SWEEPS

Experiments with band sweeps are a natural extension of the experiments with tone sweeps described above. This time the stimuli were made by sweeping a (digital) filter with a bandwidth of 200 Hz through the harmonics of a 200 Hz pulse train. Whatever the sweep rate or the duration, the centre frequency of 1300 Hz occurred exactly halfway along the time course of a stimulus. Instead of adding noise, the onsets and offsets were smoothed this time by means of a 3-ms

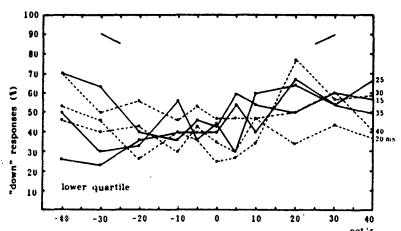
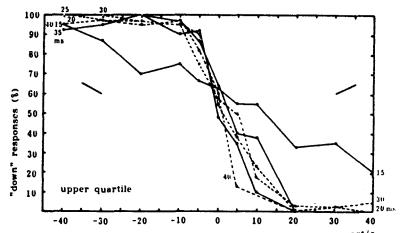
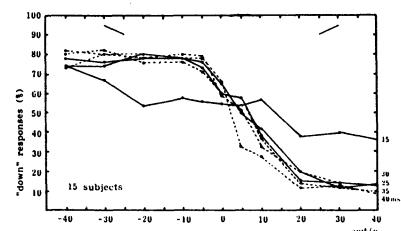


FIGURE 4, Percentages of "down" responses, averaged over all 15 subjects and over the upper and lower quartiles. The direction of the band sweep is indicated in each panel. The centre frequency was always 1300 Hz and the bandwidth 200 Hz.

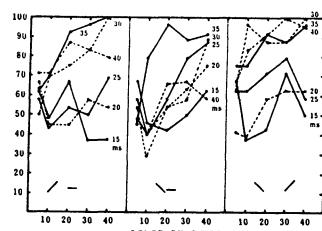
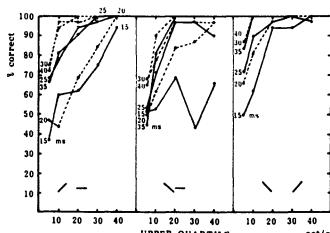
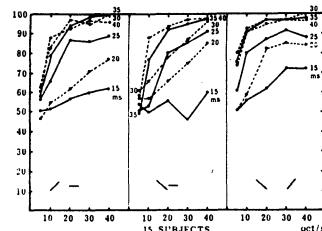


FIGURE 5, Correct sweep discrimination scores, averaged over all 15 subjects and over the upper and lower quartiles, for the three indicated pairs of band sweeps.

cosine window. Some specifications are given in Table Ia, for more details see Schouten and Pols (1986). The percentage of "down" responses for all 15 subjects, as well as for the upper and lower quartiles, are given in Fig. 4. Contrary to the tone sweeps, the responses from the lower quartile this time appear to be completely random. In the other two panels, 15-ms stimuli show substantially lower scores than do all longer on durations. There is again a degree of asymmetry: the scores reach asymptotic values at around -5 oct/s and +20 oct/s, and these asymptotic values themselves are also different: 80% and 90% respectively. There is again some tendency to perceive zero and slowly rising sweeps as falling, although this effect is not nearly as pronounced as it was for the tonal sweeps.

We next performed a discrimination experiment, similar in design to that for the tone sweeps. For some specifications see Table IIb, for more details see Schouten and Pols (1986). The percentage correct scores are given in Fig. 5. The lower panel shows that the lower quartile, which happens to be the same group of three subjects as in the identification experiment, performs substantially better than random for this discrimination task. Evidently it is easier to discriminate band sweeps from each other and from stationary bands, than it is to identify the direction of the sweep in a consistent way as going up or down. As for the tone sweeps, the difference in discrimination scores is largest between falling vs. level and falling vs. rising, indicating that level and rising stimuli are most different from each other. The tendency to perceive level and slowly rising stimuli as falling is still there but less pronounced with these band sweeps than with the tone sweeps. Using 75% correct as a discriminability measure we come up with the critical values as given in Table III. Comparison with the tone sweeps shows that most of the time it is easier to discriminate tone sweeps than band sweeps; this is most apparent for the rising versus level condition.

4. SUBSEQUENT STEPS

The tendency to perceive level and slightly rising stimuli as falling, which was quite strong in the case of tone sweeps, is much less strong for band sweeps. It does exist however, and it can be expected to play a role in speech perception. Our expectation is that it will not diminish if we attach steady signals (e.g. 100 ms) to the onsets or offsets of our single-band sweeps; for strict comparison we will first rerun the earlier experiment but now with 1300 Hz as the initial or final frequency of the band sweep (see Tables Ie and IIc). It is much more difficult however, to foresee what will happen if two band sweeps are sounded simultaneously: the difference between excitation patterns possibly caused by falling and rising sweeps may then disappear altogether, especially if the two band sweeps are in different directions, as often happens in speech. A subset of these stimuli will subsequently be used to make them even more speech-like, for instance by adding a stationary first and fourth formant; we can then start using speech labels as in CV-and VC-stimuli, instead of "up/down" labels.

5. MATCHING SYNTHETIC FORMANT TRANSITIONS

Apart from the function of formant transitions in CVC-type syllables to point towards neighbouring consonants (in this way providing information for consonant identification), they of course serve the more basic purpose of signalling the vowel itself. However, because of articulatory dynamics and idiosyncrasies, there is a great deal of variation in vowel formant frequencies. Lindblom and Studdert-Kennedy (1967) assume context-dependent perceptual compensation for articulatory undershoot. Apart from identification (Lindblom and Studdert-Kennedy, 1967) and discrimination experiments (Danaher et al., 1973; Mermelstein, 1978), matching could be a sensitive method (Brady et al., 1961; Fujisaki and Sekimoto, 1975).

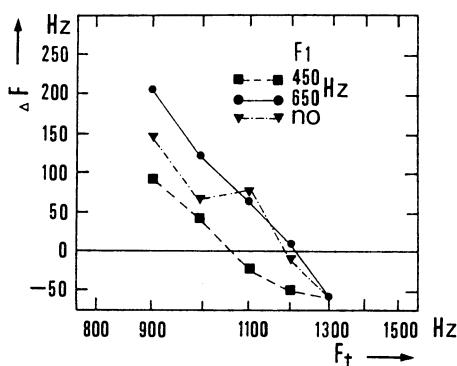


FIGURE 6. Matching results for one- and four-formant stimuli, in which F_2 drops from 1800 Hz to the indicated target frequency F_t . The difference between matched and target frequency (ΔF) is presented along the vertical axis, averaged over 12 subjects and two settings per subject. Two conditions refer to four-formant stimuli with F_1 at either 450 Hz or 650 Hz, whereas in the third condition the F_2 -transition only is present.

characteristics of the signal, like a difference in F_1 or absence of all other formants, the settings change. Various other formant trajectories were studied as well. The same stimuli which were used in the matching experiment were also presented for vowel identification. We have not been able to discover a straightforward relation between matching and identification data.

Whereas in vowel identification one frequently only has a choice of two vowel categories, in matching a much more specific judgment can be made. We used 200-ms four-formant synthetic stimuli in which one formant changed in the final 100-ms from a stationary value to a target value. The subject was requested to match a 70-ms stationary signal, of which the same formant could be varied in frequency, to the final value of the dynamic stimulus. Some typical results are given in Fig. 6; for more details see Pols et al. (1984). The deviation from the final frequency is represented as a function of that final or target frequency F_t . The initial frequency of F_2 was always 1800 Hz in this situation. If subjects had been able to match the final frequency very well, all points would have scattered around zero. However, for the smallest band sweeps (from 1800 Hz to 1300 Hz) there is some indication of overshoot, whereas for all other sweeps the matched frequency lagged behind. With different overall

6. GENERAL DISCUSSION

In this session of the symposium on the "Psychophysics of speech perception", in which "Dynamic aspects" get special attention, we have focussed in our paper on the perception of tone, band, and formant sweeps. We started from a psychophysical viewpoint by systematically studying the identification and discrimination of single and multiple tone and band sweeps. For untrained listeners we found a lower sensitivity than for trained listeners for such sweeps in the speech-like range and also a substantial asymmetry. Zero- or low-sweep rates are judged to move "down", irrespective of the actual direction. This effect is less prominent for band sweeps than it is for tone sweeps. Whether it is of general importance in speech perception remains to be seen when we go via double band sweeps to more speech-like signals. The inclusion of a stationary part preceding or following the formant transition might be another parameter for higher accuracy on the part of the listener.

In consonant identification experiments, in which only the vocalic parts of VC- and CV-type natural segments are presented, we see a reasonable sensitivity for the interpretation of formant transitions in terms of consonant labels. In these types of signals most of the time a more or less stationary part is included. But probably more important than the stationary part is the fact that in these experiments labels were used for which we have had life-long training.

The matching experiments so far have not given a clear indication about how formant transitions are interpreted and extrapolated, under conditions in which not an absolute identification is required but a psychophysical match.

For the time being we must accept a strong interaction between stimulus characteristics and task variables, while an easy extrapolation from psychophysical results to speech perception results does not seem to be possible.

REFERENCES

1. Brady, P.T., House, A.S., and Stevens, K.N. (1961). Perception of sounds characterized by rapidly changing resonant frequency. Journal of the Acoustical Society of America, 33, 1357-1362.
2. Buiting, H.J.A.G., Koopmans-van Beinum, F.J., and Pols, L.C.W. (1983). The role of transitions in the identifiability of vowels taken from conversational speech. Abstracts 10th Int. Congr. Phonetic Sciences, 507.
3. Carlson, R. and Granström, B. (1982). The representation of speech in the peripheral auditory system. Elsevier Biomedical Press, Amsterdam.
4. Danaher, F.M., Osberger, M.J., and Pickett, J.M. (1973). Discrimination of formant frequency transitions in synthetic vowels. Journal of Speech and Hearing Research, 16, 439-451.
5. Fujisaki, H. and Sekimoto, S. (1975). Perception of time-varying resonance frequencies in speech and non-speech stimuli. In: A. Cohen and S. Nooteboom (Eds.), Structure and process in speech perception, 269-282. Springer Verlag, Berlin.
6. Klaassen-Don, L.E.O. (1983). The influence of vowels on the perception of consonants. Ph.D. thesis, Univ. of Leyden.

7. Klaassen-Don, L.E.O. and Pols, L.C.W. (1984). The role of coarticulation in the identification of consonants. In: M.P.R. van den Broecke and A. Cohen (Eds.), Proc. 10th Int. Congr. Phonetic Sciences, 451-454. Foris Publications, Dordrecht.
8. Koopmans-van Beinum, F.J., Wouters, H.A.L., Buiting, H.J.A.G., and Pols, L.C.W. (1984). The influence of response categories on the identification of vowels excerpted from conversational speech. Proc. Inst. Acoustics, Vol 6, Part 4, 363-370.
9. Liberman, A.M. and Mattingly, I.G. (1985). The motor theory of speech perception revised. Cognition, 21 1-36.
10. Lindblom, B.E.F. and Studdert-Kennedy, M. (1967). On the role of formant transitions in vowel recognition. Journal of the Acoustical Society of America, 42, 830-843.
11. Mermelstein, P. (1978). Difference limens for formant frequencies of steady-state and consonant-bound vowels. Journal of the Acoustical Society of America, 63, 572-580.
12. Pols, L.C.W. (1979). Coarticulation and the identification of initial and final plosives. In: J. Wolff and D.H. Klatt (Eds.), ASA-50 Speech Communication Papers, 459-462.
13. Pols, L.C.W. (1986). Variation and interaction in speech. In: J.S. Perkell and D.H. Klatt (Eds.), Invariance and variability in speech processes, 140-154. Lawrence Erlbaum Associates, Hillsdale, N.J.
14. Pols, L.C.W., Boxelaar, G.W., and Koopmans-van Beinum, F.J. (1984). Study of the role of formant transitions in vowel recognition using the matching paradigm. Proc. Inst. Acoustics, Vol. 6, Part 4, 371-378.
15. Pols, L.C.W. and Schouten, M.E.H. (1978). Identification of deleted plosives. Journal of the Acoustical Society of America, 64, 1333-1337.
16. Pols, L.C.W. and Schouten, M.E.H. (1981). Identification of deleted plosives: The effect of adding noise or applying a time window. A reply to Ohde and Sharf. Journal of the Acoustical Society of America, 69, 301-303 (L).
17. Pols, L.C.W. and Schouten, M.E.H. (1982). Perceptual relevance of coarticulation. In: R. Carlson and B. Granström (Eds.) The representation of speech in the peripheral auditory system, 203-208. Elsevier Biomedical Press, Amsterdam.
18. Pols, L.C.W. and Schouten, M.E.H. (1985). Plosive consonant identification in ambiguous sentences. Journal of the Acoustical Society of America, 78, 33-39.
19. Schouten, M.E.H. (1985). Identification and discrimination of sweep tones. Perception & Psychophysics, 37, 369-376.
20. Schouten, M.E.H. (1986). Three-way identification of sweep tones. Perception & Psychophysics, 40, 359-361.
21. Schouten, M.E.H. and Pols, L.C.W. (1983). Perception of plosive consonants - The relative contributions of bursts and vocalic transitions. In: M.P.R. van den Broecke, V.J. van Heuven, and W. Zonneveld (Eds.), Sound structures: Studies for Antonie Cohen, 227-243. Foris Publications, Dordrecht.
22. Schouten, M.E.H. and Pols, L.C.W. (1984). Identification of intervocalic plosive consonants: The importance of plosive bursts vs. vocalic transitions. In: M.P.R. van den Broecke and A. Cohen (Eds.). Proc. 10th Int. Congr. Phonetic Sciences, 464-468. Foris Publications, Dordrecht.
23. Schouten, M.E.H. and Pols, L.C.W. (1986). Identification and discrimination of single formant sweeps. PRIPU, Vol. 11, Nr. 2, 21-33.

PSYCHOPHYSICAL REPRESENTATION OF STOP CONSONANT AND TEMPORAL MASKING IN SPEECH*

Christel Sorin
CNET, 22301 Lannion, France

There is at present a strong demand from colleagues working in Automatic Speech Recognition for a "better" representation of the speech signal as input for their systems. After the adoption of the psychoacoustical Mel scale, some of them hope that the knowledge of a "truer" internal auditory representation of the speech signal could reduce the variability they have to handle, and/or make more salient some useful cues for phonetic identification.

On the other hand, the explosion of physiological research on processing by the peripheral auditory system provides data on the internal representation of speech that can be of great usefulness for speech psychoacousticians willing to replace "the black boxes", too often used to describe auditory perceptual processes, by more precise models.

In this context, we are trying to measure, with psychoacoustical procedures, the internal spectral representation of time-varying speech signals.

Among the several experiments that have been reported, most have been restricted to the internal spectral representation of stationary vowels (Moore and Glasberg, 1983, Tyler and Lindblom, 1982, among others). We decided to investigate rather the spectral representation of intervocalic consonants.

However, the classical experimental techniques are not well suited to the study of time-varying signals. Measurements using simultaneous masking paradigms are liable to phase interactions between the complex masker components and the test-tone (Schroeder, 1982). These phase interactions cannot be controlled if natural speech maskers are used, and we insisted on using natural speech stimuli! Secondly, neither non-simultaneous masking nor pulsation threshold paradigms allow an "on-line" investigation of time-varying signals in context.

However, motivated by the issue of absolute vs relative invariance of burst spectral properties (Fant, 1985), we designed a first set of experiments on the perceptual processing of a velar intervocalic consonant. The questions we raised were the following:

*Many thanks to Martine Boyer for having been a long-suffering subject and to Francois Lonchamp for his help in writing the English version of this paper.

- What is (are) the spectral internal representation(s) of the burst of this intervocalic plosive?
- Is it dependent on the presence of a following sound through backward masking?

More generally:

- Is there more invariance at the psychoacoustic level than at the acoustic level?
- Does backward masking really play a role in speech perception?

I. MEASUREMENTS OF THE AUDITORY PROFILE OF THE INTERVOCALIC STOP BURST

A. The experiment

The "auditory spectrum" of the burst was measured using the forward masking procedure. As masker we used two natural tokens /AK/ and /UK/, obtained by removing, in both cases, the second vowel from the entire original token, /AKA/ and /UKU/ pronounced by a male speaker (figures 1a/ and 2a/). The cut was made at a waveform zero-crossing point and the end of the release burst was determined on the basis of short-term spectral change. The center of the 16 ms pure tone probe was located 10 ms after the end of the burst. The details of the masked threshold measurements are given in Appendix 1.

B. The results

The amounts of masking plotted in Figures 1b/ and 2b/ correspond to the differences between the probe masked levels and probe absolute thresholds for each frequency (i.e. the probe masked level was around 55 dB SPL at 2 kHz). They can be regarded as the auditory profiles of the bursts in the two contexts /AKA/ and /UKU/.

The 2 profiles are clearly different. Their global shapes reproduce well the burst's LPC spectra shown in Figures 1c/ and 2c/. These LPC burst shapes check with observations on velar consonants made by several authors, among whom Blumstein (1986): "for velars, there are several patterns of spectral compactness as a function of vocalic context. Velars in front of back vowels (/U/) display a frequency peak around 900 Hz and a high-frequency peak around 4.2 kHz with little energy in between. Velars in front of front vowels (such as /A/) display a fairly broad mid-frequency peak between 2.0 and 3.0 kHz."

This contrast is clearly maintained in the masking profiles. The auditory burst shape in /AK/ shows a relatively high-level maximum around 1.6 kHz, although the two small LPC peaks visible on both sides of 2 kHz are well reproduced in this profile. One should note also the strong masking effect of the low frequency part (below 1 kHz) of the burst.

The burst shape in the /UKU/ profile has a strong peak at 900 Hz, just as in the LPC spectrum. In this case, the "auditory contrast" is of the same order of magnitude as the acoustic (LPC) spectral contrast (roughly 30 dB between 900 Hz and 2.5 kHz). This result

supports the conclusions of Moore and Glasberg (1983) as to the preservation of spectral contrast in forward masking measurements, possibly because of suppression activity.

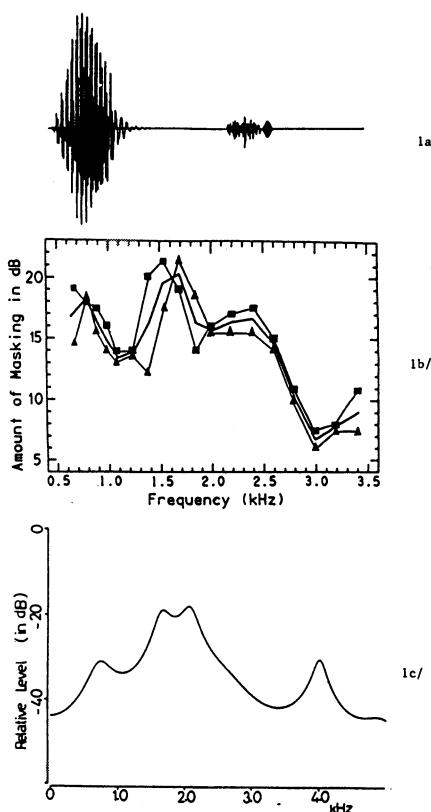


FIGURE 1
 1a/ Waveform of the /AK/ stimulus and of a 60 dB test tone at 1 kHz located just after the burst's offset.
 1b/ Amount of forward masking of the /K/ in the /AKA/ context for both subjects (symbols: individual data; bold line: mean over the two subjects).
 1c/ LPC Spectrum of the 40 ms /K/ burst in the /AKA/ context.

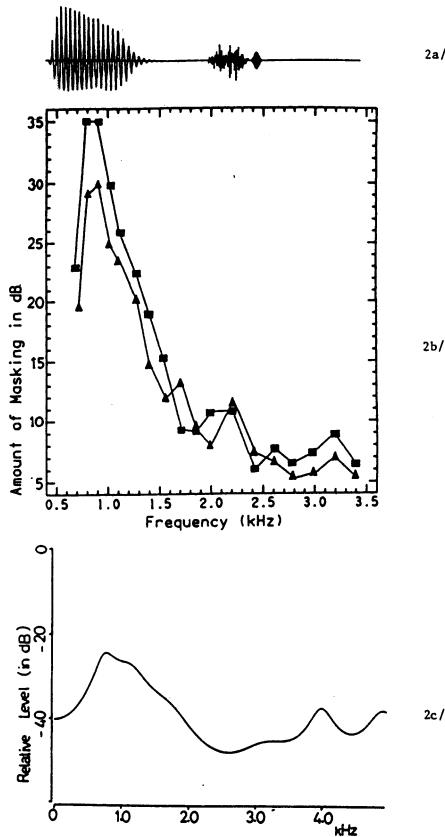


FIGURE 2
 2a/ Waveform of the /UK/ stimulus and of a 60 dB test tone at 1 kHz located just after the burst's offset.
 2b/ Amount of forward masking of the /K/ in the /UKU/ context for both subjects (symbols: individual data; bold line: mean over the two subjects).
 2c/ LPC Spectrum of the 40 ms /K/ burst in the /UKU/ context.

C. Discussion

The gross shape of the burst spectrum is well reproduced in the auditory profile obtained by the forward masking procedure. The spectral contrasts are even improved in the mid-frequency range. The different patterns of the burst's spectral compactness as a function of vocalic context are maintained in the auditory profiles. It can therefore be concluded that, in the absence of the following vowel, the internal spectral representation of the intervocalic burst is no more invariant than its acoustic spectrum.

Let us see now if the backward masking effect of the following vowel can strongly modify these internal spectral representations as hypothesized by Pastore (1981).

II MEASUREMENTS OF VOWEL'S BACKWARD MASKING EFFECT

A. The 3 experiments

We investigated the potential backward masking (BM) effect of the following vowel on the velar release burst in the /AKA/ context.

Three different paradigms were used to evaluate the amount of BM due to the second /A/ vowel.

In Experiment 1, the consonantal release burst was cut out from the speech signal as shown in Figure 3a/. The 16 ms pure tone probe was located just before the vocalic onset, so that the amount of BM due to the second /A/ vowel could be measured as the difference between the test-tone detection threshold in the presence of /A/ and the absolute threshold of the probe obtained when the vowel was removed (see Appendix 1).

In Experiment 2, the probe was located exactly at the same position on the /AKA/ token, but the consonantal burst was now present (Figure 4a/). Here, the probe is subjected to both simultaneous masking from the burst and to BM from the succeeding vowel. Note that the burst itself is also subjected to the BM effect of the vowel.

Therefore, the difference between the test-tone detection threshold with and without the following /A/ is a complex estimate of the BM of the vowel that acts on both the probe and possibly the burst. As concerns phase effects between burst components and probe, care was taken to locate the probe at exactly the same position within the burst and to control its phase at onset. Since we only computed the difference between the 2 conditions, the phase effects should be of minimal importance.

In experiment 3, a 16 ms silence interval was introduced between the burst offset and the vocalic onset. The probe was located in the silence interval (Figure 5a/). Therefore, the probe was subjected to both forward masking from the burst and to BM of the vowel. Here again, the amount of BM is taken as the difference between test-tone detection, with and without the vowel.

B. The results

We have plotted in Figure 3b/ the difference between probe masked levels in the 2 conditions of Experiment 1: when the vowel is present and when the /A/ vowel is removed. It is in fact the relative amount of BM that can be attributed to the final vowel. We can observe:

- that the amount of masking reaches 8 to 10 dB at the vowel formant frequencies,
- that the locations of the spectral shape maxima correspond to those of the vowel formants at onset (see, in Figure 3c/, the LPC spectrum of the vowel onset). However, the relative amplitudes of the formants are not respected (compensation for spectral tilt).
- a greater masking effect in the region of the first formant.

These results suggest that BM effects can influence the internal representation of weaker intervocalic sounds.

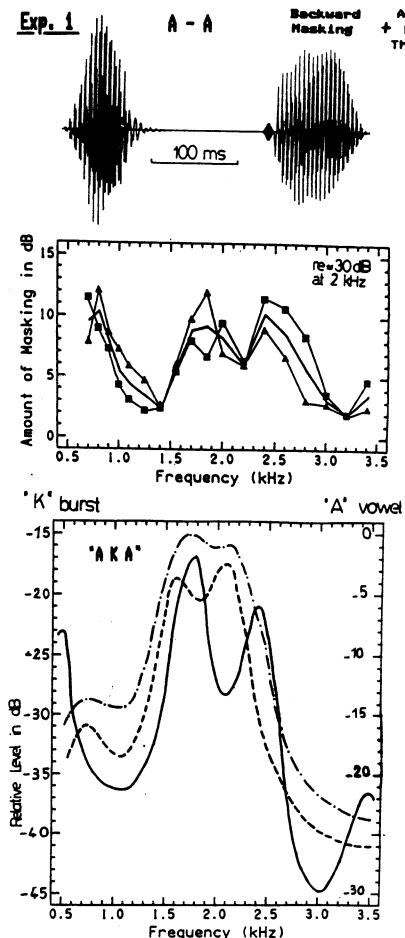


FIGURE 3,
Experiment 1

3a/ Waveform of the /A-A/ stimulus (burst removed) and of the probe located just before vocalic onset.

3b/ Amount of backward masking due to the second /A/ vowel, burst absent, for both subjects (symbols: individual data; bold line: mean over the two subjects).

3c/ LPC spectra of the /K/ burst and the /A/ vowel onset (— first 16 ms of /A/, - - - whole 40 ms /K/ burst, - · - last 20 ms of /K/ burst).

Figures 4b/ and 5b/ show, in the same manner, the relative amount of BM observed in experiments 2 and 3. In the upper right corner of each figure, the SPL masked levels of the probe are indicated. Note that the level at which the 2 kHz probe was just detected was only 30 dB in Experiment 1 (burst removed), went up to 45 dB in Experiment 3 (silence introduced) but reached 70 dB in Experiment 2 (in the presence of the burst). This makes it clear that the potential BM effect of the vowel was measured at 3 widely different probe levels.

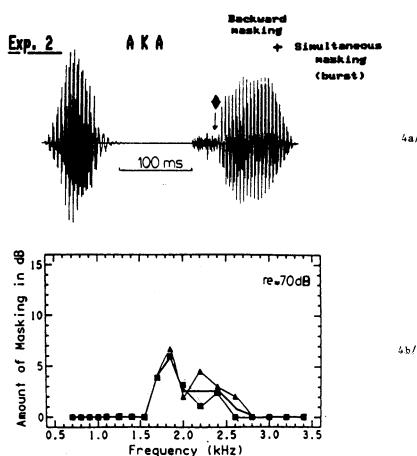


FIGURE 4, Experiment 2
 4a/ Waveform of the /AKA/ stimulus (arrow shows position of the probe tone).
 4b/ Amount of backward masking due to the second /A/ vowel, burst present (symbols: individual data; bold line: mean over the two subjects).

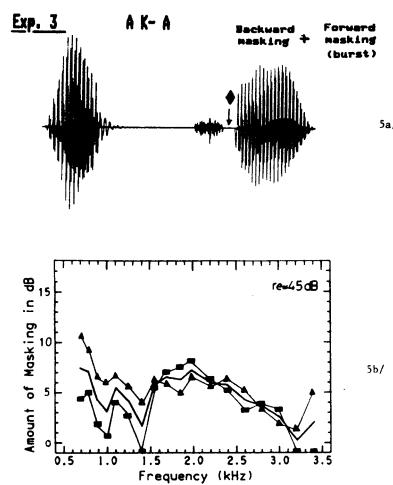


FIGURE 5, Experiment 3
 5a/ Waveform of the /AK-A/ stimulus, a 16 ms silence being introduced just before the vocalic onset (arrow shows position of the probe tone).
 5b/ Amount of backward masking due to the second /A/ vowel, a 16 ms silence being introduced just before the vocalic onset (symbols: individual data; bold line: mean over the two subjects).

It is obvious that the amount of BM revealed by Experiment 2 is very weak in comparison to that of the two other conditions. There may be several reasons for that. Firstly, one could say that the simultaneous masking effect of the consonantal burst on the probe overrides the BM from the vowel. More precisely, it increases the detection threshold of the probe to a level where BM of the vowel has

almost no effect. Secondly, one could also say that the BM effect acted on both the burst and the probe, so that the two effects are cancelled.

The dynamic range of the BM pattern obtained in Experiment 3 is also compressed in comparison to that obtained in Experiment 1. There are strong interindividual differences in the low frequency part and, in the F₂ - F₃ area, the formant structure has disappeared to be replaced by a broad plateau: the auditory spectral shape is smoothed relative to the acoustic spectral shape of the vowel.

C. Discussion

Concerning the potential role of vocalic BM on the gross shape of the auditory spectrum of an intervocalic stop-burst, we would like to offer the following comments.

Even if the real amount of vocalic BM is underevaluated by our simultaneous masking procedure, it cannot be greater than that observed in Experiment 1, where the burst was entirely removed. It is probably nearer to the masking amount observed in Experiment 3. Therefore, its effect could be a 5 to 10 dB reduction in the amplitude of the low frequency part of the burst (that is below 1.4 kHz). It could have a similar, but less pronounced effect (less than 5 dB) on the amplitude in the mid-frequency range (1.5 to 3 kHz). It is clear that subtracting this small amount of masking from the burst auditory profile in the /AKA/ context (fig. 1b/) cannot strongly modify its gross shape.

These results suggest that, at least for velar intervocalic consonants, the BM effect of the succeeding vowels cannot possibly modify the burst auditory profiles to the extent of making them similar (although the "correction" goes in the right direction). The specific velar burst characteristics, that is, a strong peak around 0.9 kHz for /UKU/ as opposed to a higher and wider one in the 1.5 - 2.5 kHz range in /AKA/, will be preserved in the auditory profiles. So the hypothesis attributing to BM an important function in preserving a context-independent burst shape across different vocalic contexts must probably be rejected.

To summarise, our results suggest that the BM effect caused by a vowel can be neglected

- if the vowel is preceded by a non-weak sound (BM is then overridden by a simultaneous masking effect)
- for delays, before vocalic onset, greater than 15 ms.

When BM interacts with Forward Masking from another sound (/AK-A/, Exp. 3):

- BM is weaker than Forward Masking
- its spectral shape is smoothed relative to the acoustic spectral shape of the vowel.

CONCLUSION

Our results do not reveal a strong effect of the following vowel on the stop burst auditory spectrum that could be attributed to backward masking. The "invariance" of the burst spectral shape seems no more "absolute" at the psychophysical than at the acoustic level.

Measuring, psychoacoustically, the "internal representations" of running speech sounds raises methodological questions. Further work will definitely require a close collaboration with psychoacousticians in order to refine or create procedures furnishing a realistic "internal representation" of non-stationary speech signals.

APPENDIX 1

Masked threshold measurements

The test-tone was in all conditions a 16 ms pure tone burst shaped with a Hanning window. The frequencies of the test tone were varied by 100 Hz steps from 0.7 to 1.1 kHz, by 150 Hz steps from 1.25 to 2.0 kHz and by 200 Hz steps from 2.0 to 3.4 kHz, in order to approximately match a critical-band spacing.

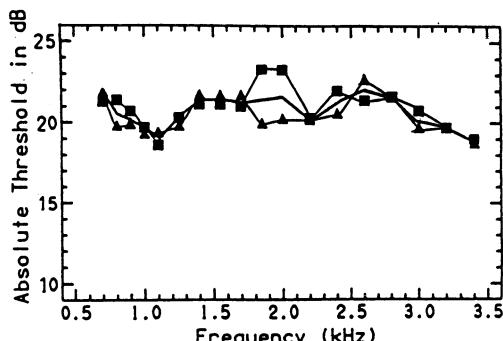
Thresholds were measured with an adaptive forced-choice method. The speech and "speech + test-tone" stimuli were presented in pairs, with a 500 ms inter-stimulus silence interval. A 3.5 s silence interval followed the subject's response. The initial level of the test tone was adjusted so that it was clearly audible.

The speech signals were presented monaurally via TDH 39 headsets at a 75 dB SPL overall level. Speech signals and test tones were added digitally and the modifications of the test tone levels were performed on line by the computer during the post-response interval.

In all the experiments, the mean thresholds for each subject were computed over at least 4 runs.

Two subjects were used, both with extensive practice in similar tasks. The absolute thresholds of the test tones were measured in the following way: half of the /K/ closure phase and the final /A/ were deleted and the test-tone was added 100 ms after the offset of the first /A/ vowel to avoid any forward masking effect of the /A/ vowel. This first /A/ stimulus acted as a warning signal. Figure 6 gives the test-tone absolute thresholds for each subject. Intra-subject standard deviation was about 1 dB.

FIGURE 6 (right column), Absolute thresholds of the 16 ms tone for both subjects (— mean over the two subjects, ▲— individual thresholds).



REFERENCES

1. Blumstein, S.E. (1986). On acoustic invariance in speech. In: J. Perkel and D. Klatt (Eds.) Invariance and Variability in Speech Processes, 178-197. Lawrence Erlbaum Ass.
2. Fant, G. (1986). Features - Fiction and Facts. In: J.S. Perkell and D.H. Klatt (Eds.) Invariance and Variability in Speech Processes, 480-492. Lawrence Erlbaum Ass.
3. Moore, B.C.J. and Glasberg, B.R. (1983). Masking patterns for synthetic vowels in simultaneous and forward masking. Journal of the Acoustical Society of America, 73, (3) 906-917.
4. Pastore, R.E. (1981). Possible psychoacoustic factors in speech perception. In: P.D. Eimas and J.L. Miller (Eds.) Perspectives on the Study of Speech, 165-205. Lawrence Erlbaum Ass., Hillsdale.
5. Schroeder, M.R. and Mehrgart, S. (1982). Auditory masking phenomena in the perception of speech. In: R. Carlson and B. Granström (Eds.) The Representation of Speech in the Peripheral Auditory System, 78- 88. Elsevier Biomedical Press, Amsterdam.
6. Tyler, R.S. and Lindblom, B. (1982). Preliminary study of simultaneous-masking and pulsation-threshold patterns of vowels. Journal of the Acoustical Society of America, 71, 220-224.

EFFECTS OF STIMULUS DYNAMICS ON FREQUENCY DISCRIMINATION

Francisco Lacerda

Institute of Linguistics, University of Stockholm,
S-106 91 Stockholm, Sweden

and

Centro de Linguistica da Universidade de Lisboa,
Av. 5 de Out. 85-6, P-1000 Lisboa, Portugal

INTRODUCTION

Physiological data shows that for certain stimulation levels, the discharge rate of an auditory-nerve fiber is roughly proportional to the level of stimulation. However, the same data shows as well that this proportional response attains a plateau if the stimulus level is increased above a certain limit (Schalk and Sachs, 1980). In other words, an auditory-nerve fiber has a certain dynamic range within which it can represent variations in the intensity of stimulation by varying its discharge rate, but for intensities above a certain level the discharge rate saturates and all the stimulation above that level is represented by the same firing rate. The dynamic range of a single fiber is small (of the order of a few tens of dB), although it varies somewhat with the fiber's spontaneous rate (SR) (Schalk and Sachs, 1980). As a consequence, the dynamic range of a frequency channel - even considering the combined action of fibers with different SRs - is rather small, and the auditory patterns that can be predicted on the basis of the single-fiber responses will appear with very broad spectral peaks even at moderate levels. Of course, non-linearities like lateral suppression play a role in the sharpening of the auditory patterns (Houtgast, 1974; Duifhuis, 1980) but nevertheless the degradation of the levels in those auditory representations is probably too big to be recovered by lateral suppression mechanisms.

An alternative explanation is that the auditory representations take advantage of the initial responses of auditory-nerve fibers. Since the response of an auditory-nerve fiber to a stimulus of a constant level begins with a wide dynamic response which is followed, shortly after, by a reduction in the dynamic range, as the fiber adapts to the stimulation level (Kiang, 1968; Delgutte, 1980; Smith and Brachman, 1980; Miller and Sachs, 1983), auditory representations before adaptation takes place are likely to produce more favourable representations of spectral peaks and valleys.

An experiment reported by Chistovich (1971) gives support to the idea that adaptation of the auditory-nerve fibers has direct consequences for the auditory representation of speech stimuli. The experiment consisted of adjusting the level of the F_2 of a four-formant vowel so that vowel quality would remain constant, even when the onset of the second formant was delayed relative to the onset of the other three formants. The results showed that as the onset of F_2 was delayed, its level had to be decreased in order to keep the vowel quality constant. The outcome of the experiment is interpreted in the

light of the adaptation phenomenon of the auditory-nerve fibers: F₂ intensity is enhanced in relation to the intensities of the other formant peaks because the second formant is delivered to non-adapted channels, when all the other channels involved in the auditory representation of the vowel have already adapted to their stimulation levels.

Another experiment by Chistovich et al. (1982) indicates again that spectral components appearing later have their auditory representations favoured, when compared with those already producing adapted responses. More recently, an experiment by Summerfield et al. (1984), exploring an auditory after-effect, gives further support to the idea that adaptation of auditory-nerve fibers can play an important role in the selective enhancement of spectral cues.

The results of all these experiments show that there are psychoacoustic counterparts to the adaptation phenomenon of the auditory-nerve fibers. The important aspect they reveal is that the auditory system, as a consequence of the "automatic gain control" mechanisms of the nerve fibers, produces a stream of auditory patterns in which the relevant (new) information is enhanced.

Knowing that adaptation provides a means of enhancing new acoustic events, a new question may be asked, concerning the quality of the non-adapted and adapted auditory representations. Stimulus onsets must be optimal for the generation of "good" auditory representations since adaptation has not enough time to reduce the dynamic range of the frequency channels involved.

The hypothesis of a sharper auditory representation at stimulus onsets, as compared to that obtained when the frequency channels are already adapted to the stimulation level, was first studied by Lacerda and Lindblom (1983) in a frequency discrimination task, replicating Flanagan's (1955) frequency DL's experiment. Since Lacerda and Lindblom's (1983) experiments and results were recently reported in an extended version of their original experiments (Lacerda, 1986), the present paper will summarize the stimulus description and test procedures, and present further related experiments.

METHODS

The experiments to be discussed below were designed to study frequency discrimination as a function of stimulus onset characteristics. The stimuli used can be divided in two groups: One group of brief, constant F-pattern stimuli and another group of CV and VC synthetic stimuli, with time-varying F-patterns.

The brief, constant F-pattern stimuli were designed to test the hypothesis that the sharpness of an auditory representation is greater during the non-adapted auditory-nerve fibers' responses than after adaptation. A problem to be solved here is generating a series of stimuli to measure frequency discrimination under non-adapted responses and another series to measure discrimination under adapted responses, using stimuli of the same energy and duration in both series. To meet these design constraints, the onset characteristics of the stimuli were manipulated - a series of brief (50 ms), constant F-pattern, abrupt onset stimuli (eliciting non-adapted responses-

Delgutte, 1980; Delgutte and Kiang, 1984) and an equivalent series of gradual onset stimuli were created to assess frequency discrimination under non-adapted and adapted conditions.

To study the role that non-adapted responses at stimulus onset might play in the representation of speech stimuli, further discrimination experiments were designed using synthetic CV and VC syllables. The stimuli consisted of formant transitions between a series of "locus" patterns (differing only in their F_2 values) and a target vowel pattern (the same for all the stimuli). With this design, the relevant information concerning the discrimination within the CV or VC series was delivered at the first pulse of the CV syllables or at the last one, for the VC syllables. The hypothesis is that discrimination is better among CV loci than among VC loci because in the former the relevant information is delivered to nerve fibers at rest, while for the latter the nerve fibers are likely to already adapted by the previous stimulation (at least for moderate transition rates).

Subjects

All the subjects in the perception experiments were normal hearing adults, fluent speakers of both Swedish and English and native speakers of either Danish, English, Estonian, German, Portuguese, or Swedish. Most of the subjects were native Swedish or English speakers. The experience of the subjects ranged from very experienced phoneticians to absolutely naive. The author was not counted as a subject.

Stimuli

Brief, constant F-pattern stimuli:

Two series of 50 ms vowel-like stimuli were generated using a source function with an abrupt onset for one of the series and a gradual onset of the source for the other. The source excitation was obtained from five pulses delivered at $F_0=120$ Hz. F_0 was constant and the same for all the stimuli (Lacerda and Lindblom, 1983). The pulses were amplitude modulated by an exponential function that produced a 20 dB difference between the first and the fifth pulses. In a later version of this experiment, the periodic source was replaced by a pseudo-noise source (Lacerda, 1986). The same pseudo-noise source was used in all the stimuli. The duration and amplitude modulation of the noise source were the same as in the case of the harmonic source. The stimuli were generated by feeding the amplitude modulated source output through a four-formant filter function. F_1 , F_3 , and F_4 were the same for all the stimuli ($F_1=200$ Hz, $F_3=2600$ Hz and $F_4=3600$ Hz). F_2 was varied in 30 Hz steps between 1510 Hz, and 1690 Hz to generate the seven stimuli of each series.

Speech stimuli (/da/-/ad/ and /ba/-/ab/):

The vowel target was always the same ($F_1=750$ Hz, $F_2=1250$ Hz, $F_3=2400$ Hz, and $F_4=3050$ Hz), but the stimuli differed in their consonant loci: The /da/-/ad/ stimuli had loci at $F_1=200$ Hz, $F_3=2600$ Hz, $F_4=3600$ Hz, with F_2 varying from 1450 Hz up to 1750 Hz in steps of 50 Hz; the /ba/-/ab/ stimuli were organized in the same way. Their loci were defined by $F_1=200$ Hz, $F_3=2200$ Hz, $F_4=2700$ Hz while F_2 varied from 950 Hz to 1250 Hz, also in steps of 50 Hz. F_0 was fixed at 120 Hz for all the stimuli. The duration of the stimuli was always

50 ms. The formant transitions were exponential and had 90% of their total excursion during that time. The actual formant values at the consonant loci were the nominal ones. In the design of these CV and VC stimuli special care was taken to ensure that the corresponding CV and VC syllables were as symmetrical as possible (see Lacerda, 1986).

Procedure

The discrimination experiments were all carried out using a non-adaptive 2AFC paradigm without feedback. The stimuli were presented in (AX) and (XA) pairs where one of the elements of each series was the reference stimulus (the one with the lowest F_2 value). The pairs contained either two abrupt-onset or two gradual-onset stimuli, in the case of experiment 1, and either two CV or two VC stimuli, in the case of experiment 2, which could only differ in their F_2 values. The pairs were randomized and presented binaurally, via headphones. The presentation level was 74 dB SPL for a 1 kHz calibration tone that had a peak amplitude 1 dB above the peak level of the highest amplitude peak in the stimuli.

DISCRIMINATION AMONG BRIEF, CONSTANT F-PATTERN STIMULI

Fig. 1a shows the results of previous experiments (Lacerda and Lindblom 1983; Lacerda 1986) carried out with brief, constant F-pattern stimuli, with a harmonic source.

The discrimination among the 50 ms vowel-like stimuli with abrupt onsets was significantly better than among the corresponding stimuli with gradual onsets, for F_2 differences of 60 Hz and 90 Hz. A comparison between the frequency difference limens for these brief, constant F-pattern stimuli and the DLs obtained by Flanagan (1955) (open circles, fig. 1a) shows that Flanagan's results follow closely those now obtained from the gradual onset stimuli, indicating that both the gradual onset stimuli and the long steady-state vowel stimuli used by Flanagan adapt the auditory-nerve fibers.

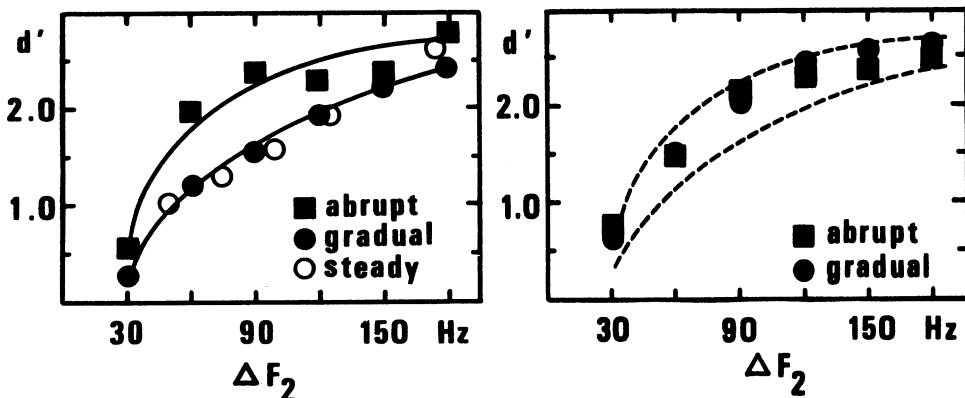


FIGURE 1, (a) Left: Discrimination scores for brief, constant F-pattern stimuli with abrupt and gradual onsets and for steady state vowels (harmonic source). (b) Right: Discrimination for brief, constant F-pattern stimuli with noise source.

When the pseudo-noise source was used, the difference between the discrimination scores for the abrupt and gradual onset stimuli became very small. The two sets of data clustered (fig. 1b), but a significant advantage in the abrupt onset discrimination was observed for ΔF_2 of 120 Hz and 180 Hz.

The conclusion from this set of experiments is that the initial portion of stimuli with abrupt onsets is the most favourable concerning the sharpness of the auditory representations.

HOW SHARP IS THE REPRESENTATION OF CV AND VC CONSONANT LOCI?

The discrimination among /da/ stimuli was compared to that among /ad/ stimuli. The performance dropped drastically when time-varying F-patterns were introduced. Furthermore, contrary to the expectation based both on the outcome of the brief constant F-pattern stimuli and on that of other discrimination experiments with CV and VC syllables (Sidwell and Summerfield, 1986), discriminability among the /da/ stimuli turned out to be worse than among the /ad/ ones (fig. 2a).

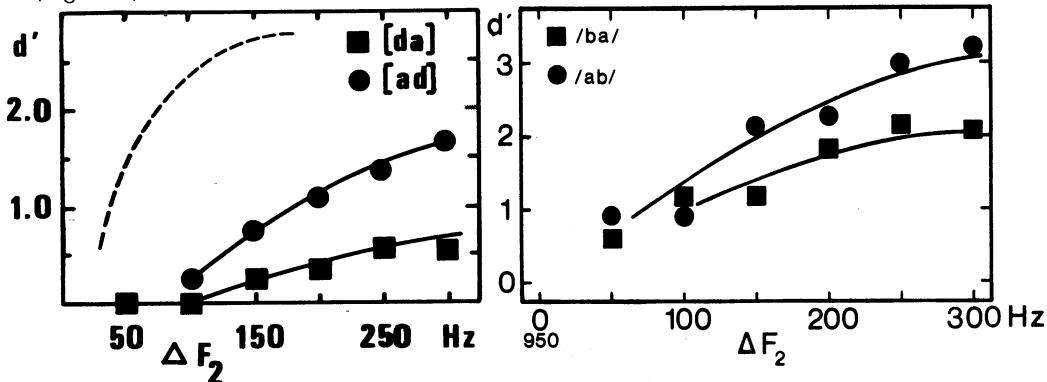


FIGURE 2, (a) Left: Discrimination among /da/ and /ad/ syllables. (b) Right: Same but for /ba/ and /ab/ syllables.

A possible explanation for the low scores observed with the /da/-/ad/ stimuli is that this discrimination task always involves stimuli drawn from the same phonetic category (Lacerda, 1986). This seems to be a reasonable explanation, but the outcome of the experiment with /ba/ and /ab/ stimuli raises some problems for this hypothesis (fig. 2b): /ba/-/ab/ scores are in the same order of magnitude as those obtained from the constant F-pattern stimuli. These high scores may be due to the particular skill of the four subjects who did the task, but there may be other reasons, as discussed below.

The disagreement between the predictions based on the brief, constant F-pattern stimuli and the results obtained from discrimination among CV and VC syllables raises some interesting questions. It may be argued that the rising F_1 transition may have the effect of turning the onset of the CV syllables into a gradual onset, but this seems to be ruled out by an informal check-up with flat F_1 transitions which

showed the same trend. Another explanation is based on the assumption that the initial portions of the stimuli are in fact sharply represented but that due to the formant transitions, sensory smearing blurs the information during the transition. This is consistent with the onset effect described by Smith and Brachman (1980) and with the idea that "the onset effect might last through the transition" (Miller and Sachs, 1983, p. 516) of the CV syllables.

If it is further assumed that discrimination of the final VC loci can be based on the post-stimulation responses (Harris and Dallos, 1979) at the auditory channels where the final loci are delivered, then it becomes possible to explain the fact that VC syllables are better discriminated than their mirror image CV syllables.

IS SENSORY SMEARING SUPPORTED EXPERIMENTALLY?

If discrimination among the CV syllables is degraded due to sensory smearing during the F_2 transitions, then it should be possible to improve the scores by decreasing the transition speed after the consonantal release. As far as the VC syllables are concerned, if the discrimination is based upon the residual excitation, then extending the final loci should not affect the scores since the relevant information was already present in the residual excitation.

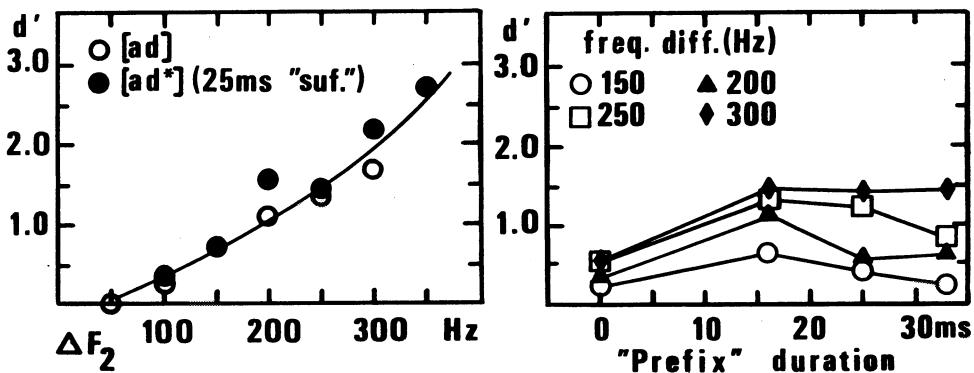


FIGURE 3, (a) Left: Discrimination among /ad/ syllables without and with extension of the final loci. (b) Right: Discrimination among /da/ as a function of the duration of the "stable loci".

Discrimination for the extended final loci

In the case of the extended final loci no significant improvement in the discrimination performance was observed. As fig. 3a illustrates, the scores, with or without extension, follow the same trend. This result indicates that the relevant information was already present in the residual masking pattern.

Discrimination for the prefixed initial loci

The initial consonant loci were preceded by two, three, or four excitation pulses (at the same F_0), anticipating the F-patterns at the onset of the transitions. The discrimination scores obtained from this experiment show interesting maxima for prefix durations of 16 ms

(fig. 3b). If it is assumed that the improved discrimination scores correspond to a situation in which sensory smearing is minimal, then the duration of the time window of the auditory system must be of this order of magnitude. This value is bigger than the time during which the auditory-nerve fibers exhibit a wide dynamic range, just after the onset of stimulation (Smith and Brachman, 1980), but it is of the order of magnitude of an integration time measured by Green (1973) (inferred from the detectability threshold of a click immersed in noise).

It may be concluded that the increased sharpness of the initial loci representations is degraded by blurring due to the integration time of the auditory system. The fact that natural CV utterances are better discriminated than VC may have to do with the broadening of the formant bandwidths occurring towards the natural vowel-consonant boundaries. This particular aspect and the disagreement with Sidwell and Summerfield's discrimination data are the object of current research.

REFERENCES

1. Chistovich, L. (1971). Auditory processing of speech stimuli-evidences from psychoacoustics and neurophysiology. Seventh Int. Cong. on Acoust., Budapest, 27-41.
2. Chistovich, L., Lublinskaya, V., Malannikova, T., Ogorodnikova, E., Stoljarova, E., and Zhukov, S. (1982). Temporal processing of peripheral auditory patterns of speech. In: R. Carlson and B. Granström (Eds.), The Representation of Speech in the Peripheral Auditory System, 165-180. Elsevier Biomedical Press.
3. Delgutte, B. (1980). Representation of speech-like sounds in the discharge patterns of the auditory-nerve fibers. Journal of the Acoustical Society of America, 68, 843-857.
4. Delgutte, B. and Kiang, N. (1984). Speech coding in the auditory nerve: IV - Sounds with consonant-like dynamic characteristics. Journal of the Acoustical Society of America, 75, 897-907.
5. Duifhuis, H. (1980). Level effects in psychophysical two-tone suppression. Journal of the Acoustical Society of America, 67, 914-927.
6. Flanagan, J. (1955). A difference limen for vowel formant frequency. Journal of the Acoustical Society of America, 27, 613-617.
7. Green, D. (1973). Minimum integration time. In: Aage Møller (Ed.), Basic Mechanisms in Hearing, 829-846. Academic Press, New York.
8. Harris, D. and Dallos, P. (1979). Forward masking of auditory-nerve fiber responses. Journal of Neurophysiology, 42, 1083-1107.
9. Houtgast, T. (1974). Lateral suppression in hearing. Academische Pers B.V., Amsterdam.
10. Kiang, N. (1968). A survey of recent developments in the study of auditory physiology. Ann. Otol., Rhinol., Laryngol., 77, 656-676.
11. Lacerda, F. (1986). Categories of speech sounds and the dynamics of the auditory system. Speech Input/Output; Techniques and Applications, IEE Conference Publication, 258, IEE, U.K., 88-93.

12. Lacerda, F. and Lindblom, B. (1983). How do stimulus onset characteristics influence formant frequency difference limens? In: A. Cohen and M.P.R. v.d. Broecke (Eds.), Abstracts of the Tenth International Congress of Phonetics Sciences IIA, p. 491. Foris Publications, Dordrecht, Holland.
13. Miller, M.I. and Sachs, M.B. (1983). Representation of stop-consonants in the discharge patterns of auditory-nerve fibers. Journal of the Acoustical Society of America, 74, 502-517.
14. Schalk, T. and Sachs, M. (1980). Nonlinearities in auditory-nerve fiber responses to band-limited noise. Journal of the Acoustical Society of America, 67, 903-913.
15. Sidwell, A. and Summerfield, Q. (1986). The auditory representation of symmetrical CVC syllables. Speech Communication, submitted.
16. Smith, R. and Brachman, M.L. (1980). Dynamic response of single auditory-nerve fibers: Some effects of intensity and time. In: G. van den Brink and F.A. Bilsen (Eds.), Psychophysical, Physiological and Behavioural Studies in Hearing, 312-219. Delft University Press, Delft, Holland.
17. Summerfield, Q., Haggard, M., Foster, J., and Gray, S. (1984). Perceiving vowels from uniform spectra: Phonetic exploration of an auditory aftereffect. Perception and Psychophysics, 35, 204-213.

EXTENDING THE SEARCH FOR A PSYCHOPHYSICAL BASIS FOR DYNAMIC PHONETIC PATTERNS

Anthony Bladon
Phonetics Laboratory, University of Oxford,
41 Wellington Square, Oxford, OX1 2JF, U.K.

1. CONTEXT AND ORIENTATION OF THIS PAPER

A prominent flavour of the research arguments summarized in several papers at this Workshop (Diehl, Howell and Rosen, Massaro, Pastore, and discussion sessions) has in my opinion been the following: (1) examine certain well-known phenomena of speech perception, (2) attempt to show that effects of an analogous kind can be demonstrated outside the human speech realm, then (3) argue that the phenomena in question owe their basis not to any 'specialness' of the speech code but to properties of auditory analysis more generally. This treatment, as the current volume reveals, has been meted out to such phenomena as categorical perception, fricative versus affricate perception, and voice onset time distinctions. In this way, a limited class of dynamic behaviours in speech, all of them with a long pedigree of experimental investigation (albeit predominantly in English), have been shown to be compatible with auditory-perceptual capabilities.

The philosophy of this paper is quite different. I take it as axiomatic that, if there is anything in the psychophysics of the transmission of speech which underlies speech behaviour patterns, anything which shapes and predisposes them in specific ways, then it should show up in phonetic universals. I therefore propose in this paper to identify some universal trends in the dynamic structure of sound patterns in language, selecting those which appear to offer promise for a methodology of psychophysical explanation. These areas of promise are largely new areas, in the sense that speech perception experiments have rarely been done on them - and psychophysical experiments virtually not at all.

To take this argument further, mention should briefly be made of some of the relevant characteristics of the human auditory response to dynamically changing events. For instance, consider onsets and offsets of spectral energy. If the human auditory system tracked onsets and offsets with equal efficiency, the two tasks would be equivalent. We might then expect to find psychophysical data illustrating the similarity of response. We might expect onsets and offsets to show similar detection thresholds. However, as reviewed and further demonstrated by Tyler et al. (1982), this expectation would not be well supported. Among the evidence is that the DL of a gap in a signal (typically 50 to 75 ms gap duration) is some three times as long as the DL of a signal itself. Filled intervals are some three times easier to discriminate than unfilled intervals; they are "processed differently". More generally, it seems, spectral changes which yield

predominantly an onset of neural firing are much more perceptually salient than those producing an offset.

This asymmetry of the auditory response to onsets versus offsets seems also to be consistent with the phenomena of forward masking (a stronger signal partially masking a weaker successor), of short-term adaptation and of recovery in the auditory nerve. By neural recovery, we understand the facility whereby silent intervals promote a rapid, high-amplitude auditory discharge, at the instant when they are interrupted by upcoming energy. Adaptation however shows itself as a decay in response to a moderate level of discharge, when the stimulus itself continues at a constant level (Delgutte, 1982, and references summarized therein).

It is of interest to phoneticians to examine properties of auditory processing, such as adaptation, recovery and the asymmetry of response to onsets versus offsets, for their implications on linguistic systems. There turn out to be several areas where explanatory inferences can be made, where the psychophysics of the transmission channel could well be implicated in observed trends in speech behaviour.

2. NASALIZATION OF VOWELS

As a first example: it is an observed cross-language fact that vowels nasalize much more readily before a nasal consonant than they do after it. Specifically, the sequence vowel+nasal is "overwhelmingly" more likely to suffer weakening of the nasal consonant or nasalization of the vowel, than is the reverse sequence nasal+vowel, which is weakened only occasionally (Ruhlen, 1968). Phoneticians would like to know why. Numerous articulatory explanations have been sought (e.g. in differential velocities of velum opening and closing), but none seems altogether persuasive (Al-Bamerni, 1983). Is there then a psychophysically-based reason? I suggest that there is, as follows.

It is probably true in running speech that neither the nasal+vowel boundary nor the vowel+nasal boundary is especially strongly signalled to the listener, given that the gross similarities in the spectral content of the two sounds induce some adaptation in the auditory nerve. This indistinctness of the successive spectral representations can be inferred from psychoacoustic manipulations on nasal and vowel sequences, as reported by Chistovich et al. (1982). Nevertheless, other evidence would lead us to expect that the auditory response to the nasal+vowel boundary would be different from that to vowel+nasal. One indication is that forward masking should assert itself more powerfully when the stronger signal precedes, than when it follows. In fact there is the general asymmetry such that the spectral offset of energy is demoted in perceptual weight in comparison to an onset. What are the implications of these statements for nasal+vowel and vowel+nasal sequences? Since the sequence vowel+nasal consists largely of spectral energy offsets rather than onsets, it ought predictably to be more vulnerable than the reverse sequence would, to temporal auditory smear. Vowel+nasal ought to be more susceptible to forward masking. Therefore, vowels should more readily nasalize after nasals than before them. This, of course, is precisely the cross-language trend we began by noting.

3. VOCALIZATION OF LATERALS

A somewhat similar piece of evidence is provided by the sequence vowel+lateral. Acoustically, the tendency is for a lateral's spectrum to adopt something of the quality of the adjacent vowel, while being somewhat weaker in energy level (Bladon and Al-Bamerni, 1976). Now it can be noted that the transition onto a postvocalic lateral is coded auditorily as mostly neural offsets (whereas a lateral+vowel sequence consists mainly of the more salient onsets). Thus one might predict in vowel+lateral sequences some overt evidence of an extra tendency towards assimilation. Indeed this is just what is observed. I suggest that this could well underlie why laterals vocalize so much more readily after a vowel (e.g. in Cockney, Old French, Dutch, Portuguese etc.) than they do before a vowel. To the perceiver, the lateral quality is less distinct in the postvocalic position.

It may not be too far-fetched to draw some support for the detail of this interpretation of laterals (and of nasalization, above) from the experiments reported in this volume by Summerfield and Assmann. They described an effect of the perceptual enhancement of temporal contrast, whereby a flat spectrum, when preceded by a precursor spectrum with valleys in it, was perceived to have not a flat quality but the quality of a sound whose peaks would be at the precursor valley frequencies. From this striking demonstration, the point most relevant to the present discussion is that this percept did not result so readily when stimuli were in the opposite order. The flat spectrum had to follow the valley spectrum. The auditory system seemed more sensitive to the spectral amplitude changes arising from a valley being subsequently filled, than to a plateau being subsequently cratered. If this is so, then there is a clear analogy with the lateral or nasality data. The dynamics of /IV-/ or /nV-/ can both be said to involve, from a physical point of view, some 'spectral valley filling': they should therefore be represented auditorily with some enhancement of their amplitude change. Their counterparts /-VI/ or /-Vn/ will, however, not be able to benefit from this enhancement. Consequently, we can account for something of the latter's greater propensity to smear (vocalize, nasalize etc.).

4. PREASPIRATION

Preaspiration of a stop is said to occur when a preceding vowel gives way to an aspirated phase prior to the stop closure. It would be hard to imagine a speech pattern less favourably designed for long-term perceptual survival. The aspiration spectrum has essentially the same shape as that of the preceding vowel: there is therefore widespread short-term adaptation following that vowel. There is no apparent prospect of neural recovery during the preaspiration. Such temporal information as is imparted by preaspiration must in any case depend wholly on the detection of offsets - which does not make for a robust outlook. And so, given that preaspiration suffers from an accumulation of auditory handicaps, it would not be a risky prediction that languages would rarely make use of this auditory-phonetic dinosaur. Indeed, from the 317 languages of the UCLA database (Maddieson, 1981), preaspiration is attested in only 2 or 3.

5. THE GLOTTAL FRICATIVE

The same arguments apply also to the distributional asymmetry of /h/ in languages. It has often been said that syllable-final and word-final occurrences of [h] are much rarer than initial [h] (Holmberg and Gibson, 1979). A related fact is the well attested report (Lindblom, 1978) from languages having /h/, that a word ending in a vowel (with no glottal stop after it), when played to listeners in reverse, is heard as having an initial [h]. Now there probably is a physiologically-based tendency for the vocal folds to open, at vowel end, more gradually than they approximate for a vowel beginning; this would contribute to the impression (in the reversal condition) of an initial [h]. However, this does not explain why we do not hear the vowel offset as a final [h]. But there is a perfectly good psychophysical reason for these /h/ behaviours. Like preaspiration in the previous paragraph, /h/ after a vowel is highly nonsalient to the hearer. With its vowel-like spectrum it suffers heavily from short-term adaptation by the preceding vowel, and it depends on offset detection, which is inefficient. In short, temporal auditory smear prevents a final [h] from being easily heard. Hence, presumably, its rarity.

6. FORMATION OF SYLLABIC CONSONANTS

I have been suggesting that where psychophysically the transitions between somewhat similar sounds are temporally indistinct, language can be expected to show an assimilatory effect. If the syllable in question is unstressed, one might predict that the effect will go further. In these respects, conditions are just right for the generation of syllabic consonants. A vowel which 'overnasalizes' could readily be heard as a (syllabic) nasal; a vowel+lateral sequence could readily be heard as a 'lateral-like vowel', from which the step to a perceived syllabic consonant is indeed short. One would expect these syllabic consonants to form most readily after vowels (because of adaptation), and with nasals and laterals rather than other consonants because their auditory spectra are more similar to a vowel's. Now, from Bell's study (1978) of 85 languages, it is noticeable how well this scenario is borne out. Bell shows (a) that syllabic consonants tend to occur in unstressed positions; (b) that much the commonest syllabic consonants are nasals, followed by laterals; and (c) that these syllabic consonants inevitably result from vowel syncope. It seems fair to conclude that dynamic auditory phonetics can predict this behaviour rather well. Notice in particular the asymmetry: syllabic consonants arise very rarely from the syncope of a following vowel e.g. in such a sequence as /nVC-/. This can of course be attributed to the relatively greater auditory strength of the spectral onsets in the /nV-/, than in the more susceptible /-Vn/.

7. DISSIMILATION

Of the observed cross-language trends in consonant cluster types, a few (though emphatically not all) can be seen to have a psychophysical grounding. The dynamic phenomena which seem most consistently to be attributable to this kind of influence are those of epenthesis and of dissimilation.

Greenberg (1978), from a 104-language sample, showed a strong tendency for sibilants not to combine in sequence. A familiar

illustration of this is provided in English which inserts an epenthetic vowel to separate the sibilants in /wɪʃɪz, wɪtʃɪz/ but not otherwise /wɪts, wɪfs/. Now the articulatory difficulty argument does not seem very attractive here, since English happily tolerates /θs/, a sequence with a nonsibilant but surely as difficult to produce as /ʃs/. Rather, we can draw on our knowledge of the auditory-spectral properties of sibilants, and suggest that the succession of sibilant sounds is disfavoured psychophysically, because of excessive auditory adaptation. Sibilants have a similar spectral shape to each other, much more so than to a (low-intensity) front fricative. In this way, the hearer's identification of a second sibilant in a sequence of two is prejudiced.

A second escape route which languages use, apart from epenthesis, to avoid a succession of sibilants, is dissimilation. For example, German sechs [zeks] 'six' shows a dissimilation of sibilant [ç] to [k] before [s]. Compare however sechzehn [zɛçtse:n] 'sixteen' which keeps [ç] before [ts]. This example fits rather interestingly with psychophysical expectations. While the adjacent sibilants dissimilate, the nonadjacent sibilants [ç-s] in sechzehn do not: they are robust enough to the perceiver, because the intervening plosive permits the necessary neural recovery.

Ohala (1981) has advanced the interesting view that dissimilations are all to be laid at the hearer's door. More particularly they are, he claims, the product of the listener's (mistaken) compensations. Ohala's presentation is worth closer study as a source of evidence for the psychophysical basis of phonetics. However we shall not further elaborate Ohala's reasoning here; rather, it seems more important to note a reservation about it. The dissimilating sibilants cited above are suggestive of an altogether less exotic line of thinking (which, nevertheless, still implicates the auditory channel). Dissimilations such as my illustration, which concern sounds immediately adjacent and hence within the range of temporal smear effects, may arise not (as Ohala assumes) out of some active compensation process hypothetically attributed to listeners, but, much more probably, as a way simply to avoid an auditory insufficiency. We need suppose no more than that hearers (passively) need to maintain some temporal auditory distinctiveness. Dissimilation can be an expedient to that end.

8. CONCLUSION

This discussion of dissimilation combines with the discussions of the preceding sections to form a consistent picture. They all give evidence of an underlying influence of the perceptual system on the long-term syntagmatic properties of phonetic systems. Auditory adaptation, masking and failure to achieve neural recovery can lead to temporal smear, which hearers resolve in a variety of phonological ways. In addition, the temporal asymmetry, by which adaptation follows (but does not precede) a neural event, and by which energy onsets are much more salient than offsets, turns out to be relatable, in the sound patterns of languages, to a variety of structural asymmetries.

Of course, one might object (as Repp did, in discussion) that there are many other properties of speech behaviour which have no obvious correlation with the psychophysics of their transmission. True; but articulatory, social, and other factors are not being denied their

role. This paper aims to do no more than identify some favourable points of contact and of explanation between phonetics and psychophysics. A more important and urgent shortcoming of my search for such correlations is that the ones currently proposed are mainly justified by extrapolation or by inference from studies whose purpose is incidental to them. It would be nice to begin to redress that shortcoming by embarking on the specific experimental tasks which are implied. However, my sense from this Workshop is that we are hardly yet equipped to do so. The study of peripheral auditory dynamics may need more honing first. If so, then the extension of the search for a psychophysical basis of sound patterns will remain rather programmatic for some time to come. Meanwhile, speech perception methodology can usefully gain from a more general awareness of broad, cross-linguistic horizons.

REFERENCES

1. Al-Bamerni, A. (1983). Oral, velic and laryngeal coarticulation across languages. D. Phil. thesis, University of Oxford.
2. Bell, A. (1978). Syllabic consonants. In Greenberg, J.H. (ed.), Universals of Human Language, Vol. 2: Phonology (Stanford, Stanford Univ. Press.), 153-201.
3. Bladon, R.A.W. and Al-Bamerni, A. (1976). Coarticulation resistance in English /l/. Journal of Phonetics, 4, 137-150.
4. Chistovich, L.A., Lublinskaya, V.V., Malinnikova T.G., Ogorodnikova, E.A., Stoljarova, E.I., and Zhukov, S.J. (1982). Temporal processing of peripheral auditory patterns of speech. In Carlson, R. and Granström, B. (eds.), The Representation of Speech in the Peripheral Auditory System (Amsterdam, Elsevier Biomedical), 165-180.
5. Delgutte, B. (1982). Some correlates of phonetic distinctions at the level of the auditory nerve. In Carlson, R., and Granström, B. (eds.), The Representation of Speech in the Peripheral Auditory System (Amsterdam, Elsevier Biomedical), 131-150.
6. Greenberg, J. (1978). Some generalizations concerning initial and final consonant clusters. In Greenberg, J.H. (ed.), Universals of Human Language, Vol. 2: Phonology (Stanford, Stanford University Press), 243-279.
7. Holmberg, E. and Gibson, A. (1979). On the distribution of [h] in the languages of the world. Phonet. Exper. Res. Inst. Ling. Univ. Stockholm, 1, 68-82.
8. Lindblom, B. (1978). Phonetic aspects of linguistic explanation. Studia Linguistica, 32, 137-153.
9. Maddieson, I. (1981). UPSID: The UCLA phonological segment inventory database: Data and Index. UCLA Working Papers in Phonetics, 53.
10. Ohala, J. (1981). The listener as a source of sound change. In Masek C.S., Hendrick, R.A., and Miller, M.F. (eds.), Papers from the Parasession on Language and Behavior (Chicago, Chicago Linguistic Society), 178-203.
11. Tyler, R.S., Summerfield, Q., Wood, E.J., and Fernandes, M.A. (1982). Psychoacoustic and phonetic temporal processing in normal and hearing-impaired listeners. Journal of the Acoustical Society of America, 72, 740-752.

GENERAL DISCUSSION OF SESSION 3: DYNAMIC ASPECTS

Chairman: D.B. Pisoni

Session 3 of the NATO Workshop on the Psychophysics of Speech Perception focused on the dynamic aspects of speech perception. By dynamic aspects, we mean an interest in issues related to the "time-varying" properties of complex speech and nonspeech signals. The papers and discussions covered a very wide range of topics and problems in the field, ranging from studies of the perceptual cues to fricatives to temporal order identification to masking of burst spectra in consonants. Despite the rather wide diversity of interests and topics covered, a number of general issues emerged from the papers and discussions. In the sections below, we attempt to summarize these briefly and, hopefully, by doing so, establish the potential significance of these findings for future research in speech perception.

Although there are rather substantial differences in the theoretical goals and experimental methodologies of speech scientists and hearing scientists (psychoacousticians), there is now very strong evidence for the beginnings of some fruitful interactions and potential exchanges of ideas, methods, and research findings between the two disciplines. For the most part, speech scientists have been primarily interested in research questions surrounding the identification, recognition, and categorization of speech signals, and the acoustic correlates of phonetic features. Psychoacousticians, on the other hand, have traditionally been interested in the limits of sensory resolution of discriminating nonspeech signals and in developing models to characterize the sensory processing of these signals by the auditory system.

In this session of the workshop, we saw evidence for both approaches as well as a rapprochement between the two research philosophies. In recent years, more and more psychophysical studies have been carried out using complex speech and speech-like signals as compared to the earlier emphasis and concern with relatively simple nonspeech signals like tones and noise bursts. It was also apparent from several of the presentations in this session that a large number of recent studies have been carried out using traditional techniques and methodologies from speech perception (i.e., identification and discrimination) with complex nonspeech signals. Although there have been some changes in research philosophy and methodology, the two approaches still remain quite distinct in their overall goals and orientation. This is not surprising because the research questions and philosophies deal with quite different levels of description.

For example, numerous studies have been carried out by psychoacousticians to investigate the sensory encoding of complex signals like speech. Although speech signals are used as stimuli, the major focus in this research is on discrimination rather than identification or categorization. The researchers are simply using speech as "complex" acoustic stimuli and they display little, if any, interest or concern with the significance of the signals to the listener as linguistic entities or events in spoken language. This appears to be a common concern of those hearing scientists working with speech signals, and this trend seemed to characterize a number of the studies reported in this session as well.

The goals pursued by speech scientists in using complex nonspeech signals that have properties in common with speech are quite different in a number of respects. In these studies which are concerned primarily with identification and categorization, investigators have tried to describe the important acoustic properties in the speech waveform that could be used to account for the labeling performance and the location of the boundaries between perceptual categories. Indeed, since the publication of the paper by J.D. Miller et al. (1976) demonstrating categorical perception of complex nonspeech signals differing in noise/buzz onset time, and the suggestions offered by K.N. Stevens on the underlying perceptual basis of phonetic categories in terms of processing constraints by the auditory system, numerous nonspeech "control" studies have been carried out. These recent nonspeech studies have raised important questions about the previous accounts of a number of phenomena in speech perception, although there is currently no universal agreement about the significance of the nonspeech results for explaining all the relevant phenomena in speech perception in a principled way (by recourse to psychophysical properties and/or mechanisms such as masking, sensory integration or saturation).

There appeared to be fairly good agreement among members of the workshop that blanket assertions concerning "specialness" of speech perception were no longer adequate to account for the differences and similarities observed in perception between speech and nonspeech signals. Some of the recent nonspeech studies figured prominently in discussions at this session. Unfortunately, there was still skepticism about whether psychophysical accounts of the identification and discrimination of complex nonspeech signals would be sufficient to account for the results obtained in speech perception studies on trading relations, context effects, and speech rate. No general principles or consensus appeared to emerge in discussions about these issues, although there were extensive interactions among members of the workshop.

In addition to these general observations and conclusions, a number of more specific research topics were discussed by members of the workshop. Two studies dealing with the acoustic cues for the affricate-fricative distinction were presented, and a proposal framed in terms of masking was offered as an explanation of the psychophysical basis for this contrast.

A detailed summary of data and findings on the voicing distinction was presented for both speech and nonspeech signals, and suggestions for future research were outlined. A number of years ago

a suggestion was made that the VOT contrast between stops in initial could be accounted for by reference to constraints on temporal order identification as described by Hirsh in his classic article. In this session, an extensive discussion of Hirsh's data and its relevance to voicing perception took place. It was suggested that Hirsh's temporal order data have been misinterpreted by researchers working on the psychophysical basis of voicing cues. The discussion centered on a reexamination of Hirsh's old data and assessment of the claims about VOT and TOT stimuli in labeling and discrimination tasks.

The criticisms of the misinterpretation of Hirsh's data were based on both the specific task employed and the nature of the psychometric functions. The task was essentially an identification task rather than a discrimination task, and was therefore highly susceptible to task demands and expectations. The resulting psychometric function was symmetric around onset synchrony (zero ms TOT), and indicated a gradual change in perception for increasing asynchrony. The 17 ms threshold thus indicates a statistically defined relative threshold rather than a natural discontinuity which might serve as a strong basis for higher order speech contrasts. The essence of Hirsh's conclusions has now been replicated in a number of laboratories using both identification and discrimination tasks. With simple stationary stimuli, the TOT threshold, labeling boundary, and discrimination peak consistently seem to be at locations which are too brief to provide a reasonable primary basis for voicing contrasts in English stop consonants. Recent research has tended to focus on TOT for nonstationary analogs to speech stimuli.

In addition to studies on the presumed psychophysical basis of voicing perception in initial position, there was also presentation and discussion of nonspeech data on voicing perception in medial position in words like "rapid--rabid". An issue was raised in the discussion concerning the theoretical significance of these nonspeech control studies and the relevance of these results for understanding the processing of complex nonspeech signals by the auditory system. Some reluctance was expressed about future research using nonspeech control stimuli and the usefulness of comparisons between speech and nonspeech signals. If a parallel is shown between speech and nonspeech perception (e.g., an interaction among cues), the nonspeech interaction is in need of explanation in terms of auditory or psychophysical principles. Alternatively, it may be that nonspeech stimuli that are sufficiently speechlike are processed by central processes involved in speech perception, even when the listener is not aware of the speech-likeness of the stimuli. That is, the extent to which speech mechanisms are engaged may be a function of stimulus properties, not (or not entirely) of whether a conscious speech percept is reached by the listener.

Other psychophysical data were presented on the identification of brief tone sweeps that resemble formant transitions in consonant-vowel syllables. As with the previous nonspeech studies, some questions were raised about the relevance of these data to speech perception. Data were also presented on the spectral representation of intervocalic consonants and the possible role of the temporal masking in the perception of stop consonants. Other studies examined vowel perception and frequency discrimination for complex spectra such as those found in natural speech.

Concern was raised over the tendency to study the perception of only a single token of a given stimulus contrast. Normal perception of speech involves stimuli which differ along a number of relevant and irrelevant dimensions, and thus represents a high-uncertainty situation. These experimental conditions often encourage subjects to focus on subtle characteristics of stimuli and thus possibly to perform the required task on the basis of cues which may be relatively unimportant or even irrelevant for speech perception.

The identification of parallel phenomena for speech and nonspeech control signals does not represent an explanation for speech perception, but only the potential for a common explanation for both phenomena. It was suggested that the strategy for the psychoacoustic investigation of nonspeech analogs should be to define the nature of the perception of the parallel speech and nonspeech phenomena, with the ultimate goal of developing testable models for the perceptual phenomena. This research strategy ultimately will define the nature of the front-end of the complex system which results in the perception of speech. If the resulting models accurately describe and predict both the nonspeech and speech phenomena, then a principled explanation of the phenomena will have been provided that can serve to account for both speech and nonspeech signals.

Finally, some speculations were offered about the psychophysical basis of phonetic and phonological patterns in different languages and the potential importance of studies examining phonetic universals in language other than English. Much of the research on speech reception has been done with English, and it is apparent that much more data is needed from other languages as well in order to support claims about any biologically or genetically based perceptual process.

In summary, the papers presented in this session and the ensuing discussions emphasized the importance of studies designed to investigate the underlying sensory and psychophysical basis for phonetic features and phonetic categories in language. This work is still very new and the exact implications of these findings, particularly the results of the nonspeech control experiments, remain unclear at the present time. However, there was fairly good agreement that this research was important and interesting and would provide alternative explanatory accounts to the now traditional and dominant view that phenomena observed in speech perception result from the operation of specialized perceptual mechanisms that entail some form of articulatory mediation. Not everyone at the workshop was convinced that it was worth continuing these nonspeech studies in the future. But most participants agreed that work on the sensory and/or psychophysical basis of speech perception was a worthwhile goal for the future. The implications for understanding both normal and pathological speech processing were apparent throughout this session.

Chapter 4

TIMBRE

**(PERIPHERAL CONSTRAINTS AND CENTRAL PROCESSES IN
THE PERCEPTION OF COMPLEX SIGNALS).**

PSYCHOPHYSICS OF AUDIO SIGNAL PROCESSING AND THE ROLE OF PITCH IN SPEECH*

Ernst Terhardt,
Institute for Electroacoustics, Technical University, München,
F.R. Germany

1. INTRODUCTION

This paper is going to describe a broad and general approach to a problem which at first sight may appear relatively special, namely, the psychoacoustic representation of the first two vowel formants. The problem emerges from two facts, the first of which is pertinent to the speech source, the second to the auditory receiver: (1) the relatively high oscillation frequency of the vocal cords ("voice fundamental frequency") which renders the physical representation of vocal-tract resonances quite scanty; (2) the auditory system's relatively high frequency selectivity, which in the lower frequency region analyses the signal into individual harmonics rather than formants. An illustration is given in Fig. 1, showing schematically the boundary separating the area of auditory analysis into spectral pitches of individual harmonics from that of melted perception (It should be noted that actually there is not a sharp boundary but a transient zone that depends on the vowel's spectral envelope and SPL; cf. Plomp & Mimpens, 1968; Terhardt, 1979a,b; Stoll, 1982; Houtgast, 1974). It is apparent that in practically the whole oscillation-frequency range typical of natural speech resolution into individual harmonics takes place in the existence region of the first two formants. On the right side of the diagram is indicated the weight which is assigned by the central auditory system to various frequency regions. That weighting function was obtained from two different types of data: (1) Nonsense-syllable recognition (articulation index: cf. Fletcher, 1953), and (2) virtual-pitch perception ("spectral dominance", cf. Plomp, 1967; Ritsma, 1967; Terhardt et al., 1982). It is apparent that it is essentially the existence region of the first two formants which is most important to the central auditory system.

While it is evident that the most important vowel formants are aurally resolved into individual harmonics, the frequency spacing of the latter (equal to fundamental frequency) turns out to have little influence on vowel recognition. In our recent experiments (see appendix), we did not find any significant effect in the fundamental-frequency range of about 80 to 200 Hz. Apparently it is only at considerably higher fundamental frequencies that an effect can be measured (cf. Ryalls & Lieberman, 1982; Gottfried, 1986). On the other hand, we found a pronounced effect of inharmonicity (see appendix); a

*I am indebted to W. Heinbach for providing Fig. 3, and to H. Ortner and G. Stoll for carrying out the experiments described. This research was carried out in the Sonderforschungsbereich 204, "Gehör", München, supported by the Deutsche Forschungsgemeinschaft.

finding which is consistent with results of Elman (1981) on vowel reproduction with inharmonic feedback.

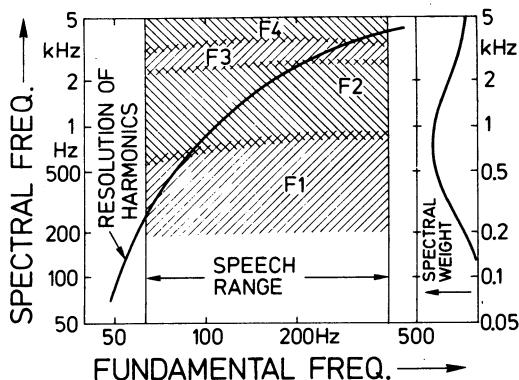


FIGURE 1, "Spectral frequency vs. fundamental frequency" plane, showing schematically the formant existence regions and the area in which harmonics are aurally resolved into individual spectral lines (below the solid line). Right: "Importance of frequency regions" in articulation index and virtual pitch perception.

These and other observations suggest that aural detection of lower formants may be a process of higher sophistication than mere maximum- or center-of-gravity-detection of a smeared spectral envelope. The seeming paradox may have conceptual, mathematical, and psychophysical implications, which for the understanding of speech perception are of high significance. It is one of the present paper's goals to point out some of those implications and ways of resolving them. First, basic conceptual aspects will be considered. Then it is pointed out that aurally adequate representation of the auditory stimulus, playing a crucial role in psychophysics, is as yet unavailable, and a solution is offered. Finally, the role of spectral and virtual pitch in speech communication is considered. In the appendix, recent experiments on perception of isolated vowels with harmonic, inharmonic, and random-noise spectra, relevant to the present topic, are described.

2. SIGNALS, SENSATIONS, AND SYMBOLS: THE THREE "WORLDS"

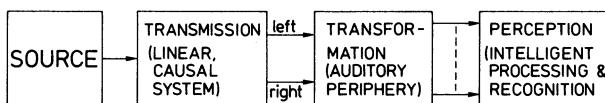


FIGURE 2, The essential parts of the communication chain, with regard to speech perception.

Consider the essential parts of the speech perception chain shown in Fig. 2. Its most complex and unexplored part is the ultimate perception process (box on the right). The input to that process is provided by certain basic auditory attributes (e.g., pitch, loudness, roughness, sharpness) which in turn are produced by stimulus-to-

sensation transformations on a peripheral level (box termed "transformation"). The latter part has two input channels corresponding to the two ears. The respective auditory stimuli are sound-pressure oscillations which are dependent on the characteristics of both the source signal (produced by a talker) and the transmission path (which in the "natural" case is a more or less reverberant sound field). In many cases there are additional influences corrupting the source signal, e.g., other voices, and noise.

Exploration of the speech communication chain is attempted with basically different approaches (physical, neurophysiological, psychological, etc.). Those approaches follow certain general concepts which however are seldom explicitly reflected and mentioned (cf. Repp, 1981). Therefore, let us begin with an attempt to clarify the respective roles of the different approaches.

In psychophysics, the dichotomy of stimulus and sensation provides a useful and important concept. The existence of two basically different "worlds" (namely, that of the physical signal versus that of subjective sensations) is widely accepted. The psychophysical approach to speech perception makes particularly apparent that Popper (1972) and Eccles (1973) are right in saying that there is a third "world", i.e. the world of information, or symbols. The distinction between the speech signal (i.e., the physical aspect) on the one hand, and "what it means" on the other, is conceptually identical with the dichotomy of speech versus language, which is familiar to everybody concerned with the topic. Following Popper & Eccles's concept, we can thus make the following assignments:

- World 1 = The world of matter and energy, including the speech signal.
- World 2 = The world of subjective experiences, sensations, and emotions, including those pertinent to speech.
- World 3 = The world of information, including symbolic manifestations of speech, i.e., language.

It should be noted that the "three-world-concept" essentially is merely a reformulation of basic and naive human experiences: it is not a theory in the sense that it would make decisions on questions which should remain open for research. On the contrary, as an approach to interdisciplinary problems such as that of speech communication, it is the most unbiased concept I know of.

Primarily, the three-worlds concept makes apparent that it is impossible to measure objects and processes in one world using methods of another. By physico-chemical methods alone, as used in physical acoustics and neurophysiology, one will never find anything beyond physics and chemistry, i.e. data and objects of world 1. However, by incorporating the information aspect, neurophysiological research can elucidate relationships between worlds 1 and 3. Sensations can be measured only by experiments with (human) subjects using psychophysical methods. Therefore, exploration of relationships between objects and/or processes in worlds 1 and 2 is the domain of sensory psychology and psychoacoustics. By making the respective roles, merits, and limitations of different approaches explicit, the efficiency of their mutual complementation may be enhanced.

The crucial criterion of making a distinction between worlds 2 and 3 can be seen in categorization. Categorization is a decision process in which to a certain group of psychophysical entities, which can be both physically and sensorially different from each other, is assigned one and the same categorical label. The speech vowels provide a typical example; another example is tone categorization in music. Gestalt perception may be regarded as an "implicit categorization", i.e., categorization without explicit labeling. Gestalt perception thus can be seen as a natural abstraction and categorization process, establishing a relationship between objects in worlds 2 and 3. In that view, any categorization process implies an abstraction; categorization and abstraction are more or less synonymous. The most elementary case of abstraction may be seen in the formation of visual contours. In the present view, formation of contours is an elementary case of abstraction and as such a basic element of "intelligent processing". Since formation of spectral pitches in audition is seen as equivalent to visual-contour formation (cf. section 4), the same can be said about spectral pitch: formation of spectral pitches in the auditory periphery may be regarded as a most elementary abstraction process.

In that view, auditory Gestalt perception ("intelligent processing and recognition" in Fig. 2) is conceptualized as hierarchically organized processing of categorized sensory variables, such that a Gestalt that in one hierarchical layer is composed of certain elements can in the next- higher layer provide an element to another Gestalt of higher order, and so on. That process, though typically psychological, is concerned with objects of world 3. In that context, I recommend taking into account the ideas of Hofstadter (1979).

With regard to auditory Gestalt perception, it is interesting to note that time plays a twofold role. First, time is that conceptual variable that measures changes of spatial patterns. Second, it provides another Gestalt dimension in the sense that subsequent events (or objects, or elements) are synthesized within a certain time interval into an entity, i.e., Gestalt. Rhythmic patterns provide a most familiar example of that phenomenon. In the view outlined, formation of a rhythmic pattern requires previous formation of temporal elements, or events, which is equivalent to temporal categorization. In speech, formation of syllables is an example of temporal categorization. It was experimentally verified that the perceived rhythm of natural speech is entirely congruent to syllable structure (Köhlman, 1985).

As far as the strategies of auditory Gestalt perception are concerned, much can be learned from visual perception. Conceptually it appears plausible to make three basic assumptions:

1. The Gestalt processor is flexible and "intelligent", adapting its strategies to any perceptual situation and problem. In that view, it appears not quite plausible to conceptualize two perceptual modes, i.e. for speech and non-speech. Rather, there may exist many more "modes", which however reflect nothing but the processor's flexibility.
2. The Gestalt processor is primarily source-of-information-directed, where ordinarily sources of information are objects of the external world.
3. The Gestalt processor is entirely dependent on the input provided by the peripheral system. Therefore, adequate stimulus representation

and psychophysical transformation are of crucial importance. By exploring what the processor basically can do, given the limitations of its input, one may - taking account of (1) and (2) - find out step by step how it in fact works.

3. ADEQUATE REPRESENTATION AND EVALUATION OF THE AUDITORY STIMULUS

The auditory stimuli (one to each ear) play a key role, the importance of which can hardly be overestimated. While at first sight description and physical-mathematical evaluation of the stimuli may appear as more or less trivial (i.e., state-of-the-art), closer conceptual inspection of the problem reveals that this part of the whole process leaves much to be clarified.

To what extent, and in what sense, are the harmonics of a voiced speech signal "real"? We can perceive several of them with distinct individual pitches (spectral pitches); but in what sense do they objectively exist? The classical answer to that question (which to date is still accepted and used too readily) was given by G.S. Ohm (1843), saying that it is the Fourier series of periodic signals which assigns reality to the harmonics. However, this is not an explanation which is compatible with linear-systems theory, since the Fourier-series representation assumes an unrealistic property of the signal, namely, infinite duration. Plain everyday experience demonstrates that the auditory system carries out a type of Fourier analysis which is a continuous variable of both frequency and time. Here a conceptual deficit becomes apparent, which up to now often has been ignored.

A mathematical signal transformation which meets the aspects and requirements mentioned is provided by the formula

$$P(f, t) = \int_0^t p(x) e^{-a(t-x)} e^{-j2\pi fx} dx , \quad (1)$$

where $p(x)$ is the audio signal, and $P(f, t)$ the resulting complex time-variant spectrum. The signal is multiplied with a weighting function which decays exponentially with the time constant $1/a$ toward past times relative to the actual moment of analysis, t . That type of time-variant Fourier-transformation was earlier proposed by Flanagan (1972) and can be traced back to Gambardella (1971), Schroeder & Atal (1962), and Fano (1950) (see also Searle, 1982). From the aspect of aurally adequate signal representation, the transformation with the particular weighting function included in Eq. (1) was recently reconsidered and conceptually justified, and its relevant characteristics were outlined (Terhardt, 1985a,b). The main advantages of the so-called Fourier-t-transformation (FTT) Eq. (1) are:

- Its characteristics are in a close relationship with those of "natural" spectrum analyzing systems such as filter banks, resonators, and the inner ear. The effective analysis bandwidth is $B=a/\pi$, and the effective time window length is $T=1/a$.
- It can readily be adapted to the characteristics of the ear. We found that an adequate analysis bandwidth is of the order of 1/10 critical bandwidth. This means that at frequencies higher than about 1 kHz B

is about 2% of frequency, while at low frequencies (below about 500 Hz), it is about 10 Hz. Consequently, effective time-window length T is about 30 ms at low frequencies and gets proportionally shorter at high frequencies.

- It is optimal with respect to both frequency- and time-resolution; this is expressed by the BT-product, which equals 1/., i.e. about 0.3. Compared to other methods of spectrum analysis, that is an outstandingly small value.
- The FTT-spectrum can be digitally calculated by an efficient recursive algorithm described earlier (Terhardt, 1985a).

Signal representation by the frequency- and time-continuous FTT spectrum makes apparent that discrete harmonics of periodic signals such as voiced speech sounds are not "objectively present" in an immediate physico-mathematical sense. They are implicitly represented by the maxima, as a function of frequency, of the absolute magnitude of $P(f,t)$. To become "real", they must be discerned from the continuous spectrum by a categorization process. That categorization process is conceptually significant, i.e., non-trivial; it may be regarded as a most elementary achievement of abstraction.

It is interesting to note that the signal characteristics are essentially represented by the absolute magnitude of the FTT-spectrum alone, provided that the constant a is properly specified as a function of frequency. With the aurally adequate specifications given above, the magnitude spectrum reflects very well both the aurally observed frequency- and time-resolution of speech.

Regarding the communication chain in Fig. 2., the next question to be considered is: In what way is the ear's stimulus (described by FTT) affected and changed by the transmission path, and what are the possibilities and limitations of a source-directed receiver to make a distinction between stimulus properties pertinent to the source, and properties pertinent to transmission (e.g. room acoustics)? Provided that the transmission channel does not have nonlinear distortion, that problem clearly is one of linear systems theory and the FTT-concept may be useful in its aurally adequate solution. That type of linear systems theory has not yet been fully developed.

In the diffuse sound field of a reverberating room the auditory stimuli can be considerably distorted. From the investigations of Schroeder & Kuttruff (1962), Plomp & Steeneken (1971), and others one can draw the following conclusions concerning sound transmission from source to receiver.

- The absolute magnitude of the transmission factor, as a function of frequency, shows pronounced peaks and dips at intervals of a few Hz. The mean magnitude difference between adjacent peaks and dips is about 10 dB. The mean frequency spacing of peaks is about $4/T_N$ (T_N = reverberation time).
- The phase angle, as a function of frequency, can attain large values and is essentially random.
- When either the source or the receiver moves in the diffuse sound field, the absolute magnitude of the transmission factor varies in a quasi-random way, with a standard deviation of about 5.6 dB.

4. ON THE ROLE OF PITCH IN SPEECH

The above review makes apparent that the precision and reliability with which the auditory stimuli represent the original source signal is quite limited. Under these conditions it is only the "frequency-time contours" of the ear-signal magnitude spectra, i.e., differential rather than absolute values, which can provide information about the source signal. The "auditory Gestalt processor", if it is as "intelligent" and flexible as appears plausible to presume, will consequently react with strategies of "auditory contour perception".

As mentioned already, the spectral pitches produced by the peripheral auditory system can be seen as auditory analogs of visual contours. While we have used the concept of the spectral-pitch pattern earlier in modeling specific auditory phenomena, i.e., perception of virtual pitch (Terhardt, 1972; 1974; 1979a,b; Stoll, 1982) and influence of spectral fine structure on timbre (Benedini, 1979), the present approach is more general and ambitious, suggesting that in auditory perception in general spectral pitches play a role which is analogous to that of contours in visual perception. We put forward a concept according to which in natural speech the time-variant pattern of spectral pitches as such essentially provides the information about formant structures (and perhaps other clues of the speech signal). Spectral pitches are produced by sufficiently pronounced maxima of the stimulus FTT spectrum, no matter whether those maxima pertain to individual harmonics (which is the case at lower spectral frequencies), or formants (at higher frequencies). Like visual contours, spectral pitches provide to the "central processor of peripheral information" binary cues, the occurrence - or nonoccurrence - of which characterizes an auditory Gestalt. It is suggested that that type of auditory Gestalt is of ultimate importance in the perception of both music and speech.

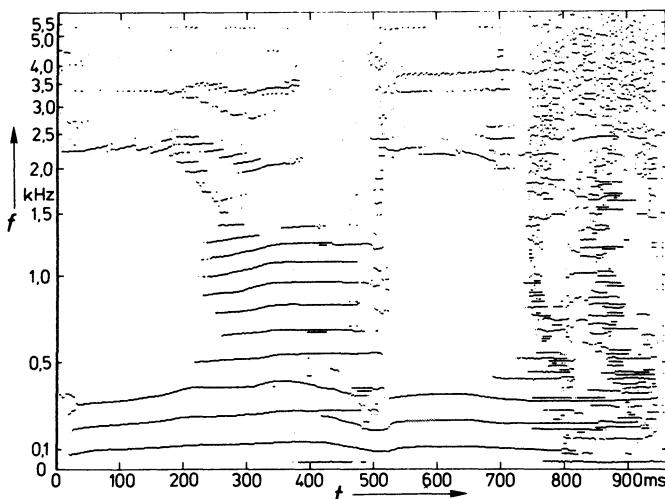


FIGURE 3, Traces of maxima detected in the FTT magnitude spectrum of the phrase "Er geht" (male speaker); from Heinbach (1986).

Fig. 3 shows an example of spectral-pitch analysis of speech (kindly provided by W. Heinbach). In an investigation of the FTT spectrum, Heinbach (1986) designed an algorithm detecting quasi-instantaneously spectral maxima from the smoothed magnitude spectrum. Fig. 3 shows the time- frequency contours of spectral maxima from the German utterance "Er geht" (male speaker), reflecting with a close approximation the spectral-pitch-time pattern. Analysis bandwidth and frequency scaling are based on the critical-bandwidth scale, as outlined in section 3. One can see a sort of skeleton representation of the speech signal, which nicely reflects essential characteristics. In particular, the vowel formants are represented by the mere presence of detectable maxima (i.e., spectral pitches) in the corresponding frequency regions, while outside the formant regions spectral pitches are missing. By resynthesis, yielding a speech signal of fair to good quality, Heinbach was able to demonstrate that the spectral-pitch-time pattern includes most of the relevant information.

It is apparent that the "intelligent Gestalt processor", once it has detected the lowest existing spectral pitch of a voiced speech segment, can in principle predict the positions, on the spectral-frequency scale, where other spectral pitches, i.e. those of the second and higher harmonics, must lie. From the actual occurrence or absence of those spectral pitches, the processor can draw conclusions on the gross spectral envelope, in particular, formant regions. Obviously, at pursuing that strategy, knowledge of the actual "fundamental pitch" in every instant of time is of decisive importance to the Gestalt processor. That knowledge can in many cases be acquired by simply detecting the lowest existing spectral pitch. However, when the auditory stimulus is weak, even in natural situations the lowest harmonic's spectral pitch may be little pronounced or even missing (In a telephone channel, this is true all the time, because of band limitation). Fig. 4 shows theoretical spectral-pitch patterns of the vowels /ə/ and /a/, with 200 Hz fundamental frequency and SPL's of 40, 60, and 80 dB. These calculations are based on mutual masking of components, following the procedure described earlier (Terhardt, 1979a). The prominence of each spectral pitch is expressed by the so-called sound-pressure-level excess (ordinate). It is apparent that the prominence of the fundamental's spectral pitch is distinctly dependent on SPL. While at higher SPL's it is the most prominent of all spectral pitches, it gets insignificant below about 40 dB SPL, due to the influence of the threshold of hearing.

Pursuing the concept of auditory Gestalt perception reveals that for an intelligent processor familiar with basic characteristics of the speech signal, such as harmonicity of vowel spectra, it is not necessary that the fundamental pitch is actually represented by a detectable spectral maximum (spectral pitch). Rather, fundamental pitch can be inferred from the spectral pitches of higher harmonics. The process which makes that particular contribution was conceptualized and psychoacoustically justified many years ago (Terhardt, 1972; 1974), and is called virtual-pitch formation. Though even a harmonic complex tone such as a speech vowel does not have just one pitch but is pitch-ambiguous, the spectral pitches of harmonics can be safely

predicted, because the pitch-ambiguity is not random but systematic and typical of harmonic complex tones (cf. Terhardt et al., 1982).

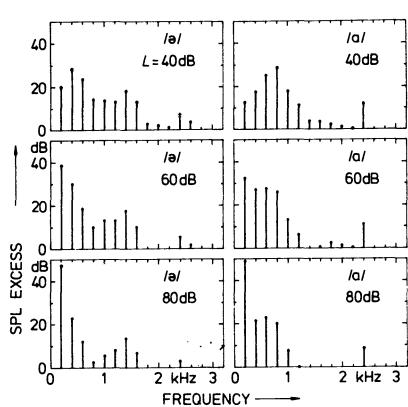


FIGURE 4, Calculated spectral pitch patterns of the vowels /ə/ and /a/ at 40, 60, and 80 dB SPL.

Procedure of calculation see in Terhardt (1979a).

Note that the prominence of the fundamental's spectral pitch at lower SPL is distinctly dependent on SPL.

In the strategy of "formant detection by spectral-pitch prediction" proposed here, harmonicity of the vowel spectrum is a crucial precondition, which is based on the fact that natural voiced segments are spectrally harmonic. Therefore it is not surprising that, on the one hand, identification of natural vowels is, over a wide range, independent of fundamental frequency, while on the other hand, pronounced inharmonicity reduces both vowel identification and goodness rating (see appendix). Our experimental data suggest that vowels with tonal but inharmonic spectra are even less well identified and goodness-rated than random-noise vowels (which are comparable to whispered vowels).

The above concept significantly contributes to the problem, recently addressed by several authors, of perceptually separating simultaneous voices (see the corresponding contributions in this volume). Since the virtual-pitch algorithm is not dependent on a priori information about the number of sources, it "automatically" provides the Gestalt processor in the next higher layer with information about commonality or non-commonality of spectral pitches. Although we have not yet tested that aspect in detail, the prospects of extracting phonetic information from intermingled spectral-pitch patterns produced by different speech sources appear promising.

The present view on the role of pitch (in the psychoacoustic sense of the term) throws a new light on the role of fundamental pitch (in the phonetic sense): instead of being just a by-product which for speech communication is of secondary (prosodic) significance, virtual pitch (fundamental pitch) gains the position of an essential tool in auditory extraction of phonetic information.

APPENDIX: EXPERIMENTS ON THE INFLUENCE OF SPECTRAL AND VIRTUAL PITCH ON PERCEPTION OF ISOLATED SYNTHETIC VOWELS

The following variables were considered:

1. Fundamental frequency of periodically pulse-excited, i.e. spectrally harmonic, vowels. This variable is equivalent to width of harmonic spacing, i.e. sampling of spectral envelope.
2. The relationship between fundamental and formant-frequencies. Fundamental frequencies of harmonic vowels were chosen such that certain harmonic frequencies were in controlled relationships to either the first or second formant frequencies, i.e., either coincided or deviated by a definite amount.
3. Prominence of pitch: By random-noise excitation of vowels any type of pitch dependent on the source as such was removed.
4. Harmonicity of spectral pitches. Vowels were generated by superimposing partials whose frequencies were quasi-random ("inharmonic" spectra).

Perceptual criteria were identification and goodness rating (see below).

Table 1. Formant frequencies of synthesized vowels; kHz

	F1	F2	F3	F4
/a/	0.85	1.20	2.8	3.5
/e/	0.35	2.25	3.0	3.5
/i/	0.25	2.20	3.0	3.5
/o/	0.35	0.53	2.6	3.5
/u/	0.30	0.87	2.2	3.5
/ae/	0.86	2.05	2.8	3.5
/oe/	0.40	1.70	2.0	3.5
/y/	0.25	1.67	2.0	3.5

Table 2. Fundamental frequencies used; Hz

/a/	85	200	100	130	163
/e/	87	107	144	175	195
/i/	82	116	150	162	200
/o/	117	133	88	156	200
/u/	97	121	125	174	200
/ae/	86	100	121	171	195
/oe/	81	100	117	160	189
/y/	83	100	128	167	200

Steady vowels were computer-generated using spectral envelopes calculated on the basis of Fant's (1960) theory by the algorithm described earlier (Terhardt, 1979a). The signals were produced through a 12-bit D/A converter and were low-pass filtered with a 5 kHz cut-off frequency. In previous experiments the formant frequencies shown in Table 1 yielded good quality vowels. The fundamental frequencies used in the harmonic condition are shown in Table 2. Noise excitation was simulated using many spectral components with narrow, random frequency spacing and randomization of amplitudes. By means of separate listening experiments it was verified that the quasi-random noise so obtained was aurally indistinguishable from true random noise. Inharmonic spectra were synthesized by means of spectral components

with random frequencies such that their mean and minimal frequency spacing were 120 Hz and 8 Hz, respectively. The test stimuli were presented binaurally through earphones, with 300 ms duration and 60 dB SPL.

Seven subjects (male, age 24-28, with normal hearing) participated. In the identification experiments they were informed about the set of vowels to be presented, and were asked to identify, after each presentation, each test vowel by pressing a labeled button. In the pulse-excited, i.e. harmonic condition, each vowel provided five different test stimuli corresponding to the five fundamental frequencies. In each of two runs, each test stimulus was presented five times, and the stimuli were presented in random order.

In the goodness rating experiment, the procedure was the same as before, but subjects were asked to evaluate the degree to which each test stimulus was typical of the vowel category perceived, whatever that category was. Information on vowel category itself was neither given to, nor required from, the Ss. Subjects were instructed to respond by writing down, after each test stimulus presentation, a figure ranging from 0 (untypical of any vowel category) to 10 (most typical). In this experiment, only five (instead of eight) inharmonic vowels were included.

Table 3. Identification percentages of vowels in the harmonic, random-noise, and inharmonic conditions.

	/a/	/e/	/i/	/o/	/u/	/ae/	/oe/	/y/	mean
HARMONIC	100	98	97	88	95	100	83	95	94
NOISE	100	72	95	63	93	98	83	93	87
INHARMONIC	98	28	90	34	78	77	25	80	64

Table 3 shows the results of the identification experiment, i.e. the relative numbers of correct identifications. In the harmonic condition, the influence of fundamental frequency (in the range investigated) was insignificant. Therefore, the means of all fundamental frequencies are shown. One can see that in the random-noise and inharmonic conditions identification is, for some vowels, reduced as compared to harmonic excitation. In the inharmonic condition this effect is more pronounced than in the random noise condition. It was verified that these differences are significant on an error level of less than 1% (The mean values in the harmonic condition are based on 350 responses per vowel, i.e. 10 responses, 7 Ss, 5 fundamental frequencies; those of the noise and inharmonic conditions correspond to 70 responses). The following correlations between the data from different conditions (across vowels) are found:

- harmonic vs. noise: insignificant ($R=0.59$);
- harmonic vs. inharmonic: significant (5%; $R=0.72$);
- noise vs. inharmonic: significant (< 5%; $R=0.88$).

These figures indicate that identification performance is dependent both on vowel category (i.e., spectrum envelope) and type of excitation (i.e., spectral fine structure).

Table 4. Goodness ratings in the harmonic, noise-excited, and inharmonic conditions (arbitrary units 0-10).

	/a/	/e/	/i/	/o/	/u/	/ae/	/oe/	/y/	mean
HARMONIC	7.7	7.0	5.6	5.1	4.5	6.5	5.2	6.6	6.0
NOISE	5.0	3.5	3.2	1.9	2.7	4.4	4.1	3.8	3.6
subset	"	"	"	"	---	---	"	---	3.5
INHARMONIC	6.6	0.2	3.5	1.1	---	---	1.5	---	2.5

The results of the goodness-rating experiment are shown in Table 4. In the harmonic condition the mean values of the fundamental frequencies are given, since there was no significant influence of fundamental frequency. (Correlation of rating with fundamental frequency was practically zero). Although in the inharmonic condition only a subset of vowels was included, it is evident that the data essentially follow the same pattern as those of the identification experiment. In particular, goodness rating in the noise and inharmonic conditions is distinctly less than in the harmonic condition, and that reduction is most pronounced in the inharmonic condition. The identification scores and goodness ratings (within conditions, across vowels) correlate with fair to high significance (harmonic condition: R=0.67; noise: 0.69; inharmonic: 0.91).

In the harmonic conditions of both experiments, particular attention was paid to the influence of the harmonic-vs.-formant-frequency relationship. Statistical analysis revealed that there is no significant influence of the degree to which individual harmonics depart from either the first or the second formant frequency. The corresponding correlation coefficients are practically zero at an error level of less than 1%.

REFERENCES

1. Benedini, K. (1979). Ein Funktionsschema zur Beschreibung von Klangfarbenunterschieden. *Biol. Cybernetics*, 34, 111-117.
2. Eccles, J.C. (1973). The Understanding of the Brain. McGraw-Hill, New York.
3. Elman, J.L. (1981). Effects of frequency-shifted feedback on the pitch of vocal productions. *J. Acoust. Soc. Am.*, 70, 45-50.
4. Fano, R.M. (1950). Short-time autocorrelation functions and power spectra. *J. Acoust. Soc. Am.*, 22, 546-550.
5. Fant, G. (1960). Acoustic Theory of Speech Production. Mouton, The Hague.
6. Flanagan, J.L. (1972). Speech Analysis, Synthesis and Perception. Springer, New York, 2nd ed.
7. Fletcher, H. (1953). Speech and Hearing in Communication. Van Nostrand, New York, 287-290.
8. Gambardella, G. (1971). A contribution to the theory of short-time spectral analysis with nonuniform bandwidth filters. *IEEE Trans. CT-* 18, 455-460.
9. Gottfried, T.L. (1986). Intelligibility of vowels sung by a countertenor. *J. Acoust. Soc. Am.*, 79, 124-130.
10. Heinbach, W. (1986). Untersuchung einer gehörbezogenen Spektralanalyse mittels Resynthese. In: Fortschritte der Akustik, (DAGA'86), Bad Honnef, pp. 453-456.

11. Hofstadter, D.R. (1979). Gödel, Escher, Bach: An Eternal Golden Braid. Basic Books, New York.
12. Houtgast, T. (1974). Auditory analysis of vowel-like sounds. Acustica, 31, 320-324.
13. Köhlmann, M. (1985). Bestimmung der Silbenstruktur von fliessender Sprache mit Hilfe der Rhythmuswahrnehmung. Acustica, 56, 120-125.
14. Ohm, G.S. (1843). Über die Definition des Tones, nebst daran gekloppter Theorie der Sirene und ähnlicher tonbildender Vorrichtungen. Ann. Phys. Chem., 59, 513.
15. Plomp, R. (1967). Pitch of complex tones. J. Acoust. Soc. Am., 41, 1526-1533.
16. Plomp, R. and Mimpin, A.M. (1968). The ear as a frequency analyzer. J. Acoust. Soc. Am., 43, 764-767.
17. Popper, K. (1972). Objective Knowledge - an Evolutionary Approach. Clarendon Press, Oxford.
18. Repp, B.H. (1981). On levels of description in speech research. J. Acoust. Soc. Am., 69, 1462-1464.
19. Ritsma, R.J. (1967). Frequencies dominant in the perception of the pitch of complex sounds. J. Acoust. Soc. Am., 42, 191-198.
20. Ryalls, J.H. and Lieberman, P. (1982). Fundamental frequency and vowel perception. J. Acoust. Soc. Am., 72, 1631-1634.
21. Schroeder, M.R. and Kuttruff, H. (1962). On frequency response curves in rooms. Comparison of experiment, theoretical, and Monte Carlo results for the average frequency spacing between maxima. J. Acoust. Soc. Am., 34, 76-80.
22. Schroeder, M.R. and Atal (1962). Generalized short-time power spectra and autocorrelation functions. J. Acoust. Soc. Am., 34, 1679-1683.
23. Searle, C.L. (1982). Speech perception from an auditory and visual viewpoint. Canadian J. Psychol., 36, 402-409.
24. Stoll, G. (1982). Spectral-pitch pattern. A concept representing the tonal features of sounds. In: M. Clynes (Ed.) Music, Mind, and Brain, 271-278. Plenum Press, New York.
25. Terhardt, E. (1972). Zur Tonhöhenwahrnehmung von Klängen. Acustica, 26, 173-199.
26. Terhardt, E. (1974). Pitch, consonance, and harmony. J. Acoust. Soc. Am., 55, 1061-1069.
27. Terhardt, E. (1979a). Calculating virtual pitch. Hearing Research, 1, 155-182.
28. Terhardt, E. (1979b). On the perception of spectral information in speech. In: O. Creutzfeldt, H. Scheich, and Chr. Schreiner (Eds.), Hearing Mechanisms and Speech, 281-291. Springer, Heidelberg/New York.
29. Terhardt, E., Stoll, G., and Seewann, M. (1982). Algorithm for extraction of pitch and pitch salience from complex tonal signals. J. Acoust. Soc. Am., 71, 679-688.
30. Terhardt, E. (1985a). Fourier transformation of time signals: Conceptual revision. Acustica, 57, 242-256.
31. Terhardt, E. (1985b). Verfahren zur gehörbezogenen Frequenzanalyse. In: Fortschritte der Akustik, (DAGA'85), 811-814.

DOES THE HUMAN AUDITORY SYSTEM INCLUDE LARGE SCALE SPECTRAL INTEGRATION?*

J.L. Schwartz, and P. Escudier

Institut de la Communication Parlée, INPG-CNRS
46 Av. Félix Viallet, 38031 Grenoble Cedex - France

INTRODUCTION

The hypothesis of a large scale frequency integration mechanism in the auditory system was introduced long ago in vowel perception theories. It underwent two main developments in the last 15 years, first through the F'2-concept (see 2, 5), secondly with the "center of gravity effect" and the 3.5 Bark critical distance of Chistovich and colleagues (6, 7). This 3.5 Bark critical distance could provide a very powerful tool for understanding some of the main properties of vowel systems (see 10, 18, and Bladon, this volume), but experimental evidence for the critical distance concept is very scarce. We shall try here to reinforce this experimental basis, around three questions. First, can we understand data about F'2 matching tests and about the center of gravity effect without a large scale spectral integration with a 3.5 Bark critical distance? Second, does such a mechanism explain the seemingly contradictory results about the part played by formant amplitudes in front vowel identification? Third, what could be its contribution in the auditory representation of the labiality opposition for front vowels?

I. ABOUT TWO "REDUCING" ASSUMPTIONS CONCERNING THE F'2 - CONCEPT AND THE CENTER OF GRAVITY EFFECT

Bladon (3) proposed that a common mechanism could be responsible for both F'2 experimental data and the center of gravity effect, owing to 3.5 Bark integration. This corresponds to a structure of the auditory system such as the one described in Figure 1a, and used in several studies (9, 19). Indeed, psychophysical matching tests leading to both sets of results are quite similar in principle. This assumption implies that F'2 - concept and the critical distance do correspond to a perceptual reality. But we find in the literature at least two other hypotheses that we can describe as "reducing", since they implicitly introduce a dependence relationship between these experimental facts, and two very general psychophysical mechanisms: the critical band concept on one hand, and phonemic identification on the other. We shall test both of them.

*The influence of C. Abry, from the Phonetic Institute of Grenoble, on this whole work is particularly important: every discussion with him on the topic led to new insights or proposals. Thanks also are due to R. Carré, for discussions... and for his ear as the third subject.

I.1 Could the 3.5 Bark critical distance be deduced from 1 Bark critical bands?

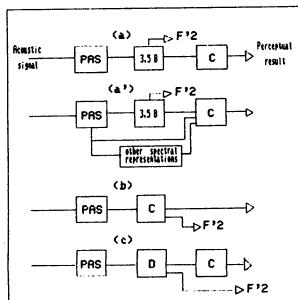


FIGURE 1, Possible structures for the auditory system (see text).

- P.A.S. Signal analysis in the peripheral auditory system
- 3.5 B 3.5 Bark spectral integration
- C Phonemic classification mechanism
- D Distance for classification

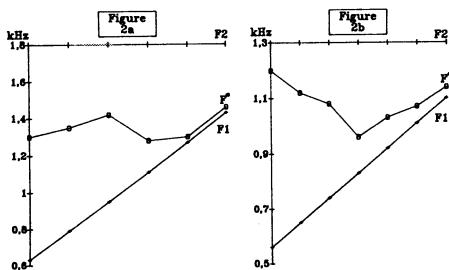


FIGURE 2, Center of gravity effect.

- a/ Experimental results from (6)
- b/ Simulation results, B & L model

This possible conclusion from B & L's work, which was not explicit in their paper, is very interesting because it is quite "economical": the 3.5 Bark critical distance would be only a byproduct of the peripheral 1 Bark critical bands, and, in consequence, we would not need any large scale frequency integration mechanism to understand this whole set of data!

In 1981, Bladon and Lindblom (4) (B & L in the following) used a rather classical model for computing auditory spectral distances, defined as integrated absolute differences between Sone/Bark representations, for predicting F'2 experimental results. They prepared two 4-formant reference signals corresponding to [i] and [y], and for each of them a set of 2-formant test signals, with F1 fixed to the first formant of the reference signal and varying F2 values, and showed that the specific F2 value producing the least auditory difference between reference and test signal in psychoacoustic experiments also produced the least estimated distance in their model. Hence, F'2 values could be predicted on the basis of the Euclidian distance between Sone/Bark internal representations. This corresponds to structure 1b in Fig. 1, where F'2 would be a by-product of a "distance computer" used by the auditory system for phonemic classification, and hence would have no intrinsic reality. Keeping in mind our first assumption that F'2 data and the center of gravity effect are two manifestations of a single mechanism, we may wonder whether this model would also predict this second effect, and the 3.5 Bark critical distance.

We tried to test this hypothesis by systematically reproducing experiments of Chistovich and colleagues on B & L's model. Their first critical distance estimates (7) were made with two-formant (F1, F2) reference signals of fixed F2 and decreasing F1 values, by systematically determining for each F1 value the F* formant value of a one-formant test signal giving the smallest auditory distance between reference and test signal. For F1-F2 separations lower than the critical distance, F* decreased with F1, and for greater F1-F2 separations, F* "left" F1 and increased back towards F2. This experimental pattern is reproduced in Fig. 2a for F2 = 1.8 kHz, with an arrow pointing to the beginning of the F* rise. In our simulation, for each (F1, F2) reference signal we determined the F* value producing the smallest distance in the B & L model. Results for F2 = 1.3 kHz are displayed in Fig. 2b.

Table I: observed and predicted critical distances

F2(kHz)	observed (from (6))	predicted (from Fig. 2)
0.7	3.0 Barks	2.5-3.0 Barks
1.0	3.4 Barks	2.5-3.0 Barks
1.3	3.2 Barks	3.0-3.5 Barks

In Table I, we give experimental critical distance estimates together with our estimates. It is quite difficult to conclude whether our results are compatible with those of Chistovich and colleagues, all the more so since quantitatively precise estimates are very difficult both in psychophysical experiments and in simulation, where they strongly depend on quantitative data such as formant bandpasses, critical bands in peripheral analysis, exact definition of amplitude compression linked to the Sone function, lack of two-tone suppression phenomena in the B & L model, etc. We shall come back to this question in the last part of this paper.

I.2. Could F'2 adjustments be deduced from categorical identification?

In their first paper on the topic, Carlson and colleagues studied F'2 variations in the [i]-[y] region, with four-formant synthetic signals characterized by F1, F2, and F4 fixed respectively to 255, 2000 and 3350 Hz, and F3 values increasing from 2300 to 3000 Hz. They noticed a 900 Hz F'2-shift from 2400 to 3300 Hz for a 300 Hz F3-variation around 2700 Hz, and suggested that classification could be involved in the matching: they judged that such a large F'2 variation was difficult to explain on the basis of purely auditory mechanisms, and assumed that the subjects first identified the sound in a given phonetic category (namely [i] or [y]) and then adjusted F'2 values according to their internal model of each category. Other experiments about formant levels and spectrum identification, to which we shall come back in chapter II, reinforced their conclusion that classification came before F'2 extraction. This corresponds to an auditory system structure as described in Fig. 1c, in which F'2 would be a product of a phonemic decision, and not a cue for decision, and, hence, would once more have no intrinsic perceptual reality.

But we can propose another hypothesis to account for this F'2 instability around an F3 position of 2700 Hz. Indeed, this position corresponds more or less to the middle, in Bark, of the F2 and F4

positions. Therefore, a spectral integration model can predict such an F'2 jump, whether the main energy mass exists in the F2-F3 group or in the F3-F4 group. This was clearly proved by a previous study carried out by our group on such a model (9).

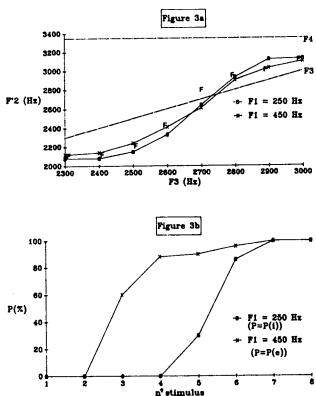


FIGURE 3, F'2 and identification tests.

In a, F = formula predictions for F1=250 or 450 Hz (see III.1). Stimuli used in b:

- 1 F1=250 or 450, F2=1700,
 F3=2000, F4=3350Hz
- 2 F1=250 or 450, F2=1700,
 F3=2300, F4=3350Hz
- 3-8 F1=250 or 450, F2=2000,
 F3=2300, 2500, 2600,
 2700, 2800, 3000,
 F4=3350 Hz

To determine which of these two propositions was correct, we reproduced Carlson and colleagues' F'2 experiments with a 450 Hz F1 value, so as to enter the [e]-[ø] region. In Fig. 3a, we display the mean F'2 adjustments we obtained with three subjects, for both F1 values. Clearly, the relationship between F3 and F'2 is the same, with the same F'2 jump of about 700 Hz for a 300 Hz F3 shift from 2500 to 2800 Hz. On the other hand, identification tests on the same signals show very different frontiers between the [y]-[i] case and the [ø]-[e] one (Fig.3b). Clearly, the F'2 jump cannot be a product of categorical identification, while the second hypothesis easily accounts for this whole set of data. Hence, structure (c) in Fig. 1 does not account for this new experiment, while structure (a) does.

II. FORMANT AMPLITUDES AND FRONT VOWEL IDENTIFICATION

The proposition by Carlson and colleagues on the part played by categorization in F'2 adjustments was also motivated by their identification results, in which the [y]-[i] frontier for 4-formant signals with F1, F2, F4 fixed at the same values as in Fig.2 was obtained for an F3 value (around 2750 Hz) which was almost constant for F2 and F4 levels which had been increased or decreased by as much as 24 dB! They concluded that it appeared difficult to imagine any spectral integration-like formalism extracting F'2 value before the vowel would be identified. Indeed, how can a model based on Fig. 1a account for the seeming lack of a contribution of formant amplitudes to front vowel identification? Notice, however, that, identification data obtained by Aaltonen (1) show on the contrary a clear formant level effect. We shall show that a 3.5 Bark integration mechanism does give an explanation for these apparent contradictions.

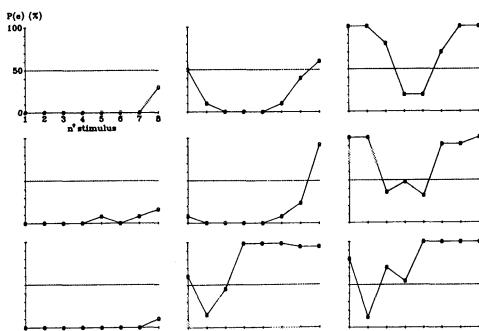


FIGURE 4, F2 and F3 amplitudes and e/ identification.

One subject/row, one condition/column (see text).

Stimuli used:

all stimuli: $F_1 = 450$, $F_2 = 1700$, $F_4 = 3350$ Hz. From 1 to 8, $F_3 = 2000, 2100, 2200, 2300, 2400, 2500, 2600, 2700$ Hz

important energy grouping characteristics. In the case of Aaltonen's 3-formant signals such groupings do not occur, since F_4 is absent.

This conclusion is reinforced by identification data reported in Fig.4, where we submitted three subjects to 4-formant signals with F_1 , F_2 and F_4 fixed respectively to 450, 1700, and 3350 Hz (hence, F_2 and F_4 are at a distance of more than 3.5 Barks, and only F_2 and F_3 should contribute to the F'_2 estimate), with varying F_3 , and for three different formant amplitude conditions: first a "standard" one, with A_2 and A_3 (F_2 and F_3 amplitudes) deduced from a cascade model (C1), then for $A'_2 = A_2 - 12$ dB (C2), and finally for $A'_2 = A_2 - 12$ dB and $A'_3 = A_3 + 12$ dB (C3). Subjects were instructed to classify these sounds as [e] or [ø]. Results vary from one subject to another, but the trend is clear: for such a situation where only two formants should be integrated by the 3.5 Bark integration mechanism, F_2 and F_3 amplitudes play an obvious part in phonemic identification. We can conclude that formant amplitudes do contribute to front vowel identification, at least for non-low front vowels, and that once more the whole set of experimental data on this topic can be explained with the structure proposed in Fig. 1a.

We should notice in this figure that a simple frequency integration mechanism with no other previous spectral treatment cannot account for the non-monotonous dependence of [e]-identification responses on F_3 values. An analysis of signal spectra shows that high levels of [e]-responses for the first sound in condition (C2) and for the first 2 or 3 sounds (depending on subjects) in condition (C3) correspond to cases where F_2 does not appear as a local spectral maximum. This can be linked to a crucial point of Chistovich's theory (8) : before 3.5 Bark integration there would be a special inhibition-like mechanism reinforcing spectral peaks. This mechanism probably

We showed in Fig. 3a that, with the 4-formant signals used by Carlson et al., there was a steep F'_2 increase around an F_3 value of about 2700 Hz, corresponding to the middle of F_2 and F_4 in Bark. When F_2 or F_4 amplitudes are changed, the position of this sudden F'_2 increase will be only slightly modified, since it is mainly due to the passage from an F_2 - F_3 grouping to an F_3 - F_4 grouping and, hence, determined by the F_3 position compared to the F_2 - F_4 middle. If we assume that the [y]-[i] frontier corresponds to an F'_2 value close to 2700 Hz, according to our results of Fig.3, this explains why the identification data of Carlson et al. with 4-formant signals do not seem to depend on formant levels: in fact, it is a case where formant amplitude influence is hidden by more

operates, for (C2) for example, for the second sound and not for the first one, which should provoke an $F'2$ decrease for this second stimulus, and the observed decline in [e]-responses between first and second stimulus.

III. THE $F'2$ CONCEPT AND LABIALITY OPPOSITION IN FRENCH FRONT VOWELS

We present here a first report on the representation of the labiality opposition in a hypothetical perceptual ($F1$, $F'2$) space compared with its representation in the acoustic ($F1$, $F2$) space. We shall first propose a formula for predicting $F'2$ values from formant data, and describe the ability of this formula to predict experimental data, and secondly use this transformation formula in the study of an articulatory-acoustic corpus elaborated for the study of labiality in French.

III.1. A two-formant model based on the 3.5 Bark integration concept

A first model was elaborated, according to Fig.1a, and gave good predictions of $F'2$ experimental results (9). The formula we shall use in this work was deduced from the properties of this model. First, formant amplitudes are computed from formant frequencies according to a cascade model (11), with bandpasses also estimated from formant frequencies (12). Then, all frequencies are converted to Bark, and all amplitudes to Phones, according to a classical dB-Phone transformation (20). $F'2$ values are finally estimated under three possible conditions, D being the 3.5 Bark critical distance:

- * if $F3-F2 > D$, $F'2 = F2$;
- * if $F3-F2 < D$ and $F4-F2 > D$, or $F4-F2 = < D$ and $F4-F3 > F3-F2$, $F'2$ is the center of gravity of $F2$ and $F3$;
- * if $F4-F2 < D$ and $F4-F3 < F3-F2$, $F'2$ is the center of gravity of $F3$ and $F4$.

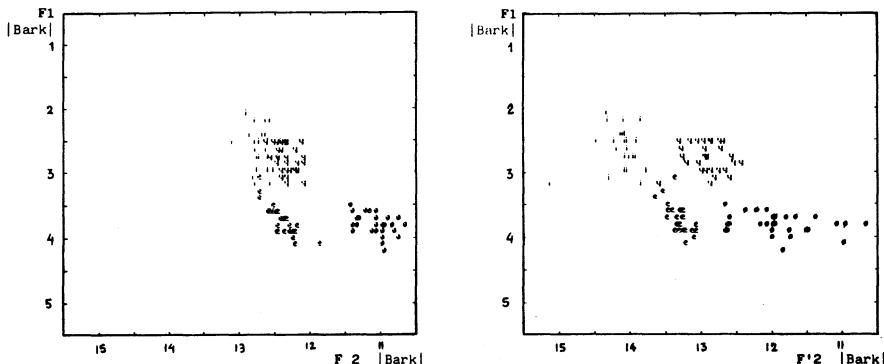


FIGURE 5. Acoustic ($F1$, $F2$) (a), and "perceptual" ($F1$, $F'2$) (b) representations of rounding opposition for one speaker of the corpus.

Transitions from one condition to another are smoothed, and two-tone suppression phenomena are introduced. A more complete description is given in (17). Prediction results are given in Table II for the two experimental $F'2$ corpuses available in the literature, together with

predictions obtained with the formula proposed by Bladon and Fant (B & F). The results are generally rather good for both formulae, with an advantage to the new model we propose, since, as Bladon notes, it is not susceptible to an ambiguity error in cases where the B & F model is (3). It also performs very well for the corpus used in I.2 (see Fig. 3a).

Table II: adjusted
and computed F'2
values

F'2a: adjusted; F'2c: computed	F1	F2	F3	F4	F'2a	F'2c NM	F'2c BF	Z'2a-Z'2c NM	Z'2a-Z'2c BF
Z'2a - Z'2c:	u	310	730	2250	3300	730	730	0.00	0.00
prediction error, in	o	400	710	2460	3150	720	710	0.07	0.06
Barks	ø	360	1690	2200	3390	1720	1879	1737	-0.59
BF: from (2), NM:	a	580	940	2480	3290	960	940	0.12	0.10
New Model, from	y	255	1930	2420	3300	2010	2119	2156	-0.35
(17)	ɛ	280	1630	2140	3310	1730	1817	1674	-0.33
top: corpus from (5)	e	375	2060	2560	3400	2370	2288	2370	0.23
bottom: corpus from	æ	605	1550	2450	3400	1960	1835	1638	0.44
(2)	i	255	2065	2960	3400	3210	3079	3064	0.26
									0.30

	F1	F2	F3	F4	F'2a	F'2c NM	F'2c BF	Z'2a-Z'2c NM	Z'2a-Z'2c BF
ɔ	670	1050	2900	3490	947	1050	1068	-0.62	-0.72
ɑ	660	1170	2770	3650	1103	1170	1180	-0.37	-0.42
ɛ	290	700	2550	3280	669	700	701	-0.23	-0.24
ɔ	570	840	2640	3310	806	840	842	-0.23	-0.25
ɔ	370	730	2670	3240	700	730	733	-0.22	-0.24
ɛ	360	1550	2430	3030	1503	1859	1910	-1.41	-1.60
ɛ	300	1320	2480	3440	1300	1320	1351	-0.10	-0.25
ʌ	620	1260	2390	3610	1284	1260	1266	0.12	0.09
OE	700	1430	2390	3350	1458	1498	1464	-0.18	-0.03
ɪ	450	1300	2640	3470	1326	1300	1343	0.13	-0.08
ø	460	1520	2290	3290	1570	1786	1570	-0.86	0.00
a	770	1400	2460	3710	1452	1400	1409	0.24	0.20
t	380	1690	2460	3570	1754	1966	1747	-0.76	0.03
e	470	2180	2720	3790	2361	2390	2326	-0.08	0.10
t	680	1890	2580	3940	2076	2145	1925	-0.22	0.51
y	300	1890	2250	3000	2101	2031	2084	0.23	0.05
œ	640	1450	2330	3030	1637	1676	1589	-0.16	0.20
i	300	2300	3070	3590	3095	3290	3154	-0.39	-0.12

III.2. Application to an acoustic corpus

We worked on a corpus prepared by the researchers of the Grenoble Phonetic Institute for a very complete articulatory study about labiality in French (15). It contains vowels [i], [y], [e], [ϕ], preceded by fricatives [s], [z], [ʃ], and [ʒ]. Each of these 16 CV combinations is represented by 3 logatoms pronounced once each by 5 speakers. Hence, this corpus is made of rounded and non-rounded front vowels in consonantal contexts, making for maximum and minimum rounding opposition. For such front vowels we expect clear differences between acoustic (F_1 , F_2) space and "perceptual" (F_1 , $F'2$) space. We present here only the very beginning of this study, obtained from a first speaker (Fig. 5). In the left panel of Fig. 5, separation between [e] and [ϕ] in the acoustic space is quite good, while there is a small overlap between [i] and [y] regions. In the right panel of Fig. 5, separation between [i] and [y] in the (F_1 , $F'2$) space is much better, almost as good as between [e] and [ϕ]: we can talk of a "normaliza-

tion" of labiality opposition between high and non-high front vowels for this speaker.

This rather spectacular result does not prove that the F'2 concept will solve all problems in this study. In fact, several problems remain, linked to the exact choice of the critical distance D in our formula, and to the classical problem of formant detection. For this last aspect, a model working directly on the whole spectrum, such as the one we propose in (9), could give very interesting results. Anyway, this first study shows very promising trends about the possible role of the F'2 parameter as a "good" perceptual parameter for phonemic classification.

CONCLUSION - SCOPE AND LIMITS OF THE 3.5 BARK INTEGRATION CONCEPT

Chapter I left us with two possible theories explaining F'2 data and the "center of gravity effect", one (see Fig.1a) assuming the perceptual reality of a large scale spectral integration mechanism, the other one (see Fig. 1b) looking more "economical", since it proposed that peripheral 1- Bark frequency integration could account for all results in this field. Chapters II and III showed that the first assumption could explain quite a lot of facts about perception and categorization of front vowels, and that we should use the very profound explanatory consequences of a 3.5 Bark frequency integration mechanism in the study of phonetic systems. Hence it appears that the more "economical" model is most probably the first one, since it can explain a lot of results in different areas with a single and very simple mechanism.

Yet, a lot of perceptual results such as those of Fujimura (13) on front vowel identification, of Aaltonen (1) on response time in identification tests, of Klatt (16) on spectral valleys and phonetic distances, of several authors on B1 and nasality perception (see (14) for example), can only be understood on the basis of a slightly modified structure of the auditory system, described in Fig. 1a', where the 3.5 Bark integration mechanism supplies only one of several different spectral representations available to the central decision system. It is only with such a multi-representational structure, very classical in descriptions of the visual system, that we can understand at the same time the perceptual role of F'2, Chistovich's center of gravity, and the experimental results cited above.

REFERENCES

1. Aaltonen, O. (1985). The effect of relative amplitude levels of F2 and F3 on the categorization of synthetic vowels. Journal of Phonetics, 13, 1-9.
2. Bladon, R.A.W. and Fant, G. (1978). A two-formant model and the cardinal vowels. STL-QPSR, 1, 1-8.
3. Bladon, A. (1983). Two-formant models of vowel perception: shortcomings and enhancements. Speech Communication, 2, 305-313.
4. Bladon, A. and Lindblom, B. (1981). Modeling the judgment of vowel quality differences. Journal of the Acoustical Society of America, 69, 1412-1422.

5. Carlson, R., Granström, B., and Fant, G. (1970). Some studies concerning perception of isolated vowels. STL-QPSR, 2-3, 19-35.
6. Chistovich, L.A., Sheikin, R.L., and Lublinskaya, V.V. (1979). 'Centers of gravity' and the spectral peaks as the determinants of vowel quality. In: B. Lindblom and S. Ohman (Eds.) Frontiers of Speech Communication Research, 143-158. Academic Press, London.
7. Chistovich, L.A. and Lublinskaya, V.V. (1979). The center of gravity effect in vowel spectra and critical distance between the formants: Psychoacoustical study of the perception of vowel-like stimuli. Hearing Research, 1, 185-195.
8. Chistovich, L.A. (1980). Auditory processing of speech. Language and Speech, 23, 67-73.
9. Escudier, P., Schwartz, J.L., and Boulogne, M. (1985). Perception of stationary vowels: internal representations of the formants in the auditory system and two-formant models. Franco-Swedish Seminar, Grenoble.
10. Fant, G. (1983). Feature analysis of Swedish vowels - a revisit. STL-QPSR, 2-3, 1-19.
11. Fant, G. (1960). Acoustic Theory of Speech Production. 's-Gravenhage.
12. Fant, G. (1985). The vocal tract in your pocket calculator. STL-QPSR, 2-3, 1-19.
13. Fujimura, O. (1967). On the second spectral peak of front vowels: a perceptual study of the role of the second and third formants. Language and Speech, 10, 181-193.
14. Hawkins, S. and Stevens, K.N. (1985). Acoustic and perceptual correlates of the nonnasal-nasal distinction for vowels. Journal of the Acoustical Society of America, 77, 1560-1575.
15. Institut de Phonétique de Grenoble. Labialité et Phonétique, Université des Langues et Lettres de Grenoble, 1980.
16. Klatt, D. (1982). Prediction of perceived phonetic distance from critical-band spectra: a first step. Proc. IEEE, 1278-1281.
17. Mantakas, M., Schwartz, J.L., and Escudier, P. (1986). Modèle de prédiction du "deuxième formant effectif" F'2 - Application à l'étude de la labialité des voyelles avant du français. 15ème JEP.
18. Stevens, K.N. (1985). Spectral prominences and phonetic distinctions in language. Speech Communication, 4, 137-144.
19. Traunmüller, H. (1981). Perceptual dimension of openness in vowels. Journal of the Acoustical Society of America, 69, 1465-1475.
20. Zwicker, E. and Feldtkeller, R. (1967). Das Ohr als Nachrichtenempfänger. Hirzel, Stuttgart.

SOME ASPECTS OF THE SOUND OF SPEECH SOUNDS*

Hartmut Traunmüller
Institutionen för lingvistik, Stockholms universitet
S-106 91 Stockholm, Sweden

PERCEPTUAL QUALITIES

The quality of any sounds, including speech sounds, can be described in terms of associative attributes such as "loud" and "sharp". Listeners are also capable of scaling sound quality in such terms, whereby, e.g., the scales of loudness and sharpness have been established. These quantities - as such - appear, however, to play only a marginal or exceptional role in auditory perception. We are seldom aware of them unless we want to describe a sound that does not mean anything to us. The determination of such psychoacoustic quantities presupposes an intentional activity by the listener. Spontaneously, we hear directly what is going on, e.g. the ball being hit by the racket of a ping-pong player. Among these audibleongoings we can often distinguish different aspects and scale the magnitude of some sense-conveying variables and, in effect, we do also this spontaneously. These variables are, however, for the most part highly specific and complex in relation to a physical description of the acoustic signal.

An important functional distinction among the perceptual aspects of speech sounds is that between "phonetic quality", "personal quality", and "transmittal quality":

"Phonetic quality" characterizes verbal information. On the basis of phonetic quality we can first distinguish among languages (dialects, sociolects) used by a speaker, which gives us a hint about which group of people the speaker belongs to. If we are competent in the language used, phonetic quality will tell us which of the conventional units of language (abstract units like phonemes and morphemes and their concrete forms of realization) the speaker is using. Speech segments can be considered to be identical in phonetic quality whenever phoneticians would transcribe the segments with the same phonetic symbols in a maximally narrow transcription.

"Personal quality" depends mainly on the physiology of individual speech organs and on the state of the speaker's mind. Thus, it tells us something about extra-linguistic features of the speaker,

*The preparation of this paper was supported by a grant from HSFR, the Swedish Council for Research in the Humanities and Social Sciences. I am grateful to BA Richard Schulman for permission to use his fresh data on shouted vowels

e.g. his age, sex, and emotional state (e.g. angry, afraid, cheerful, grieved).

"Transmittal quality", refers to the modifications of the acoustic signal on its way from the speaker to the listener. It may tell us about direction and the distance to the speaker.

Our sense of hearing enables us or even compels us to distinguish these aspects. When listening to natural speech, we have difficulties in estimating abstract variables like pitch and loudness, while this appears to be more feasible when listening to singing, albeit at the expense of accuracy in the functional aspects of speech.

DIMENSIONS OF SPEECH SOUNDS

None of the three functional types of speech sound quality can be completely described on the basis of a single parameter. It appears necessary to consider at least five basic functional dimensions of speech sounds in order to arrive at an understanding of the perceptual processes that make us distinguish the different types of sound quality. Let us consider these functional dimensions and their acoustic and psychoacoustic correlates. The two dimensions referred to in the following as "depth" and "openness" are exclusively phonetic in nature. There is a third dimension, referred to as "elevation", which is relevant to both phonetic quality and personal quality. The remaining dimensions, "size" and "distance", are the principal variables of personal and transmittal quality, respectively. All these dimensions are practically independent of each other and can thus be seen as the orthogonal coordinates of a five-dimensional space. In addition, speech signals contain durational cues, but we are here discussing spectrally cued dimensions only.

Depth: In vowels, this dimension reflects both tongue body retraction, corresponding to the distinctive feature "backness", and lip protrusion, corresponding to "roundness". In consonants articulated in the oral cavity it represents the depth of the front cavity and thereby various "places of articulation". The quantity that we perceive phonetically is, however, not given by this depth as such, but by this depth in relation to the size of the speaker's vocal tract. There is also, as we shall see, some interaction with elevation.

The most prominent audible feature that distinguishes different stationary speech sounds like vowels or fricatives from each other is their characteristic pitch, which used to be studied by investigators who were dependent on their ears as an instrument of sound analysis (Reyher, 1679; Helmholtz, 1863). The type of pitch intended here is especially easy to notice in unvoiced speech sounds, e.g., in whispered vowels. We are going to refer to this type of pitch as "sibilant pitch" in distinction from the unrelated "voice pitch" associated with glottal vibrations. In speech sounds produced with a narrow constriction, sibilant pitch is given by the quarter wave resonance of the front cavity (cf. Kuhn, 1975), i.e., by the depth of this cavity. In the extreme case of whistled speech (Busnel and Classe, 1976) which is in use in several unrelated communities, all the information is carried by a sinusoid whose pitch is at least close to what we mean by "sibilant pitch". When matching sinusoids to the pitch of vowels, Köhler (1910) obtained 3480 Hz for /i/, 2265 for /e/, 1140 for /a/, 470 for /o/, and

225 Hz for /u/. In front vowels sibilant pitch can be assumed to coincide with the pitch of F2', a substitute formant for all formants above F1 (Carlson et al., 1975), while in back vowels it is closer to F1. The salience of sibilant pitch appears to decrease with increasing uniformity in the tonotopical spacing of the formants. The perception of a second formant pitch distinct from sibilant pitch is, however, also possible. Thomas (1969) has shown that subjects can accurately match sinusoids to the second formant of vowels.

Depth affects mainly sibilant pitch and the correlated F2, but also to a small degree F1. The frequency of the first formant is somewhat higher in back vowels and in rounded vowels, as compared with unrounded front vowels of the same phonological degree of openness. In addition, loudness and, in particular, sharpness are also affected. Both decrease with increasing depth. Besides this correlation, sharpness appears to be without use in vowel perception. Among consonants, it appears to be relevant for the distinction between "voiced" and "unvoiced" fricatives. On the basis of sharpness this distinction can also be made in whispered speech.

Openness: Vowels have traditionally been represented in a diagram whose coordinates represent the "openness" of the oral cavity or inversely, the "height" of the highest point of the tongue body, in addition to the dimension that we have called "depth". As noticed by Jung (1926), these coordinates can equally well be interpreted as representing the pitches associated with the two major peaks in vowel spectra (see Figure 1). For a given speaker, the frequency position of the first formant is closely correlated with the depression of the jaw and with the vertical aperture of the labial opening. The jaw is lower in phonetically more open vowels than in closed ones but it is also lower in vowels produced with increased vocal effort. In both cases F1 increases. In addition, F1 is negatively correlated with vocal tract size. The quantity that we perceive phonetically as "openness" is not given by F1 alone, but by its position in relation to the other characteristic frequencies, in particular f0. In closed (high) vowels it may be difficult to perceive a distinct pitch of F1 (Thomas, 1960). Unless we adopt a holistic view (matching the signal against a set of complete spectral templates) we must, however, reckon with some kind of effective pitch of the first formant in order to be able to describe phonetic quality in psychoacoustic terms.

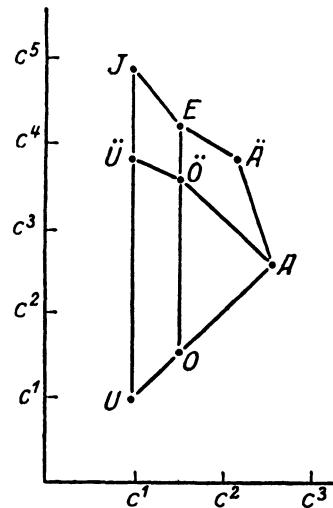


FIGURE 1, Vowels arranged in a diagram according to their audible pitches, indicated in musical notation (from H. Jung, 1926)

Openness affects mainly F_1 and to a small degree f_0 ("intrinsic pitch"). The effect on f_0 appears, however, to pass unnoticed in speech perception. On the basis of a tonotopical scale, the interaction between F_2 and F_1 looks both qualitatively and quantitatively similar to that between F_1 and f_0 if we postulate that the openness dimension should correspond to phonological openness irrespective of backness and roundedness ("depth"). In both cases a characteristic frequency (f_0 or F_1) is slightly increased when the next higher characteristic frequency (F_1 or F_2) is decreased. This suggests that we may have to do with the same underlying phenomenon. It might be that the pitch percepts corresponding to these characteristic frequencies are affected in this way.

If, ceteris paribus, F_1 is increased in a vowel signal, this is perceptually interpreted as an increase in vowel openness. Although loudness increases, perceived speaker distance and vocal effort remain practically unaffected. This fact, by the way, renders the synthesis of speech a simpler task than it otherwise would be.

Elevation: This term refers to an elevation of fundamental frequency, typically concomitant with a lowered position of the jaw and, hence, an elevation of F_1 . In some cases there is also an increased subglottal pressure. This dimension reflects vocal effort as well as emotional, phonemic, and prosodic tone. Increases in tone and in vocal effort are expected to have similar consequences for the pitches involved. Dynamically, the two types of variation may be discriminated by means of the effect on loudness which is higher when due to a change in vocal effort. Phonetically, elevation appears not to be perceived in absolute terms either, but in relation to vocal tract size. Elevation affects all pitches, but this influence decreases with increasing pitch, as can be seen in Figure 2. Elevation is reflected in more or less of an upward compression of the tonotopical pattern of auditory excitation. While the effects of increased vocal effort on f_0 have been investigated in detail (Rostolland, 1982), the effects on formant frequencies have been studied only in a few instances (Schulman, 1985). In whispered vowels, elevation is typically higher than in voiced vowels at conversational vocal effort (see Figure 3). Since elevation affects all pitches, this prosodic dimension is not lost in whispering (cf. Meyer-Eppler, 1957).

If naive listeners are asked to estimate the loudness of vowels produced naturally with varying vocal effort (as by Ladefoged and McKinney, 1963), they are not likely to estimate loudness in the sense in which this term is used in psychoacoustics, but they will estimate the quantity that in fact is subject to variation in the experiment, namely vocal effort. This has led to a terminological controversy, but the experimental results are fairly clear. While loudness N is related to sound pressure p as $N \approx p^{0.46}$ (in the general case, cf. Zwicker and Feldtkeller, 1968), the very different exponent of 1.1 holds for the relation between sound pressure and perceived vocal effort (referred to as "loudness" by Ladefoged and McKinney, 1963). The same exponent had been obtained for the relation between sound pressure and speakers' estimates of their own effort in producing speech sounds (Lane et al., 1961). It appears justified to assume that the subjects in both cases estimated the same variable. This is in line with the reasoning by Lehiste and Peterson (1959). Ladefoged and McKinney (1963) had obtained a closer correlation between subglottal pressure

and vocal effort (exponent 1.9) than between sound pressure and vocal effort. This further emphasizes the perceptual primacy of functional variables and indicates that the estimates were not based on "loudness" alone, but also on the degree of elevation of various pitches.

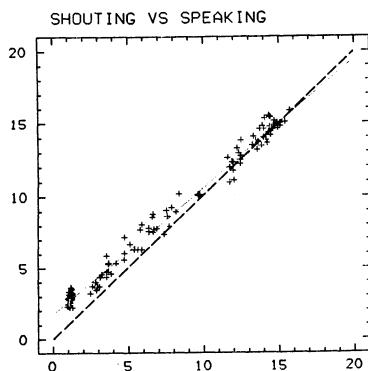


FIGURE 2, Critical band rates (in Bark) of F0, F1, F2, and F3 in ten Swedish shouted vowels (vertically) plotted against those of the same vowels spoken by the same speakers at conversational vocal effort (horizontally). Individual data from three subjects. Diagonal and regression line also shown. Frequency data from R. Schulman (1985, published for only one speaker).

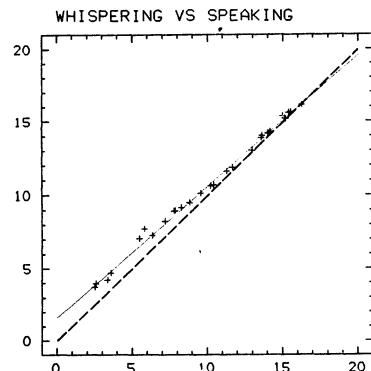


FIGURE 3, Critical band rates (in Bark) of F0, F1, F2, and F3 in five American English whispered vowels (vertically) plotted against those of the voiced versions of the same vowels spoken by the same speakers (horizontally). Mean values obtained from 20 female speakers and from 15 male speakers plotted separately. Diagonal and regression line also shown. Frequency data from K.J. Kallail and F.W. Emanuel (1984 a and b).

Size: The ontogenetic increase in the size of the vocal tract affects all tonotopical cues in a grossly uniform way, e.g., we find in the mean a difference of about 2.5 Bark in voice pitch and in all formant pitches if we compare vowels produced by children four to five years of age with those produced by adult male speakers (see Figure 4). Size differences are, then, reflected in a uniform displacement in the tonotopical dimension of the whole auditory pattern of excitation. In Figure 5, the effect of size on the pitches of F1 and F2 is shown to be large enough to dissociate the "vowel spaces" of men and children. If vowels produced by women are compared with those by men, some vowel-specific (Traunmüller, 1984a) and language-specific (Bladon et al., 1984) deviations from uniformity are found. In part, these deviations can be removed if the tonotopical scale is modified at its low-frequency end (Traunmüller, 1981). Even if that is done, there remains a sex difference implying that female vowels tend to be more peripheral than male vowels in depth vs. openness space. (Traunmüller, in prep.). Sex-differences might be reflected in elevation in addition to vocal tract size.

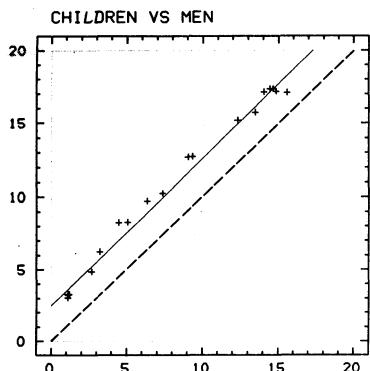


FIGURE 4, Critical band rates (in Bark) of F0, F1, F2, and F3 in five Japanese vowels produced by children aged 4 to 5 years (vertically) plotted against those of adult male speakers (horizontally). Mean values obtained from five speakers in each group. Diagonal and regression line also shown. Frequency data from H. Fujisaki (1970).

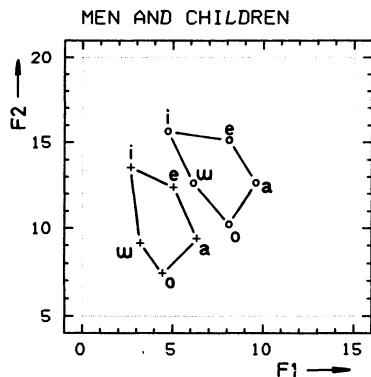


FIGURE 5, Critical band rates (in Bark) of F1 (horizontally) and F2 (vertically) in five Japanese vowels produced by children aged 4 to 5 years (o) and by adult male (+) speakers. Mean values obtained from five speakers in each group. The "vowel spaces" of these speaker groups do not overlap. Data same as in Figure 4.

In vowel-like speech sounds, vocal tract size can be estimated with reasonable precision on the basis of the frequencies of the formants above F3. They are slightly lowered by lip-protrusion, but considerably less so than F2 and F3.

Distance: There is an obvious negative correlation between loudness and the distance from a speaker, and a change in distance affects only loudness, while all the tonotopical cues remain practically unaffected.

If the loudness of a speech signal is experimentally varied in the presence of some other noise that remains unchanged, this is perceptually interpreted as a variation in the distance to the speaker. Other factors remain practically unaffected. If there is no other noise to which the signal can be related, modern man is, however, likely to perceive that somebody is manipulating a volume control knob. Loudness is almost irrelevant to phonetic quality, although it is positively correlated with vocal effort and vowel openness (F1) and negatively with depth. The loudness variations introduced by these factors would demand quite a complex algorithm in order to be removed for an estimation of speaker distance if this were to be done on the basis of isolated speech sounds.

Given connected speech, subjects are capable of estimating the distance to a speaker and even though both distance and vocal effort influence the intensity of the signal, subjects are capable of separating these factors with high accuracy. This is implicit in the results of Wilkens and Bartel (1977), who found that listeners could re-establish

the original intensity of recorded speech with a probable error of just 1 dB.

Table 1 shows the interrelations between the functional dimensions and the psychoacoustic variables. It can be seen that no one of the mentioned psychoacoustic variables is uniquely related to any one of the functional dimensions.

Table 1, Correlations (+ positive, - negative, 0 negligible, * ambivalent) between psychoacoustical variables and functional dimensions (distance, size, elevation, phonetic openness, and phonetic depth) of speech sounds.

	dis	siz	elv	opn	dpt
loudness	-	*	+	(+)	(-)
f0-pitch	0	-	+	(-)	0
F1-pitch	0	-	+	+	(+)
sibilant pitch	0	-	(+)	*	-
higher formants	0	-	0	0	0

PARAMETERS FOR SPEECH SOUNDS

Our challenge consists in describing these functional variables in acoustic or psychoacoustic terms. Which are the acoustic or psychoacoustic quantities that have the same value if any one of the functional variables (e.g., phonetic depth) is constant, no matter how the other functional variables (openness, elevation, size, and distance) are varied? In this particular case, sibilant pitch is clearly essential, but its variation due to variations in elevation and in size has to be removed from the description by normalization with respect to these two variables.

The phoneme boundaries corresponding to different degrees of phonetic depth obtained in an experiment in which subjects identified two-formant vowels with varying f_0 , F_1 , and F_2 could be described adequately by the equation

$$I = (Z_2' - (Z_0 + 3.2)) / (15.3 - (Z_0 + 3.2)), \quad (1)$$

where I is the dimensionless depth quantity and the other two quantities are the critical band rates of F_2' and f_0 (Traunmüller and Lacerda, 1986). This describes to a normalization of Z_2' with respect to the tonotopical distance between two reference points, one located 3.2 Bark above Z_0 (the voice pitch), and the other one located at 15.3 Bark, equal to 2.9 kHz. This is equivalent to a normalization of tonotopical upward compressions concomitant with increases in f_0 , i.e., with variations in elevation. Since two-formant vowels intrinsically lack the fifth variable necessary to distinguish our five functional dimensions, this result suggests that no variation in size was perceived. The reference point at 15.3 Bark might, then, represent a default choice. If we suppose that the position of this reference point in more complete or natural vowels follows the higher formants (F_4 and above), Equation (1) would without further modification also normalize size variations.

Moreover, we should not restrict the normalization process described by Equation (1) to Z_2' , but also apply it to its close relative, sibilant pitch, Z_s , and to formant pitches such as Z_1 . We then obtain the two parameters I_1 , describing phonetic openness and I_s describing phonetic depth as follows:

$$I_1 = (Z_1 - (Z_0 + 3.2)) / (Z_h - (Z_0 + 3.2)) \quad (2)$$

and

$$I_s = (Z_s - (Z_0 + 3.2)) / (Z_h - (Z_0 + 3.2)), \quad (3)$$

where Z_h represents the fifth variable we need for describing our five functional dimensions. Its value defines the hypothetical upper reference point. As for the nature of this fifth variable, we should probably not expect it to represent a definite pitch, but it might be understood as defining a reference point on the basis of a holistic representation of all the spectral peaks corresponding to the upper formants in vowel-like sounds. Since these remain at approximately the same tonotopic locations in a given speaker, it is not necessary for them to be detectable at each instant in time in order to obtain reasonable values of the quantities I_1 and I_s .

Equation (2) may be compared with the previous hypothesis that "openness" is roughly proportional with the tonotopical distance between Z_1 and Z_0 (Traunmüller 1981, 1984a, 1985, Syrdal 1985). The previous hypothesis agrees with Equation (2) for half-open vowels in which this distance is equal to 3.2 Bark. As for closed and open vowels, there is a slight discrepancy, but this is much smaller than the previously observed discrepancies due to the attraction of first formant pitch by single partials that becomes particularly pronounced at $f_0 > 350$ Hz.

Table 2 shows the derived correlates of four of our five functional variables. There is no immediately evident correlate of "distance". The position of Z_h is an inverse measure of "size", i.e. vocal tract length, and the distance between the two reference points is an inverse measure of phonational elevation. In a given speaker, i.e., if "size" is constant, the elevation variable can, of course, be estimated on the basis of Z_0 alone.

Table 2, Variables suggested to be invariant correlates of functional dimensions of speech sounds.

	dis	siz	elv	opn	dpt	
Z_h	0	-	0	0	0	
$Z_h - Z_0$	0	0	-	0	0	
I_1	0	0	0	+	0	(Eq. 2)
I_s	0	0	0	*	-	(Eq. 3)

Equations (2) and (3) are not applicable to unvoiced speech, but for unvoiced speech segments in a voiced context a lower reference point can be obtained by continuation from the voiced segments.

SIMILARITY RATING

There has been a variety of studies concerned with similarity rating of speech sounds (e.g. Plomp, 1970; Carlson and Granström, 1979). In attempts to model the process of speech sound identification it is usually assumed that representations of the sounds received by a listener are compared with a set of stored references representing different phonetic units. This comparison involves the computation of a similarity measure. The sound received is subsequently identified as that phonetic unit whose stored reference is most similar to the representation of the received sound.

In our present approach, the quantities I_1 and I_s can be seen as the coordinate values descriptive of the phonetic quality of vowels in a two-dimensional Euclidean space. Our dissimilarity measure d_{ij} is simply the distance between the points representing the received sound (index i) and each reference (index j) in turn:

$$d_{ij} = [(I_{1i} - I_{1j})^2 + (I_{si} - I_{sj})^2]^{1/2}. \quad (4)$$

This is, however, a simplified approach. Our two coordinates do not provide a complete description of the information present in vowel spectra. Moreover, the dissimilarity measure should be seen as a psychoacoustical quantity. The calculated d_{ij} should be proportional with this psychoacoustical quantity. It is clearly possible to scale "dissimilarity", although such a scaling performed with speech sounds is likely to be influenced by categorical and phonological effects. It might be necessary to weigh I_1 and I_s differently in order to obtain congruence with perceived dissimilarity. It might also develop that our d_{ij} is not a strictly linear measure of perceived dissimilarity, but this would not affect the identification results. More sophisticated approaches involve the consideration of the whole spectrum. Dissimilarity can then be computed as

$$d_{ij} = \left[\sum_{z=0}^{24} |P_{iz} - P_{jz}|^p dz \right]^{1/p}, \quad (5)$$

where z is critical band rate and P is band specific perceptual prominence. If we are interested in phonetic quality as distinct from other qualities, we have to normalize z_i (or z_j) in order to bring the reference points (cf. Eqs. 3 and 4) into agreement. An exponent $p = 1$ would correspond to area differences, while $p = 2$ corresponds to distances in a multidimensional Euclidean space. While the critical band rate (or a similar measure) is generally recognized as scaling the tonotopical dimension adequately, the adequate scale for specific prominence is in question. Plomp (1970) used band specific sound pressure levels and found an exponent $p = 1$ to be appropriate for several types of sounds, but for vowel sounds the computation gave an equally good fit to the data with higher exponents. Bladon and Lindblom (1981) used loudness density. Carlson and Granström (1979) tried both alternatives (with $p = 1$ and $p = 2$) and did not find any advantage in using the supposedly more sophisticated alternative of loudness density.

It is, of course, crucial to use the correct measure of prominence if we want to calculate perceived dissimilarity. On inspection of Equation (5) we can see that the calculated d_{ij} is proportional to P_i and P_j if both are multiplied by the same factor. If we substitute loudness density for these P -s then the calculated d_{ij} will follow a power function of sound pressure, just as loudness does. The perceived dissimilarity between, say, an [i] and an equally loud [y] is, however, practically independent of loudness. Perceived dissimilarity can evidently not be computed in this way. What about substituting specific SPL or loudness level for P in Equation (5)? On the basis of such a logarithmic measure, the computed dissimilarity between an [i] and an equally loud [y] would come out to be roughly independent of loudness. Both alternatives lead, however, to invalid results if we try to compute the dissimilarity between an [i] and an [y] whose loudness is different from that of the [i]. Both alternatives overestimate drastically the contribution of total loudness to the perceived dissimilarity.

The results of some investigations (Terhardt, 1968; Suchowersky, 1977) support the assumption that perceived dissimilarity between any auditory stimuli A and B is proportional to the number of steps of just noticeably different stimuli that can be inserted between A and B on the shortest possible route, or on the route suggested by the set of stimuli used in an experiment (cf. Traunmüller, 1984b), and we have reason to believe that this is true for any kind of stimuli. Thus all jnd-steps (Empfindungsstufen) appear to be equivalent in dissimilarity ratings. This insight can be traced back to Fechner who, on the basis of insufficient evidence, ascribed this equivalence of jnd-steps also to sensational magnitude ratings. Later, Stevens demonstrated, on the basis of ample evidence, that jnd-steps are not equivalent in magnitude ratings. Unfortunately, Stevens also denied the equivalence of jnd-steps in dissimilarity ratings (Stevens, 1965), apparently by assuming proportionality between dissimilarity and difference in magnitude. But isn't 10 sone just as similar to 5 sone, as 2 sone is to 1 sone - as predicted by Terhardt's law of equivalence, given that Webers law is valid?

There is reason to believe that the measure of P in Equation (5) should scale prominence in terms of jnd-steps. There is, however, no established method by which such a representation of sounds could be derived - except if we are satisfied with a very rough quantization of prominence: any parts of a spectrum which are below the threshold of masking, e.g. partials that can be removed without evoking an audible difference, have $P < 1$ jnd, and those above this threshold have $P > 1$ jnd. I would see it as an urgent task to develop such a method or else find an alternative way of computing dissimilarity.

GESTALT QUALITY

There are several analogies between auditory perception of phonetic quality and visual perception of shape, e.g. of graphic symbols. Perception of shape is guided by application of an inverse projective geometry by which the effects of distance, tilt (with respect to a plane orthogonal to the line of sight), and orientation within a plane are neutralized. Similar processes appear to be effective in auditory perception of phonetic quality: the effects of variations in elevation on the tonotopical pattern of auditory excitation are

analogous to those of variations in tilt. In both cases, a spatial (or tonotopic) dimension is more or less compressed. The effects of variation of speaker size are analogous to those of varying the orientation of a graphic symbol within a plane orthogonal to the line of sight. In both cases the relative locations of any details remain unaffected.

Some further analogies are best seen from the point of view proposed by gestalt theory. Gestalt qualities are said to be characterized, i.a., by their transposability. The possibility of displacing the whole pattern of excitation uniformly along the tonotopic scale without affecting phonetic quality - this corresponds to variations in speaker size - can be seen as an instance of such transposability. Of particular interest is the transposability onto different carrier structures: a spectral envelope characteristic of a certain vowel quality may be carried by complex harmonic tones or by noise, e.g. in whispering. If the envelope is carried by complex harmonic tones, a large range of wider or narrower spacings between the partials is acceptable. The perception of the gross spectral envelope as a phonetic gestalt does, however, not preclude that finer detail, e.g. partials, can simultaneously be perceptually resolved. This is required e.g., if continuity of the carrier structure is to serve as a means for tracking the voice of one speaker among other speakers or other sources of sound (Brokx and Nooteboom, 1982).

Some experimental results (Chistovich and Lublinskaya, 1981) suggest that tonotopical integration - or an equivalent process of envelope estimation, with a bandwidth of roughly 3 Bark, is effective in perception of phonetic quality. Although the general validity of this hypothesis has been questioned (Traunmüller, 1982), there are other results in favour of it. In identifications of one-formant vowels (Traunmüller, 1981) it was observed that a first formant apparently can be localized in the middle between partials as long as the distance between these is less than 3 Bark (This is always the case at $f_0 < 350$ Hz). It is also likely that sibilant pitch is perceived on the basis of such an integration. The tonotopical distance between the formants including and above F3 is less than 3.5 Bark in nearly all vowels. In most front vowels F2 is also closer than 3.5 Bark to F3. In rounded back vowels F2 is closer than 3.5 Bark to F1 (cf. Bladon, 1983; Syrdal, 1985). Given such a broad-band integration, the gross spectral envelopes of most vowel sounds contain either two prominent peaks (front vowels) or just one (back vowels), in agreement with the traditional auditory description of vowels shown in Figure 1. If this is so, we might regard the single formant peaks in the spectra of speech sounds just as a carrier structure for the gross spectral envelope that is ultimately decisive for the perception of phonetic quality. Thus, the carrier structure is not given by the voice source signal exciting the vocal tract, but by the source signal modified by the resonances of a vocal tract in position for the production of a neutral vowel. Such a vowel has apparently no definite sibilant pitch. When producing other speech sounds, the neutral shape of the vocal tract is perturbed and causes the formants carrying the signal to be modulated in frequency and - concomitantly - in amplitude, thereby giving rise to the "super-formant-structure" characteristic of the particular sounds of speech.

REFERENCES

1. Bladon, A. (1983). Two-formant models of vowel perception: shortcomings and enhancements. *Speech Comm.*, 2, 305-313.
2. Bladon, R.A.W., Henton, C.G., and Pickering, J.B. (1984). Outline of an auditory theory of speaker normalization. In: Proceedings of the Tenth International Congress of Phonetic Sciences, 313-317 Foris Publ. Dordrecht.
3. Bladon, R.A.W. and Lindblom, B. (1981). Modelling the judgement of vowel quality differences. *J. Acoust. Soc. Am.*, 69, 1414-1422.
4. Brokx, J.P.L. and Nooteboom, S.G. (1982). Intonation and the perceptual separation of simultaneous voices. *J. Phonetics*, 10, 23-26.
5. Busnel, R.G. and Classe, A. (1976). Whistled Languages. Communication and Cybernetics, Vol. 13, Springer Verlag, Berlin, Heidelberg.
6. Carlson, R., Fant, G., and Granström, B. (1975). Two-formant models, pitch and vowel perception. In: G. Fant and M.A.A. Tatham (Eds.), Auditory Analysis and Perception of Speech, 55-82. Academic, London.
7. Carlson, R. and Granström, B. (1979). Model predictions of vowel dissimilarity, STL-QPSR, 3-4/1979, 84-104. (Royal inst. technol. Stockholm).
8. Chistovich, L.A. and Lublinskaya, V.V. (1981). The "centre of gravity" effect in vowel spectra and critical distance between the formants: psychoacoustical study of the perception of vowel-like stimuli. *Hearing Res.*, 1, 185-195.
9. Fujisaki, H. (1970). Formant frequencies of sustained vowels in Japanese obtained by analysis-by-synthesis of spectral envelopes, personal communication, Univ. of Tokio.
10. Jung, H. (1926). Die neueren Vokaltheorien. *Physik. Zeitschr.*, 27, 716-723.
11. Helmholtz von, H.L.F. (1863). Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik, F. Vieweg und Sohn, Braunschweig.
12. Kallail, K.J. and Emanuel, F.W. (1984a). Formant-frequency differences between isolated whispered and phonated vowel samples produced by adult female subjects. *J. Speech Hearing Res.*, 27, 245- 251.
13. Kallail, K.J. and Emanuel, F.W. (1984b). An acoustic comparison of isolated whispered and phonated vowel samples produced by adult male subjects. *J. Phonet.*, 12, 175-186.
14. Köhler, W. (1910). Akustische Untersuchungen II. *Z. Psychol.*, 58, 59- 140; quoted from E.G. Boring, Sensation and Perception in the History of Experimental Psychology, Appleton-Century, New York 1942, p. 370.
15. Kuhn, G.M. (1975). On the front cavity resonance and its possible role in speech perception. *J. Acoust. Soc. Am.*, 58, 428-433.
16. Ladefoged, P. and McKinney, N. (1963). Loudness, sound pressure and sub-glottal pressure in speech. *J. Acoust. Soc. Am.*, 35, 454-460.
17. Lane, H.L., Catania, A.C., and Stevens, S.S. (1961). Voice level: autophonetic scale, perceived loudness, and effects of sidetone. *J. Acoust. Soc. Am.*, 33, 160-167.
18. Lehiste, I. and Peterson, G.E. (1959). Vowel amplitude and phonemic stress in American English. *J. Acoust. Soc. Am.*, 31, 428-435.

19. Meyer-Eppler, W. (1957). Realization of prosodic features in whispered speech. *J. Acoust. Soc. Am.*, 29, 104-106.
20. Plomp, R. (1970). Timbre as a multidimensional attribute of complex tones. In: R. Plomp and G. Smoorenburg (Eds.), Frequency Analysis and Periodicity Detection in Hearing, Sijthoff, Leiden, 397-414.
21. Rostoland, D. (1982). Acoustic features of shouted voice. *Acustica*, 50, 118-125.
22. Reyher, S. (1679). Mathesis mosaica, quoted from C. Stumpf, Die Sprachlaute, 6, Kap., 142-166. J. Springer, Berlin 1926.
23. Schulman, R. (1985). Articulatory targeting and perceptual constancy of loud speech. In: PERILUS 4 (Inst. linguist., Univ. Stockholm), 86-91.
24. Stevens, S.S. (1965). Matching functions between loudness and ten other continua. Percept. Psychophys., 1, 5-8.
25. Suchowersky, W. (1977). Beurteilung von Unterschieden zwischen aufeinanderfolgenden Schallen. *Acustica*, 38, 131-139.
26. Syrdal, A.K. (1985). Aspects of a model of the auditory representation of American English vowels. *Speech Comm.*, 4, 121-135.
27. Terhardt, E. (1968). über ein Äquivalenzgesetz für Intervalle akustischer Empfindungen. *Kybernetik*, 5, 127-133.
28. Thomas, I.B., (1969). Perceived pitch of whispered vowels. *J. Acoust. Soc. Am.*, 46, 468-470.
29. Traunmüller, H. (1981). Perceptual dimension of openness in vowels. *J. Acoust. Soc. Am.*, 69, 1465-1475.
30. Traunmüller, H. (1982). Perception of timbre: evidence for spectral resolution bandwidth different from critical band? In: R. Carlson and B. Granström (Eds.) The representation of Speech in the Peripheral Auditory System. Elsevier Biomed, Amsterdam, pp. 103-108.
31. Traunmüller, H. (1984a). Articulatory and perceptual factors controlling the age- and sex-conditioned variability in formant frequencies of vowels. *Speech Comm.*, 3, 49-61.
32. Traunmüller, H. (1984b). Die spektrale Auflösung bei der Wahrnehmung der Klangfarbe von Vokalen. *Acustica*, 54, 237-246.
33. Traunmüller, H. (1985). The role of the fundamental and the higher formants in the perception of speaker size, vocal effort, and vowel openness. In: PERLIUS 4, 92-102 (Inst. linguist. Univ. Stockholm).
34. Traunmüller, H. and Lacerda, F. (1986). Perceptual relativity in identification of two-formant vowels. Submitted to *Speech Comm.*
35. Traunmüller, H. (in preparation). Inherent normalization of vowel formant frequencies,
36. Wilkens, H. and Bartel, H.H. (1977). Wiedererkennbarkeit der Originallautstärke eines Sprechers bei elektroakustischer Wiedergabe, *Acustica*, 37, 45-49.
37. Zwicker, E. and Feldtkeller, R. (1967). Das Ohr als Nachrichtenempfänger, 2 Aufl., Hirzel Verlag, Stuttgart, p. 137.

INVOLVEMENT OF THE CRITICAL BAND IN IDENTIFICATION, PERCEIVED DISTANCE, AND DISCRIMINATION OF VOWELS*

B. Espinoza-Varas

Speech and Hearing Sciences, Indiana University, Bloomington,
Indiana, 47405, USA.

INTRODUCTION

Psychoacoustic models of vowel perception typically assume a critical-bandwidth filter, CBF, at the first stage of auditory frequency analysis (e.g., Zwicker et al., 1979; Searle et al., 1979). In spite of the apparent consensus about the validity of the CBF assumption, little perceptual evidence relating CBF to vowel perception is available. The most compelling perceptual evidence appears to be the high correlation observed between perceptual vowel spaces and physical vowel spaces obtained from the output of a bank of 1/3 oct filters (e.g., Pols et al., 1969). This perceptual evidence is not only indirect, but in addition is limited to one aspect of vowel perception: similarity judgments. The present study attempts to examine more directly and in greater detail the contribution of CBF to vowel perception. To obtain more direct evidence, listeners were required to judge changes in vowel formant frequency that were equal in terms of critical bandwidth units. The perceptual consequences of such acoustic changes were examined at the level of identification, ratings of phonetic goodness, and pair-wise discrimination of vowels. The study employed two different back-front place of articulation vowel continua. These continua were presented to normal listeners and to sensorineural hearing impaired listeners.

I. METHOD

a. Stimuli

The stimuli were three-formant, steady-state vowels, 400 ms in duration, generated on a digital synthesizer (Klatt, 1980). Two separate back-front vowel continua were generated, ranging either from /u/ to /i/, or from /ɔ/ to /ɛ/. Acoustic evidence (Peterson and Barney, 1952) indicates that the major cue in these back-front vowel distinctions is the frequency of F2, while the values of F1 (and to a lesser extent of F3) for the back vowel are roughly the same as those for the front vowel. Since our interest was to examine the perceptual effects of varying F2 alone in exact bark steps, the frequency of F1 and F3 was made equal for all the vowels within a back-front continuum. The values of F1 (F3) were 290 (2630) Hz, and 550 (2450) Hz for the stimuli of the /i-u/ and /ɔ-ɛ/ continua. Within each back-front vowel pair, F2 ranged from the value appropriate to

*Work performed at the Psychology Dept., Univ. of Calgary, Alberta, Canada, funded by AHFMR Fellowship. Preparation of this manuscript was supported by AFOSR and NIH Grants to Indiana University.

the back vowel (870 Hz for /u/ and 840 Hz for /ɔ/) to that appropriate to the front vowel (2290 Hz for /i/ and 1840 Hz for /ɛ/) in 6-10 steps of 0.5 critical bandwidth (Δf) each. Fundamental frequency was 120 Hz in all cases. Both the pitch and the amplitude of the glottal excitation were constant throughout the duration of a given vowel, except in the last 30 ms of the signal where the glottal-excitation amplitude decayed linearly to zero to prevent transients. All synthesis was performed with the resonators connected in parallel.

The digital stimuli were output through a D/A converter, low-pass filtered at 4.8 kHz using an analog elliptic filter, and tape recorded (Revox B-710) in randomized order. They were played back with a Revox B-710 tape recorder, amplified (Amcrom D-75), and delivered monaurally (TDH-49-P headphones) at 65 dB(A) SPL, inside of an IAC sound-attenuating booth.

b. Task and procedure

A single-interval paradigm was employed (Braida and Durlach, 1972) which makes it possible to measure three perceptual parameters within the same task. On each trial, subjects were asked first to identify the stimulus as one of the members of the vowel pair being tested, and subsequently to rate the "phonetic goodness" of the stimuli using a seven-point bidirectional scale with origin centered at rating "4". Ratings "1", and "7" represented the best exemplars of the front and back vowel, respectively. All subjects became fully proficient in the task after 10-15 min practice. The stimuli of each continuum were presented in blocks of 21-33 trials, at intervals of 0.5 s. The subjects responded on a computer keyboard. Approximately 900 trials were devoted to each continuum, yielding approximately 64 trials for each stimulus of each continuum. This trial density allowed a fairly stable estimation of the d' index. Between 2-4 experimental sessions were spent on each continuum. The experiment was controlled by a microcomputer system.

c. Data analysis

Three performance indexes were extracted from the data: Identification, average goodness rating, and discrimination. Identification functions relate the percentage of "back" vowel responses (i.e., /u/, or /ɔ/) to the frequency of F2, expressed either in Hz or in bark units. This measure informs about the definition of the perceptual classes and their boundaries. The average rating given to a stimulus plotted as a function of the frequency of F2 provides an indication of the relative perceptual distance or apparent similarity between any two stimuli. The discrimination score, d' , provides an estimate of the pairwise discriminability between two successive stimuli of the continuum. The d' scores were inferred from the ratings using the method described by Braida and Durlach (1972).

II. RESULTS

a. Normal-hearing listeners

The subjects for these experiments were college students, with normal hearing sensitivity. Four subjects were tested with the /u-i/ continuum, and three subjects with the /ɔ-e/ continuum.

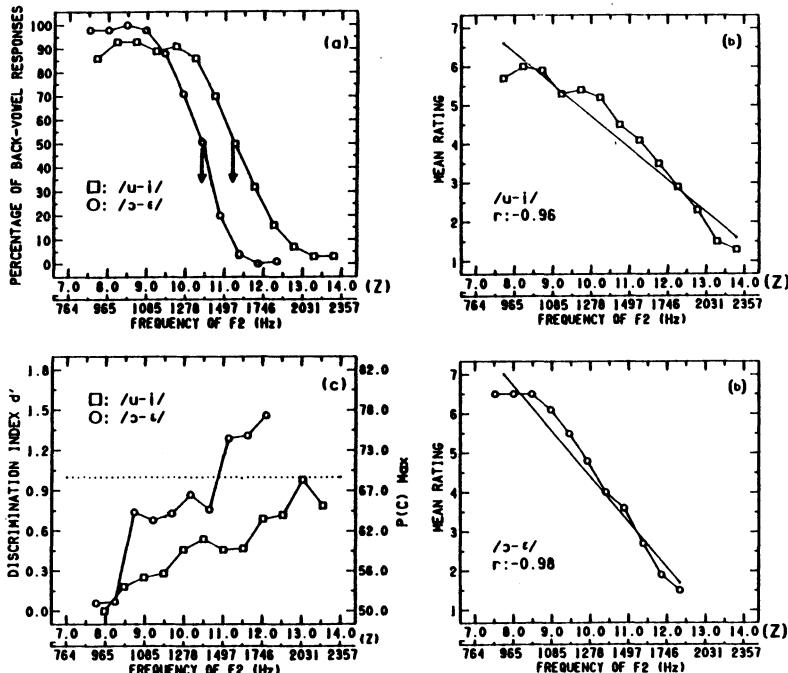


FIGURE 1, Panel (a): Identification functions relating the percentage of back-vowel responses to the frequency of F2 in units of bark (Z) and in Hz, for two vowel continua. Arrows show the values of the phonetic boundaries. Data from normal hearing listeners. Panels (b): Phonetic goodness functions relating the mean rating assigned to each stimulus, to the frequency of F2. The values of the Pearson correlations between the data and the linear fits are shown on each panel. Panel (c): Discrimination functions relating the value of d' and of $P(c)$ max to the frequency of F2, for two vowel continua. Dotted horizontal shows performance level of $d'=1.0$ or 69 percent correct. Data from normal hearing listeners.

The results were consistent across vowel continua and listeners; thus, the average performance of each group is described below. Figure 1a displays the identification functions for the two vowel continua. The abscissa of this and following figures is scaled in bark units (frequencies corresponding to each bark value are also indicated). The identification functions exhibit asymptotic regions approaching 0% and 100%, and are steep around the phonetic boundary; that is, there

was good definition of the phonetic classes. It is apparent in these functions that vowel identification is not directly related to the bark scale. Changes in F2 that are equal along this scale do not yield uniform changes in identification. In addition, the identification functions show that the frequency of F2 for the best exemplars of the back vowels (i.e., best /u/ or /ɔ/) must be changed by approximately 4.0-5.0 critical bands to obtain 100 percent front-vowel responses (i.e., /i/ or /ɛ/). Changes by one bark in the frequency of F2 of the best exemplars would be too small to cause a change in identification. In terms of bark units, the boundary between the back and front vowel falls at F2 values corresponding to 11.4 and 10.4 bark for /u-i/ and /ɔ-ɛ/. The F3-F2 distance is considered a critical parameter in front-back vowel distinctions (Syrdal, 1986; Chistovich and Lublinskaya, 1979). In our data, the phonetic boundary fell at F3-F2 distances corresponding to 3.4 and 3.9 bark for /u-i/ and /ɔ-ɛ/. There is thus fairly good agreement with the value of 3.0-3.5 bark reported by the previous authors. In this respect our results seem to agree with the proposition that vowel identification involves a very broad integration mechanism, which would combine the outputs of several critical bands (Syrdal, 1986; Chistovich and Lublinskaya, 1979).

Figure 1b displays the results of ratings of "phonetic" goodness. The rating functions show a nearly linear relation to the bark scale. The straight lines shown in the figure are the best linear fits to the data points, with Pearson correlations of -0.96 and -0.98. In general, it appears that changes in F2 that are equal along the bark scale produce equal changes in the judged phonetic goodness. This relation is quite robust and consistent across listeners and vowel continua.

The pairwise discrimination results are displayed in Figure 1c. The discrimination function shows more variability than the identification and the rating functions. Nevertheless, a very definite pattern is observed: discriminability is minimal at the "back" vowel endpoint (/u/ or /ɔ/) and maximal at the "front" vowel endpoint; thus the pairwise discrimination tends to follow the simple frequency difference between the steps. In agreement with Mermelstein (1978), this result suggests that discrimination is not limited by the critical band. If discrimination were limited by the critical band analysis, all F2 steps along the continuum should exhibit similar discriminability. In addition, all pairs of contiguous stimuli along the continuum differed by half a critical band; however, a number of them were discriminated at about the threshold performance level of $d'=1.0$. This means that vowel discrimination relies on a level of frequency analysis sharper than the critical bandwidth.

b. Hearing impaired listeners

To further examine the generality of the previous findings, an effort was made to study the effects of broadening the auditory filter bandwidth on the perception of half-a-bark changes in F2. Thus, the experiments were replicated with sensorineural hearing impaired listeners (SHIL), who are known to exhibit broadened critical bandwidths (e.g., Florentine et al., 1980). One prediction is that with an abnormally wide filter, the variation in phonetic goodness with changes in F2 would be less definite, because the unit of perceptual measure (i.e., the critical bandwidth) is altered. Another prediction is that such broadening would differentially affect each of the three

perceptual parameters, because of the differential involvement of the CBF.

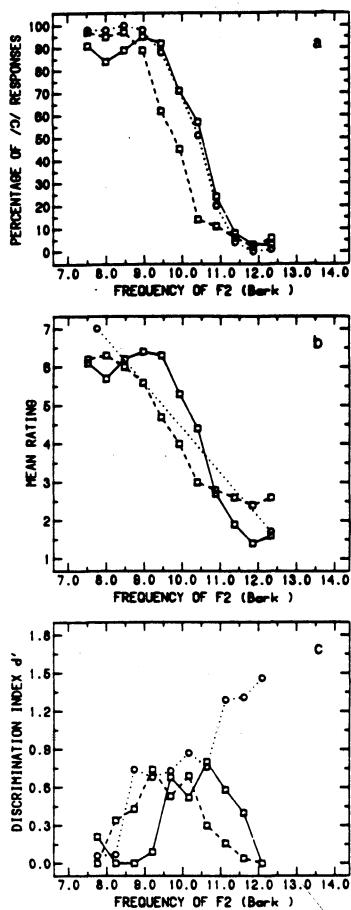


FIGURE 2. Same as figure 1, but data are from two sensorineural hearing impaired listeners (functions in solid and dashed lines). Function in dotted lines show corresponding data for normal listeners.

The same method and procedure used with normal listeners was employed. Two sensorineural hearing impaired listeners participated. In the test ear, these listeners had a moderate loss (10-15 dB) at 250 Hz and a gradual slope (about -8 dB/oct on the average) towards the higher frequencies. Thus, in addition to degraded frequency resolution, these listeners had a loss in sensitivity which increased with frequency. Besides the abnormal encoding of frequency, the representation of the relative amplitude of the formants (or spectral envelope) was also altered. For example, the second formant of the best /ɛ/ stimulus falls in an audiometric region with considerably poorer sensitivity than that of the region excited by the second formant of the best /ɔ/ stimulus. In other experiments (not reported here) steps were taken to disentangle these two factors, but they are confounded in the present study. The stimuli were presented at overall SPL's ranging from 70-80 dB.

Results obtained with the /ɔ - ε/ vowel continuum are described below. The identification functions for this continuum are shown in Fig. 2a (for ease of comparison, in all panels of Fig. 2, the data for the normal listeners are shown by means of dotted functions). Functions with solid and dashed lines describe data for two impaired listeners. For these two listeners, identification is essentially identical to that of the normal listeners, though in one listener the phonetic boundary is shifted by about 0.5 bark.

Fig. 2b shows the results of the ratings of phonetic goodness. The rating functions of impaired listeners appear to depart significantly from the linear trend obtained with normal listeners. The functions show definite asymptotes at the extremes; as a result, for most of the continuum, constant changes in F2 no longer produce constant changes in phonetic goodness. In fact, the rating functions of these listeners resemble quite closely their identification functions. This means that the subjects can categorize the stimuli, but are to a good extent unable to rate the similarity of a given stimulus to the best exemplars of the vowels. The flattening of the rating function is seen at both ends of the F2 continuum. This may suggest that the effect is not well correlated with a frequency dependent hearing loss.

Fig. 2c shows the discrimination results. Hearing impaired listeners exhibit severely depressed discrimination towards the /ɛ/ end of the continuum. This effect is clearly correlated with frequency dependent hearing loss; thus, it could have resulted simply because as F2 is raised towards the value of the front vowels, more of its energy falls into progressively worse audiometric regions. It is interesting to note that deficits in the ability to make judgments of phonetic goodness and to discriminate the vowels coexist with nearly normal identification performance.

III. CONCLUSIONS

1. The results obtained with normal listeners indicate that in front-back vowel distinctions cued by variations of F2, the critical bandwidth analysis is reflected more directly in the judgment of the phonetic goodness of vowels than in identification or discrimination judgments. The phonetic goodness of a given stimulus appears to be linearly related to the distance (in critical bandwidth units) between its F2 value and that of the prototypical F2 value (as defined acoustically). In other words, critical bandwidth filtering seems to be directly involved in the judgment of the apparent similarity between two vowel spectra. This is in agreement with previous studies using vowel similarity tasks (Pols et al., 1969; Bladon and Lindblom, 1981). Vowel identification and discrimination reflect the critical bandwidth analysis only indirectly. For example, identification could reflect the joint activity of several critical bands, whereas discrimination could reflect more directly the operation of a filtering process sharper than the critical bandwidth, such as lateral inhibition (Houtgast, 1974, Karnickaya et al., 1975).

The above results are consistent with the proposition that, at one level of representation, vowels are encoded in terms of the spatial pattern of excitation maxima on the basilar membrane (Potter and Steinberg, 1950; Syrdal, 1986). Specifically, the distances between the maxima, in critical band units, would encode the identity of a vowel. There would be specific intermaximum distances that would represent prototype vowels. The apparent similarity (or phonetic goodness) between an allophone and its prototype vowel would be proportional to the similarity between the corresponding intermaximum distances.

2. The experiments with sensorineural hearing impaired listeners with sloping audiograms do not allow a rigorous examination of the CBF assumption, with the method employed in this study. In addition to the broadening of the auditory filter, they exhibit a deterioration in

sensitivity that increases with frequency. With our method, it is not possible to determine to what extent the effects observed are due to broadening of the auditory filter, or to reduced sensitivity.

3. Keeping the above restriction in mind, the following can be proposed. Vowel identification in SHIL is very similar to that of normal listeners. However, they exhibit degraded ability to rate the apparent similarity between two different vowel spectra that belong to the same vowel class. This deficit is not correlated with the amount of hearing loss, and thus it could be interpreted as being caused by the broadening of the critical bandwidth (or equivalently, the broadening of the perceptual unit). If, as indicated above, the apparent similarity between the spectra is scaled in units of critical bandwidth (or bark), broadening of the filter would effectively increase the similarity of the vowels. On these grounds, it could be concluded that the results of SHIL's support the proposition of a close relation between the CBF and the goodness ratings.

4. Hearing impaired listeners also show degraded ability to perform pairwise discrimination between vowels, particularly towards the higher F2 frequencies. This effect seems to be correlated with the amount of hearing loss.

5. An interesting implication of these results is that, perceptually, vowels seem to possess multiple representations. These different representations can be distinguished in terms of their sensitivity to stimulus change, and in terms of the way in which they are affected by hearing loss. At the level of identification or phonetic labelling, the representation appears to be coarse, and is sensitive only to changes in the formant frequencies involving several critical bands. At the level of phonetic goodness judgments, the representation is more detailed and seems to preserve critical-band frequency detail. At the level of discrimination, the perceptual equivalent of the simple frequency difference between the F2 steps seems to be preserved. A multiple representation is suggested also by the fact that hearing impairment seems to affect differentially each of the three perceptual parameters measured. As a result, normal vowel identification can coexist with abnormalities in the ability to rate the phonetic goodness and to discriminate the vowels. Additional experiments are needed to fully substantiate this latter point.

REFERENCES

1. Braida, L.D. and Durlach, N.I. (1972). Intensity Perception. II. Resolution in one-interval paradigms. Journal of the Acoustical Society of America, 51, 483-502.
2. Chistovich, L.A. and Lublinskaya, V.V. (1979). The 'center of gravity' effect in vowel spectra and critical distance between the formants: psychoacoustical study of the perception of vowel-like stimuli. Hearing Research, 1, 185-195.
3. Florentine, M., Buus, S., Scharf, B., and Zwicker, E. (1970). Frequency selectivity in normally-hearing and hearing-impaired listeners. Journal of Speech and Hearing Research, 23, 643-669.
4. Houtgast, T. (1974). Auditory analysis of vowel-like sounds. Acustica, 31, 320-324.

5. Karnickaya, E.G., Mushnikov, V.N., Slepokurova, N.A., and Zhukov, S. Ja. (1975). Auditory processing of steady state vowels. In: G. Fant and M.A.A. Tatham (Eds), Auditory Analysis and Perception of Speech. Academic Press, London.
6. Klatt, D.H. (1980). Software for a cascade/parallel formant synthesizer. Journal of the Acoustical Society of America, 67, 971- 995.
7. Mermelstein, P. (1978). Difference limens for formant frequencies of steady-state and consonant-bound vowels, Journal of the Acoustical Society of America, 63, 572-580.
8. Peterson, G.A. and Barney, H.L. (1952). Control methods used in a study of the vowels. Journal of the Acoustical Society of America, 24, 175-184.
9. Pols, L.C.W., van der Kamp, L.K.Th., and Plomp, R. (1969). Perceptual and physical space of vowel sounds: Journal of the Acoustical Society of America, 46, 458-467.
10. Potter, R.J. and Steinberg, J.C. (1950). Toward the specification of speech. Journal of the Acoustical Society of America, 22, 807-820.
11. Searle, C.L., Jacobson, J.Z., and Rayment, S.G. (1979). Stop consonant discrimination based on human audition. Journal of the Acoustical Society of America, 65, 799-809.
12. Sharf, B. (1970). Critical Bands. In: J.V. Tobias (Ed.), Foundations of Modern Auditory Theory. Academic Press, New York, London.
13. Syrdal, A.K. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. Journal of the Acoustical Society of America, 79, 1086-1100.
14. Torgerson, W.S. (1958). Theory and Methods of Scaling, (New York, Wiley).
15. Zwicker, E., Terhardt, E., and Paulus, E. (1979). Automatic speech recognition using psychoacoustic models. Journal of the Acoustical Society of America, 65, 487-498.

PROFILE ANALYSIS AND SPEECH PERCEPTION*

David M. Green and Leslie R. Bernstein

Psychology Department, University of Florida, Gainesville, Florida
32611, USA

INTRODUCTION

There is, unfortunately, a wide gulf between research in psychoacoustics and research on speech perception. These differences arise, in part, because of the different objectives of the investigators. Understanding how the auditory system functions and understanding the speech code are different and distinct goals. But there are some areas and topics where one might expect a commonality of interest. The topics of auditory perception and the limits of certain basic auditory discrimination processes are both areas that should enjoy mutual interest and concern. But, even here, wide differences are apparent in the way these topics are approached by the speech scientist and by the psychoacoustican. These differences are especially evident in the choice of stimulus materials. The psychoacoustic stimuli are simple; the speech stimuli are complex. The just-noticeable-difference in the frequency or the amplitude of an isolated pure tone appears to have little to do with how we recognize differences between vowels or broadband consonants.

The simplicity of psychoacoustic stimuli is understandable, given the considerable emphasis placed by that discipline on the control of stimulus intensity. Psychoacoustic stimuli are presented at specific sound pressure levels, and considerable time and effort are devoted to ensuring that these levels fall within some small tolerance. A typical limit is some fraction of a decibel, since the Weber fraction for intensity of a single sinusoidal stimulus is about 1 dB. The absolute sound level of speech, on the other hand, is seldom a variable of much concern. Obviously, the sound must be intense enough to ensure that the listener can hear the utterance. But that condition can be met over a large intensity range, and 10 or 20 dB differences between presentation levels may well be regarded as secondary. The reason for such broad limits is simple to explain: the speech code involves a change in spectral composition over time and seldom depends on an absolute intensity level. Relative intensity levels at different regions of the spectrum, the definition of peaks and valleys in the spectrum, and the frequency region where the energy is present are thought to be the most important aspects of the speech code. Indeed, intensity

*The research was supported, in part, by a grant from the National Institute of Health and the Air Force Office of Scientific Research. Our thanks to Dr. Virginia M. Richards whose extensive comments on an earlier draft considerably improved this one.

level per se is generally not part of the speech code; rather, it is used to accent or embellish the utterance.

The preceding observations provide sufficient background for why we find it interesting to study the ability of the human observer to discriminate changes in the shape of the spectrum of a complex auditory stimulus. Such studies, we hope, will provide us with basic information as to how the auditory sense operates and will begin to contribute to our understanding of speech perception, which, after all, is the primary function of the auditory process. In order to understand our research on the discrimination of changes in spectral shape and, in particular, how it differs from the previous studies of intensity discrimination, we must first consider in some detail the intensity discrimination task.

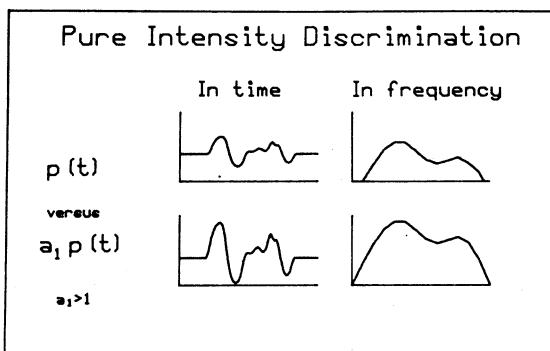


FIGURE 1. Pure intensity discrimination in which the observer discriminates a standard stimulus [$p(t)$] from a scaled version [$a_1 * p(t)$]. In the frequency domain (right side of the figure), the effect of scaling is simply to displace the spectrum along the ordinate. The temporal waveforms (left side) are identical except for the scaling factor.

Let us first consider the simplest intensity discrimination task, what we might call "pure" intensity discrimination. The two sounds used in the discrimination task are either one pressure wave, $p(t)$, or a scaled version of that same wave, $a * p(t)$, where the constant, a , is not unity. In the frequency domain, the two spectra are simply displaced from one another along the ordinate, assuming we have plotted the spectra on a logarithmic intensity scale, such as decibels. The discrimination problem is to select between the two spectra. Pure intensity discrimination, such as that illustrated in Figure 1, may be contrasted with a different task, that of discriminating a change in spectral shape, what we call "profile analysis". The stimuli to be discriminated in this task are illustrated in Figure 2. The two pressure waves, $p_1(t)$ and $p_2(t)$, may be completely unrelated. Since the waveshapes are different, the spectra of the two sounds will also be different, as illustrated in the right hand portion of Figure 2. Although the shapes of the spectra differ, the listener might use differences in intensity at a particular frequency region in order to achieve the discrimination between the two stimuli. Unless some

special precautions are taken, there is nothing to prevent the listener from discriminating a change in spectral shape on the basis of some difference in intensity at some particular frequency region. Thus, the experimenter could not, in general, guarantee that the observer's performance in discriminating a change in spectral shape is in any way different from discriminating a change in intensity.

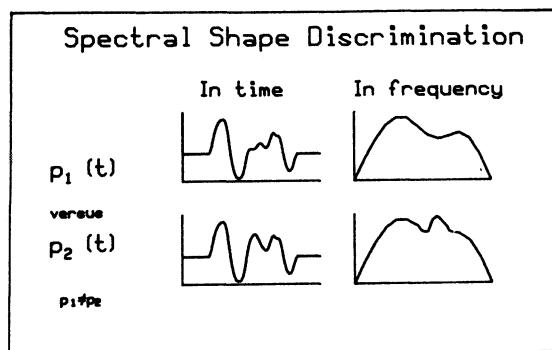


FIGURE 2, Spectral shape discrimination. The stimuli to be discriminated [$p_1(t), p_2(t)$] may be completely unrelated. Thus, in the frequency domain, their spectral shapes differ. The temporal waveforms also differ.

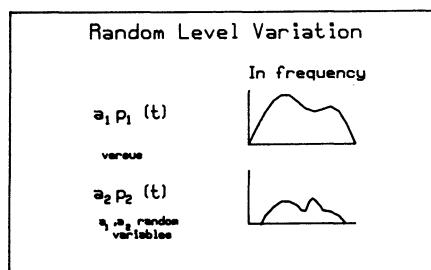


FIGURE 3, Stimuli to be discriminated are scaled by random variables (a_1, a_2) on each and every presentation to ensure that discrimination is based on spectral shape rather than intensity.

that the listener measures the sound parts of the spectra and simultaneously compares them. The absolute

The special experimental manipulation that ensures that shape, not absolute intensity, is the critical cue in the case of spectral shape discrimination is illustrated in Figure 3. It is randomizing the overall intensity level. On each and every presentation of the stimulus, the level at which they are presented is chosen at random. Thus, the scale constants, a_1 and a_2 , are random variables as the figure indicates. If the range of these random variables is sufficiently large, the stimuli heard in the discrimination task will clearly differ in intensity, and the observer will be forced to compare some other aspect of the stimuli to distinguish between them. In our case, that difference is the shape of the auditory spectra. The minimal comparison that must be made to achieve such discrimination is levels on two or more different

sound level of these measurements is largely irrelevant, because the stimuli change in absolute level on each and every presentation.

The differences in the structure of these two discrimination tasks force the observer to use somewhat different discrimination processes. In pure-intensity discrimination, the listener must construct some estimate of absolute intensity level and either compare two such estimates made at different times or compare a single estimate with some long term standard. In spectral shape discrimination, a simultaneous comparison of two or more spectral regions must be made, and from this comparison an estimate of relative level on any single presentation is largely irrelevant, because it is confounded by the randomization of overall level.

What we would like to do in this paper is review some of our research on this topic and especially emphasize what we have learned about how such spectral comparisons operate. As psychoacousticians, our primary interest is on how the auditory sense works, but we feel these experiments may provide some insight about how complex spectral discriminations are made in speech waveforms.

Procedure and Stimulus Conditions

Before proceeding with a description of the individual experiments, let us outline something about the procedure and stimulus conditions used in the research and why these experimental conditions were chosen. For almost all of the studies, we use a multitone complex. The stimuli generally cover the speech range, from 200 to 5000 Hz. The frequencies of the individual components are not, however, harmonic, as they are in speech. The tones are chosen so that successive components are equally spaced on a logarithmic frequency axis. Thus, the frequency ratio of successive components is a constant. The reason for choosing logarithmic spacing is as follows. We know the cochlea achieves a rough Fourier analysis of the stimulus in which different places along the basilar membrane are maximally sensitive to different frequencies. Roughly, this linear array is arranged so that equal spatial extent is coded as equal differences in logarithmic frequency. Our tones, therefore, provide a uniform stimulus over the linear receptor surface of the cochlea.

A typical discrimination task involves two stimuli, a "standard" complex and some alteration of the standard complex which we achieve by adding a "signal" to the standard. The signal itself consists of the in-phase addition of energy at one or more components to the corresponding component or components in the standard complex. We use equal-amplitude tones for our standard because the observers learn this standard easily. Thus, little training is needed in order to study various alterations from this standard. We use a two-alternative forced-choice procedure, and adaptively change the level of the signal to estimate the level which would yield 70.7% correct. Overall intensity is typically chosen at random over a 40-dB range in 1 dB steps. The median level is usually about 50 dB SPL per component.

In the studies reported here, the dependent variable is the level of the signal (the size of the increment) re the level of the corresponding component or components in the background. For example, if the level of the signal is equal to the level of the

corresponding component(s) of the standard, then we say the signal-to-standard ratio is 0 dB. In that case, the component to which the signal is added would be increased in level by 6 dB. In many studies, the signal is simply an increase in the intensity of a single component. But other changes have been studied as well, such as a variation in the amplitudes of all components of the standard. In the following, we recount some of the things we have learned about the perception of a change in the shape of such a complex auditory spectrum.

Effects of phase

In most of the experiments concerning profile analysis, the phase of each component of the multitone complex has been chosen at random and the same waveform (except for random variation of level) is presented during each "non-signal" interval. Therefore, the possibility exists that observers may recognize some aspect or aspects of the temporal waveform. If this were true, then discrimination could be based on some alteration of the temporal waveform during the "signal" interval rather than by a change in the spectral shape of the stimulus per se.

Green and Mason (1985) investigated this possibility directly. Multicomponent complexes were generated which consisted of 5, 11, 21, or 43 components spaced logarithmically. In all cases, the frequency of the lowest component was 200 Hz, the highest was 5 kHz. The overall level of the complex was varied randomly over a 40-dB range across presentations with a median level of 45 dB SPL. The signal consisted of an increment to the 1-kHz, central component.

In what Green and Mason termed the "fixed-phase" condition, for each number of components (5, 11, 21, and 43), four different standard waveforms were generated by randomly selecting the phases of each component of the complex. Each of these standards was fixed for a block of trials and signal thresholds were obtained for each of the different randomizations. Note that for these fixed-phase conditions, the same waveform, except for changes in overall level, occurred during each non-signal interval.

In what Green and Mason called the "random-phase" conditions, for each value of the number of components (5, 11, 21, and 43) 88 different standard waveforms were generated by randomly selecting the phase of each component of the complex. On each presentation of every trial, pairs of these 88 waveforms were selected at random (with replacement). Thus, the temporal waveforms generally differed on each presentation. The amplitude or power spectra of the stimuli were, however, identical.

The results are presented in Figure 4. For each value of component number, the open circles represent the thresholds obtained for each of the four randomizations in the fixed-phase condition. The solid triangles represent the data obtained in the random-phase conditions. The results indicate that changing the phase of the individual components and thus the characteristics of the temporal waveform has little, if any, effect on discrimination. This is true whether the same phase is used for a block of trials or if the waveform is chosen at random on each and every presentation. These data are consistent with those obtained by Green, Mason and Kidd

(1984) who had generated their waveforms using a procedure similar to the fixed-phase condition. The form of the function relating threshold to the number of components which compose the multicomponent background will be discussed in detail in a subsequent section.

The inability of changes in the phase of the individual components, and thus changes in the characteristics of the temporal waveform, to affect discrimination supports the view that, in these tasks, observers are, indeed, basing their judgements on changes in the power spectra of the stimuli.

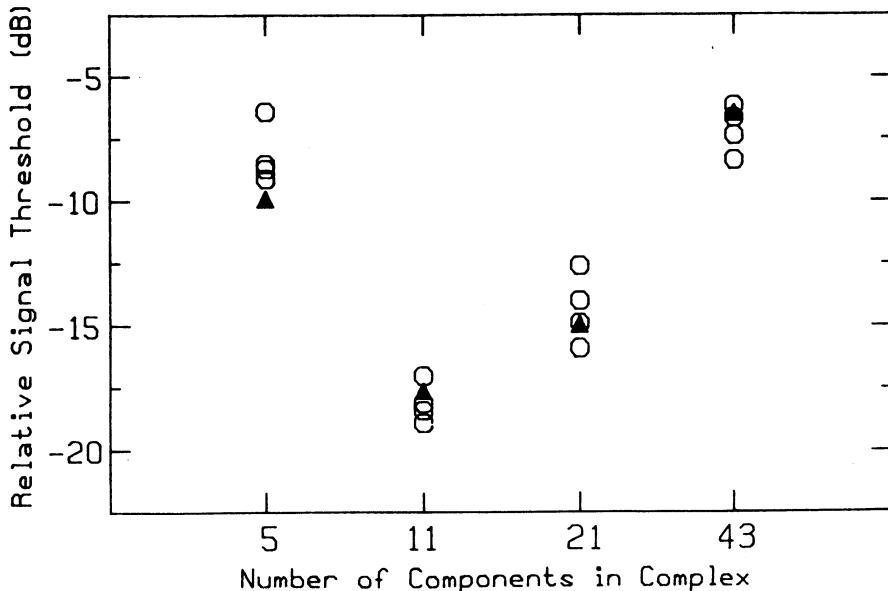


FIGURE 4. Signal threshold (dB) as a function of the frequency of the number of components in the complex. Open circles represent the data obtained for each of the four phase-randomizations when the phase of each component was fixed throughout a block of trials ("fixed-phase" condition). Filled triangles represent the data from the "random-phase" condition in which the phases of the components were chosen at random on each presentation.

Frequency Effects

So far we have demonstrated that the detection of changes in the shape of a complex auditory spectrum is based on changes in the power spectrum of the stimulus; the phase relation among the components is unimportant. The next question we consider is whether the ability to detect a change in the power spectrum is greatly influenced by the frequency region where the change occurs. Consider our complex standard composed of a number of sinusoidal components. Suppose we alter that standard spectrum by increasing the intensity of a single sinusoid. A natural question is--does the frequency locus of

the change greatly affect the ability to detect the change? The answer to this question settles an important practical issue-- to what degree are different frequency regions homogeneous? In speech, at least for vowels, the significant spectral changes typically occur within the range of 500 to 2000 Hz. As far as we are aware, there is no claim that small alterations of the spectrum are better detected at one frequency region rather than some other. Thus, we would be surprised to find that the ear's ability to detect a small change in the spectrum differs greatly as a function of frequency.

This question is also of basic interest in psychoacoustics, because it bears on the question of intensity coding and whether or not temporal factors, such as the synchrony of discharge patterns, are utilized as part of the intensity code. Sachs and Young (1979) and Young and Sachs (1979) have demonstrated that 'neural spectrograms' based on neural synchrony measures preserve the shape of speech spectra better than those based on firing rate codes. We were, therefore, particularly interested in how well the observers could detect a change in spectral shape at higher frequencies. At the highest frequencies, above 2000 Hz, neural synchrony deteriorates and, if that code were used to signal changes in spectral shape, then the ability to detect such alterations in the acoustic spectrum should also deteriorate. Certainly, differences among vowels are not signaled by changes in the location of higher frequency formats. But, in the case of speech, this frequency limitation may be the result of the production system, that is, the coding system, not the decoding system.

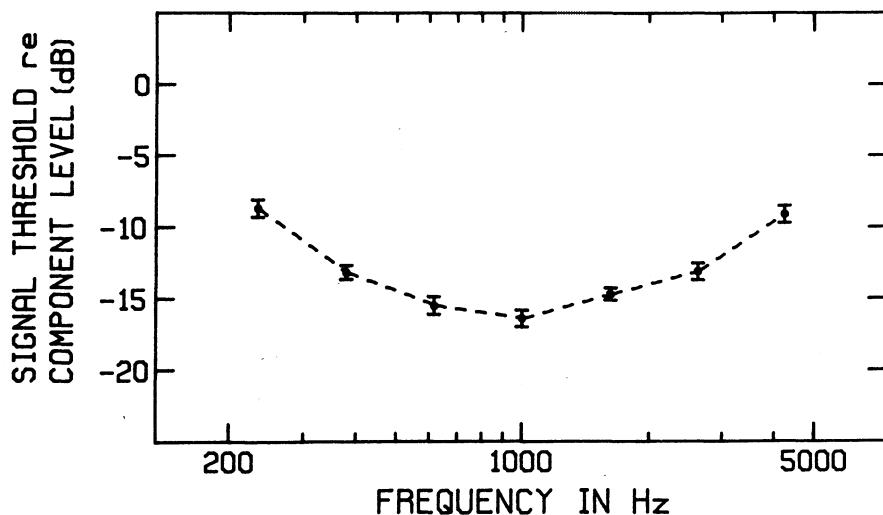


FIGURE 5. Signal threshold (dB) as a function of the frequency of the signal. A twenty-one-component complex was used as the standard. The frequency of the lowest component was 200 Hz; the frequency of the highest component was 5000 Hz. The signal, whose frequency is indicated on the abscissa, was added in-phase to the corresponding component in the complex.

In a previous study, Green and Mason (1985), we had measured how the locus of the frequency changes affects the ability to detect a change in complex spectra. These results suggested that the mid-frequency region, 500 to 2000 Hz, was the best, but variability among the different observers was sizable. Also, those data were taken after a previous experiment in which the signals were in the middle of that range. Although extensive training was given in the later experiment to all the different frequencies tested, it is conceivable that some of the data were influenced by the preceding experiment. In any case, the recent move of our laboratory provided an opportunity to recruit a new set of listeners that were truly naive with respect to the parameter of interest.

The results of our most extensive experiment (Green, Onsan, and Forrest, 1987) on this issue are shown in Figure 5. The standard spectrum was a complex of 21 components, all equal in amplitude and equally spaced in logarithmic frequency. The overall level of the standard was varied over a 20 dB range with the median value of 60 dB. The signal, whose frequency is plotted along the abscissa of the figure, was an increment in the intensity of a single component. The ordinate, like that of Figure 4, is again the signal level re the level of the component to which it was added. The results show that best detection occurs in a frequency range of 300 to 3000 Hz, with only a mild deterioration occurring at the higher and lower frequencies. These results give little support to the idea that neural synchrony is used to estimate intensity level, because, were such the case, there should be a more marked deterioration in the ability to hear a change in the spectrum as a function of frequency.

PROFILE-ANALYSIS AND THE CRITICAL BAND

The evidence presented thus far suggests the detection of a change in spectral shape, or profile-analysis, is a "global" process relying on simultaneous comparisons in two or more regions of the spectrum. An issue of central concern is the width of the spectrum over which these comparisons can be made. If one were to invoke classical "critical-band" notions, which pervade much of psychoacoustic research, it would be expected that only frequencies close to the frequency of the signal could be used in detecting an increment.

Green, Mason, and Kidd (1984) obtained data which address this issue. In their experiment, the signal consisted of an increment to the 1-kHz, central component of a multitone complex. The multitone complexes consisted of equal-amplitude, logarithmically-spaced components. In the first condition, what we will refer to as the "range" condition, the standard consisted of a three-component complex. The parameter was the range of frequencies spanned by the standard, that is, the separation in frequency between the two components which flanked the central, 1-kHz component.

In the second condition, what we will call the "range/number" condition, the number of components as well as the range was varied. Additional flanking components were added to the complex resulting in multitone complexes of 3, 5, 7, 9, and 11 components. These additional components increased the range of frequencies covered by the standard.

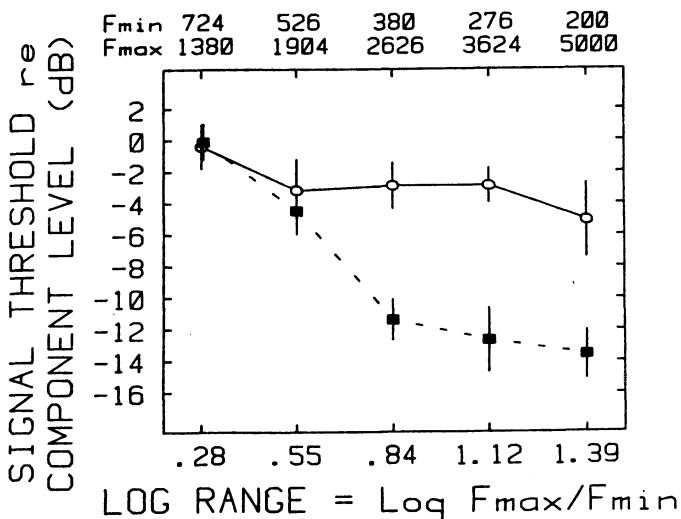


FIGURE 6, Signal threshold (dB) as a function of the logarithm of the ratio of frequencies spanned by the complex. Open circles represent the data obtained from the "range" condition, in which each complex comprised three components. The signal was always added to the central component of the complex, a 1000 Hz component. The numbers at the top of the graph give the frequency of the other two components of the complex. Solid squares represent data obtained from the "range/number" condition in which the number of components in the complex and the range was varied. Again, the signal is an increment in the central component. From the left-most portion of the graph, the squares represent complexes comprising 3, 5, 7, 9, and 11 components respectively.

The results of these two conditions are presented in Figure 6. The abscissa is the logarithm of the ratio of the highest to the lowest component in each complex. The data obtained in the range condition, with the three-component complexes, are plotted as open circles. The solid squares represent the data obtained in the range/number condition when the range of frequencies spanned by the complex and the number of components covaried. Each point is the mean of six estimates of threshold obtained from the three subjects who participated. The error bars represent the mean of the standard error computed for each observer.

Focusing on the data obtained in the range/number condition (solid squares), it is clear that as the number of components is increased, performance improves by 10 dB or more. Although only a small improvement is realized when the number of components increases beyond seven, the data obtained with seven components indicate that tones almost 1.5 octaves away from the central, 1-kHz

component (2626 Hz and 380 Hz) have a dramatic effect on performance. This result is in conflict with "critical-band" notions which would predict that energy at frequencies remote from the signal would have little effect on its detection.

The data obtained with the three-component complexes (open circles) also indicate that increasing only the range of the complex improves performance but not to the extent found when the number of components is also increased.

In short, for a given frequency range, performance is improved when the number of components which compose the profile, that is, its density is increased. This result was also obtained by Green, Kidd, and Picardi (1983). Their data showed, in addition, that if the density of components in the complex is great enough, then several components fall very close to the frequency of the signal and detection performance will deteriorate. Such an outcome is explained by simple masking and its existence supports the critical band concept. In such a case, the additional components fall within a critical-band surrounding the frequency of the signal component, and thus an increment to the signal component produces a relatively smaller increase in power in its region of the spectrum.

In summary, the conflict with classical "critical-band" concepts arises because energy at frequencies remote from that of the signal influences performance in these tasks. The data confirm the notion that profile analysis is a global process which relies upon the integration of information across many critical-bands.

Profile Analysis versus Simple Intensity Discrimination

In the concluding section of this paper, we compare the acuity of discriminating a change in the shape of a complex spectrum to the acuity of detecting a change in absolute intensity level. As reviewed in the first section of this paper, one may distinguish two separate processes for comparing intensity in a complex spectrum. The first we called pure-intensity discrimination; this process detects a change in absolute intensity level. The acuity of this process can be measured in tasks where the spectrum of the signal does not change its shape, but is simply altered in level. We have contrasted this process with the detection of a change in the shape of the complex auditory spectrum, what we have called profile analysis. In detecting a change in spectral shape, the process must be one of simultaneous comparisons of intensity levels at different regions of the spectrum, because random variation in the overall level of the spectrum on successive presentations renders the use of absolute level on any presentation an ineffective strategy. For a fixed change in intensity, can one hear that change best in a pure-intensity discrimination task or in a profile task? A clear answer to this question is of some practical importance, because for many naturally-occurring stimuli such as complex speech spectra, both processes are potentially available. Presumably, the observer uses either a combination of the two systems or the more sensitive system alone. To predict performance in a variety of realistic situations, one would have to know the relative sensitivity of the two systems.

Comparison of detection performance in the two situations is, however, complicated by the issue of prior training and experience. The situation is not unlike that of testing the ability of observers to

hear some phonemic distinction in a particular language. If one uses a group of subjects whose natural language uses this distinction, then one may expect finer discrimination capacity from that group than from another group of subjects whose native language does not use this distinction. Similarly, we have observed that listeners with a long history of training in pure intensity-discrimination experiments often do poorly when first confronted by a task involving the detection of a change in spectral shape. It is also true that observers who are well-practiced in detecting changes in spectral shape often find detection of simple intensity changes initially difficult. Recently, a well-trained profile observer complained, when asked to discriminate a change in the intensity of a single sinusoid, that the only thing he could listen for was a change in loudness!

A second factor that makes the comparison of the detection performance in the two tasks difficult is that there is more range in the ability of different people to hear simple changes in intensity level than is usually admitted in the literature. The impression that the Weber fraction is nearly constant over individuals is created largely by the use of a very compressive measure of the Weber fraction in dB [$10\log(1+\Delta I/I)$]. One often reads that the Weber fraction is about 1 dB. What is not appreciated is that a change from 0.5 to 1.5 dB corresponds to a 10 dB change on the scale of signal-to-background level which we have commonly used in profile experiments. Individual differences among listeners are sizable. Using our scale of signal-to-component level, then we often find differences of 10 dB among individuals in both pure-intensity discrimination tasks as well as profile tasks.

A final complication is that the observers we use in most of our profile tasks are not a random selection from the population; rather, they are selected on the basis of previous listening performance. Some observers find it extremely difficult to hear the change in shape of a complex spectrum. While they improve with practice, it does not appear likely that they will ever be useful participants in a series of experiments involving the comparison of thresholds obtained in a variety of different experimental conditions. Our usual procedure is to train and test subjects over a period of one or two days (two hours of listening per day) on the detection of an increment in a 1000 Hz tone in an 11- or 21-component complex. For the listener to continue in these experiments, we require that detection performance reach the -10 to -20 dB range at the end of two or three days. In general, we believe that practically all subjects could be trained to reach this level of performance, but if more than three days are required we feel that such observers would require an excessive amount of training throughout the various conditions of the experiment.

A direct comparison of the relative sensitivity of two groups of listeners was recently made by Green and Mason (1985). They compared two groups of observers--five experienced in profile listening, five who were not. The five inexperienced profile listeners had considerable training in tasks that could be classified as pure-intensity discrimination tasks. The thresholds for the ten observers were measured in two detection tasks, a pure intensity-discrimination task and a profile task.

The pure-intensity discrimination task was the detection of a change in the level of a 1000 Hz sinusoid. The sinusoid was fixed in level at 40 dB SPL. The profile task was the detection of the change in the intensity of that same component, but the 1000 Hz component was surrounded by 10 other, equal-amplitude components. We used the familiar, 11-component complex (200-5000 Hz). To make the tasks comparable, the level of all the components was also fixed on each and every presentation, at 40 dB SPL. The ratio of the frequencies between successive components of the complex was approximately 1.38. Thus, the two neighbours to the 1000 Hz components had frequencies of 1379 and 724 Hz. The signal duration in both tasks was 100 msec. The thresholds were estimated from the mean of 6 runs of 50 adaptive trials (two down-one up). Table 5-1 presents the thresholds estimated in the two tasks for the ten observers.

Table 5-1

Entry is the relative signal threshold in dB
(standard error of estimate)

Observers	Single Sinusoid	Profile	Diff (SS-P)
Profile			
Experienced			
1	-10.5 (1.4)	-18.6 (1.7)	8.1.
2	-6.4 (2.0)	-13.6 (0.6)	7.2.
3	-12.0 (0.8)	-18.5 (1.3)	6.5
4	-11.2 (1.3)	-15.8 (1.2)	4.6
5	-18.0 (1.5)	-22.7 (2.3)	4.7
mean	-11.6 (1.4)	-17.8 (1.4)	6.2
Profile			
Inexperienced			
6	-20.0 (1.6)	-10.9 (2.2)	-9.1
7	-13.2 (2.0)	-12.3 (1.6)	-0.9
8	-19.7 (1.0)	-9.2 (1.3)	-10.5
9	-14.0 (1.0)	-10.0 (1.6)	-4.
10	-17.4 (0.8)	-20.2 (1.4)	+2.8
mean	-16.9 (1.6)	-12.5 (1.6)	-4.3

As can be seen in the table, there is almost a perfect interaction between thresholds in the two tasks and previous training. The best average detection performance is about -17 dB for both groups, but it occurs for different conditions. For the experienced profile listeners, it occurs in the profile conditions. For the inexperienced profile listeners, it occurs in the single sinusoid condition. The average difference between performance on the favored and unfavored task is also very similar in the two groups, about 5 dB. The pattern of interaction between past listening experience and the two detection tasks is reflected by nearly every individual observer with one singular exception (Observer 10). That observer, whose performance level is good on both tasks, is somewhat better on the

profile task, despite the lack of previous experience. Note the range of thresholds obtained for either group within each task. Such differences among individuals are typical.

Presumably, with enough training, both groups would improve on the unfamiliar task, but, unfortunately, we have no firm data to support that conjecture. Informally, we tried to improve the performance of the inexperienced profile listeners in the profile tasks, but their thresholds, after an additional 2000 trials, did not improve very much. We are still uncertain about how best to interpret this result. The interaction present in the data reflects either a difference in training or real individual difference among observers. It may be that differences in past experience can simply not be overcome by a few thousand trials. One could argue that it is like trying to hear an acoustic distinction that is not used in one's native language. Alternatively, it is possible that there are simply two different types of observers. One type good at discriminating changes in absolute intensity, another good at discriminating changes at spectral shape.

While it is unlikely that a random sample of ten individuals would divide so perfectly, one cannot claim that the profile group was a completely random sample. As described previously, some preliminary testing was completed before selecting this group and such tests could indeed have biased the group to be good 'profile' listeners. The difference in their performance in the two tasks is reasonably uniform over all the observers experienced in profile listening. The group inexperienced in profile listening was probably a more random selection from the general population, and their results are more mixed. Observers 7 and 10 show little difference in their performance in the two tasks (their difference scores are -0.9 and +2.8 dB, respectively). Whether there are really different types of observers or simply differences in past experience remains a fascinating, but unsettled, question.

REFERENCES

1. Green, D.M., & Mason, C.R. (1985). Auditory profile analysis: Frequency, phase and Weber's Law. Journal of the Acoustical Society of America, 77, 1155-1161.
2. Green, D.M., Kidd, G., Jr., and Picardi, M.C. (1983). Successive versus simultaneous comparison in auditory intensity discrimination. Journal of the Acoustical Society of America, 73, 639-643.
3. Green, D.M., Mason, C.R. and Kidd, G., Jr. (1984). Profile analysis: Critical bands and duration. Journal of the Acoustical Society of America, 75, 1163-1167.
4. Green, D.M., Onsan, Z.A., & Forrest, T.G. (1987). Frequency effects in profile analysis. Journal of the Acoustical Society of America, in press.
5. Sachs, M.B., & Young, E.D. (1979). Encoding of steady-state vowels in the auditory nerve: Representation in terms of discharge rate. Journal of the Acoustical Society of America, 66, 470-479.

6. Young, E.D., & Sachs, M.B. (1979). Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers. Journal of the Acoustical Society of America, 66, 1381-1403.

GENERAL DISCUSSION OF SESSION 4: TIMBRE

Chairman: Louis C.W. Pols

A total of 9 papers was originally presented in this session at the conference. These papers dealt with peripheral processing and the capabilities and limitations of the auditory system, both for psychophysical stimuli and for speech.

Several specific points were addressed, such as the interaction between excitation and timbre, critical bandwidth, and large scale spectral integration. Some attempts were made to present alternative, spectrally and articulatory based, speech sound descriptions.

The animated discussion, especially in the morning session, focussed on the following major topics:

- A further clarification of details of the various papers;
- B possibilities and limitations of the ear's spectral and temporal resolving power;
- C subsequent levels of spectral interpretation towards phoneme and speech recognition, including spectral smearing;
- D clarification of the distinction between close listening and global speech identification.

re A (specific details)

Two of the papers presented during the conference are not published in this book. Although they both raised a lively discussion, they will of course not be covered here. For the remaining papers, some critical remarks made during the discussions are summarized below:

Terhardt: How exactly was the vowel identification experiment done with harmonic, inharmonic, and noise excitations? What exactly is FFT and maxigram, and how was the resynthesis done?

Schwartz-Escudier: Some people felt that there was too much emphasis in this paper on the concept of F_2' .

Traunmüller: Some people felt that the correlation between psychoacoustical variables and functional dimensions was rather speculative. A clearer distinction should be made between auditory and phonetic similarity.

Miller: The question was raised whether the system could already produce testable results.

Espinoza-Varas: There was a remark about the possible danger of using a large range (extended over several vowel categories) for vowel identification whilst using only two response categories.

One participant wondered why only F_2 was plotted irrespective of F_1 and F_3 .

Green: It should be realized that there is a large difference between the inharmonic stimuli used by Green and speech, which generally is harmonic, has a regular phase, has obvious peaks and valleys, and is not stationary. The audience would like to see a model to describe the results, especially with respect to the high resolution and the effect over a very wide frequency range.

re B (the ear's resolving power)

Speech people were somewhat surprised to hear that even psychophysicists did not seem to agree upon the ear's spectral and temporal resolution. On the one hand there is the concept of critical bandwidth, on the other hand Terhardt's model has a resolution of one tenth of a critical bandwidth (according to tuning curves); furthermore we can hear 1 dB differences and we can hear out the 20th harmonic in a complex. This seeming discrepancy is, at least partly, related to stimulus specifications and to the tasks involved. The ear is capable of doing a very detailed spectral analysis (and so is Terhardt's model); however, in everyday life the critical band is a very practical concept. The high resolution capability can be essential for pitch extraction and localization, but also for bad signal-to-noise conditions or in order to separate competing voices.

Another discussant stressed that frequency selectivity refers to the ability to analyze complex stimuli, which is basically a capability of the auditory system in the spectral domain. Frequency discrimination, on the other hand, refers to the ability to distinguish between successively presented stimuli that differ (only) in frequency. This ability is probably based on fine temporal coding, present in the trains of action potentials in the auditory nerve. It seems that the high frequency selectivity in Terhardt's model is based on a combination of these two capabilities (i.e. one in the spectral and the other one in the temporal domain).

re C (towards speech recognition)

At subsequent speech levels there is a further smearing and integration, leading to concepts like F_2' and 3 to 3.5 Bark. In fact it is probably better to forget about F_2' specifically, and to use a more general integration concept working over the whole formant range. The formant integration may be about 3 Bark; however, this does not exclude a formant frequency difference limen of about .2 Bark.

An interesting question was whether outside the speech area there are examples of a similar spectral integration. One furthermore wondered whether it has already been proven that the 3.5 Bark integration allows for error-free recognition of all vowels.

There were also several warnings that one should not forget that discrimination and categorization are two quite distinct tasks, each with its own level of signal representation and interpretation. Under natural listening conditions with reverberation and masking

noise, the signal details are not well defined. The whole-spectrum approach with global distance metrics did not get much attention.

re D (close auditory listening vs. speech perception)

After a long discussion there seemed to be some agreement that there really are different levels of signal representation. The ear is capable of doing a very detailed analysis, which is the basis for subsequent coarser levels of representation up to the cognitive level, where nothing is left but a concept. One should not confuse the psychoacoustic level with a phonemic representation.

Both psychophysicists and speech people will continue to do their own research, studying the various levels. They will not change their research programs, but hopefully every now and then they will remember this conference.

Chapter 5

PHYSIOLOGICAL CORRELATES OF SPEECH PERCEPTION

PERIPHERAL AUDITORY PROCESSING OF SPEECH INFORMATION: IMPLICATIONS FROM A PHYSIOLOGICAL STUDY OF INTENSITY DISCRIMINATION*

Bertrand Delgutte

Eaton-Peabody Laboratory, Massachusetts Eye and Ear Infirmary,
Boston, Massachusetts, and Research Laboratory of Electronics,
Massachusetts Institute of Technology, Cambridge, Massachusetts.

INTRODUCTION

A fundamental problem in speech processing by the peripheral auditory system is how to represent the short-time spectrum over the broad range of stimulus levels and signal-to-noise ratios of conversational speech. Because most auditory-nerve fibers have a limited dynamic range, profiles of average discharge rates against characteristic frequency (CF) seem to provide a poor representation of the formant frequencies of vowel-like sounds at stimulus levels and signal-to-noise ratios well within the conversational range (Sachs and Young, 1979; Sachs et al., 1983; Miller and Sachs, 1983). However, it cannot be concluded from these results that average discharge rates of all auditory-nerve fibers fail to provide sufficient information for distinguishing speech sounds, because these studies have not examined in detail the possible roles of high-threshold fibers (Liberman, 1978), and efferent feedback to the cochlea (Wiederhold and Kiang, 1970). In this paper, we approach this "dynamic range" problem by using the concepts of signal-detection theory to relate psychophysical performance in intensity discrimination to the activity of auditory-nerve fibers. This approach is motivated by the hope that the large body of detailed parametric data on intensity discrimination will provide more constraints on models of auditory processing than the relatively sparse data on the discrimination of speech parameters. From the point of view of a place model of auditory spectral representation, spectral discrimination and intensity discrimination are closely related because a change in the spectrum causes changes in the activity of individual frequency channels that are similar to those resulting from a change in the intensity of a tone at the channel center frequency. In this context, a channel is a processing element that receives inputs from fibers innervating a restricted portion of the cochlea, but free to differ in other characteristics such as sensitivity.

A number of detection-theoretic models relating intensity-discrimination performance to the discharge patterns of auditory-nerve fibers have been proposed (Siebert, 1965, 1968; Sanderson, 1975; Teich

*This work greatly benefited from discussions of psychophysical topics with H.S. Colburn, S. Buus, and M. Florentine. N.Y.S. Kiang made valuable comments on the manuscript. The assistance of P. Riley in the physiological experiments is gratefully acknowledged. This research was supported by NIH Grant NS 13126. A key portion of the work was done while the author was visiting the Centre National d'Etude des Télécommunications, Lannion, France.

and Lachs, 1979; Lachs et al., 1984; Colburn, 1981, 1984; Winslow and Sachs, 1985; Winslow 1985; Viemeister, 1986). These models are based on the idea that the stochastic behavior of auditory neurons imposes fundamental limitations on the performance achievable in any detection or discrimination task (Siebert, 1965). Siebert's (1965, 1968) optimal-processor model was based on a functional description of the responses of auditory-nerve fibers that included multiple frequency-selective channels with saturation of the rate-level functions and stochastic (Poisson) behavior of the discharges. This model predicted Weber's law for tones, i.e. the intensity difference limen (expressed in decibels) was nearly constant as a function of level. This behavior was due to the fact that, as level increased, model fibers whose CF was increasingly far from the stimulus frequency began to respond, so that at each level there was a roughly constant pool of unsaturated fibers. The necessity of such "spread of excitation" to intensity discrimination has been questioned on the basis of psychophysical results showing that intensity DL's for broadband stimuli (Miller 1947; Harris, 1963; Penner and Viemeister, 1971; Raab and Goldberg, 1975) or for tones in notched noise (Viemeister, 1972, 1983; Moore and Raab, 1974) obey Weber's law. Spread of excitation cannot occur for broadband stimuli because these stimuli excite fibers innervating the entire length of the cochlea even for moderate sound levels. A similar argument applies to tones in notched noise if it is assumed that the noise prevents the spread of tone excitation. Moreover, prediction of Weber's law for tones in Siebert's model is inaccurate because psychophysical DL's show a small but systematic decrease as intensity increases, a phenomenon known as the "near miss" to Weber's law (McGill and Goldberg, 1968; Rabinowitz et al., 1976; Jesteadt et al., 1977). It has been suggested that the change from Weber's law for broadband stimuli and tones in notched noise to the "near miss" for tones in quiet is due primarily to the increase in the number of excited channels (Bos and de Boer, 1966; Moore and Raab, 1974; Florentine and Buus, 1981). If this interpretation is correct, a key goal for a physiologically-based model of intensity discrimination would be to obtain Weber's law for individual frequency channels.

In this paper, we examine the possibility that, by incorporating into Siebert's model the differences among auditory-nerve fibers in their thresholds, spontaneous discharge rates, and dynamic ranges (Liberman, 1978; Schalk and Sachs, 1980; Evans and Palmer, 1980), we can better predict psychophysical performance in intensity discrimination. We first present physiological data on intensity difference limens (DL's) for single auditory-nerve fibers, measured using stimulus paradigms and detectability measures explicitly mimicking those of psychophysics. We then incorporate these data into a model of intensity discrimination that differs from previous models (Colburn 1981; Winslow, 1985; Viemeister, 1986) in that it includes both multiple frequency-selective channels and a physiologically realistic distribution of fiber thresholds within each channel. Finally, we show that a version of this model provides a stable representation of the spectra of speech sounds over variations in stimulus level.

I. MEASUREMENTS OF INTENSITY DL'S FOR AUDITORY-NERVE FIBERS

A. Method

Basic physiological methods for recording from auditory-nerve fibers in anesthetized cats were as described by Kiang et al. (1965). Our procedure for measuring intensity difference limens (DL's) of single auditory-nerve fibers is designed after the two-interval, two-alternative forced-choice paradigm of psychophysics. This method is illustrated in Fig. 1. Pairs of 50-ms tone bursts at the CF, respectively at levels L and $L+\Delta L$, are presented repeatedly, with random order of presentation within each pair. The numbers of spikes observed in 50-ms intervals coincident with the tone bursts are counted for each presentation, and the level increment ΔL is adjusted by means of a "PEST" adaptive procedure (Taylor and Creelman, 1967) so that the probability that the spike count $N(L+\Delta L)$ in the interval coincident with the most intense tone exceeds the count $N(L)$ in the other interval approaches a certain criterion P_{cr}^* . The PEST procedure is stopped when the increment in signal level reaches 1/2 dB. These measurements are made using a probability criterion of 0.75 for reference levels increasing in 5 dB steps from 15-20 dB below fiber threshold until $L+\Delta L$ reaches about 90 dB SPL.

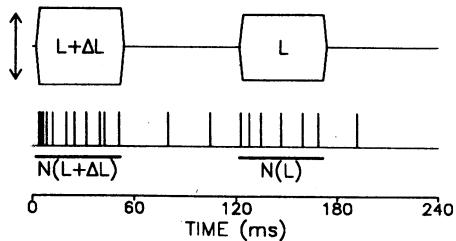


FIGURE 1, Schematic diagram showing the experimental paradigm for measuring intensity difference limens of auditory nerve fibers. The top trace shows the envelopes of 50-ms tones at the CF. The bottom trace shows spike times that might be recorded from a fiber in response to the tone bursts. The pair repetition rate is 3/s, and the silent interval between the two tones in each pair is 70 ms.

B. Results

Figure 2A shows intensity DL's as a function of level $L+\Delta L/2$ for two auditory-nerve fibers in the same cat. Both fibers have about the same CF, but differ in their spontaneous rate of discharge (SR). The DL-intensity function for the high-SR fiber (circles) has basically a U shape, with a clear minimum and a rapid growth (i.e. deterioration in performance) on both sides of the minimum. Circles in Fig. 2B

*More specifically, to take into account cases when there is an equal number of spikes in both intervals, we use the following criterion:

$$P[N(L+\Delta L) > N(L)] + 0.5 P[N(L+\Delta L) = N(L)] = P_{cr}$$

This criterion is analogous to subjects' guessing when they hear the two tones as equally loud.

show the rate-level function for a 50-ms tone at the CF for this fiber. The range of levels over which the DL is close to minimum corresponds roughly to the range in which discharge rate grows

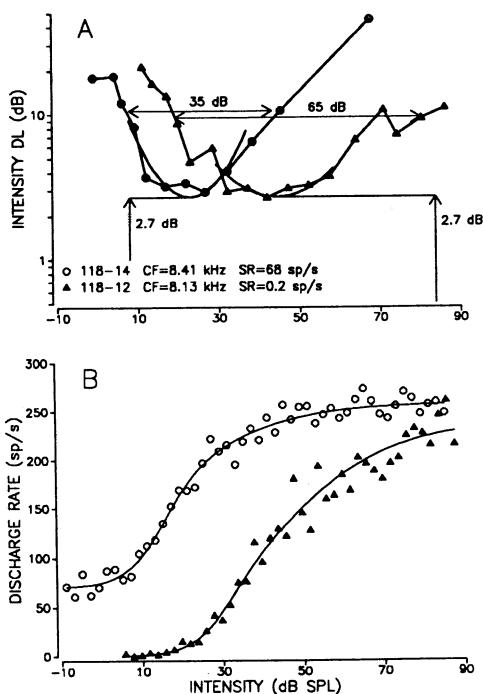


FIGURE 2,
A. Intensity DL's as a function
of level of a 50-ms tone burst
at the CF for two auditory-
nerve fibers in the same cat.
Symbols connected by thin lines
represent the measured DL's.
Smooth lines are parabolic fits
to the data used in estimating
the minimum DL's. Vertical
arrows show the minimum DL's,
while horizontal arrows indicate
the range of levels over which
the DL is less than 10 dB.
B. Discharge rate as a function
of the level of a tone at the CF
for the same fibers as in A.

rapidly with level. The DL-intensity function for the low-SR fiber (triangles) shows a minimum, as does the high-SR fiber, but the intensity at which the minimum occurs is higher, and the rate of growth of DL beyond the minimum is lower than for the high-SR fiber. As a result, the range of levels over which the DL is lower than

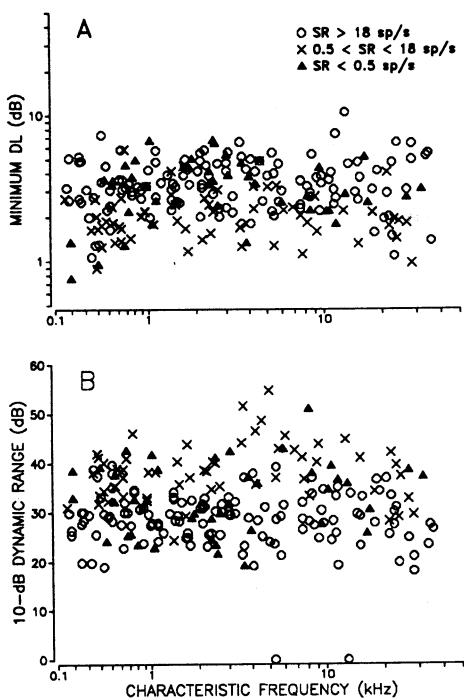


FIGURE 3,
A. Minimum intensity DL's
against characteristic frequency
for 229 auditory-nerve fibers in
17 cats.
B. Range of intensities over
which the DL is less than 10 dB
plotted against CF for the
same fibers as in A.

10 dB (which we will call the "10-dB dynamic range") is 30 dB wider for the low-SR fiber than for the high-SR fiber.

Figure 3A shows minimum DL's as a function of CF for 229 auditory-nerve fibers. Fibers are segregated according to the three groups defined by Liberman (1978) on the basis of spontaneous discharge rates: "high-SR" ($SR > 18$ spike/s), "medium-SR" ($0.5 < SR < 18$ sp/s), and "low-SR" ($SR < 0.5$ sp/s). Minimum DL's range between 1 and 7 dB, with no obvious trend as a function of CF. Most of the fibers for which the minimum DL is lower than 2 dB are from the medium-SR group. This good intensity resolution of medium-SR fibers is expected because the discharge rates of medium-SR fibers grow more rapidly with level than do those of fibers from the other two groups (Liberman, 1978). Figure 3B shows 10 dB dynamic ranges as a function of CF for the same set of fibers as in Fig. 3A. Although this measure is obviously correlated with the minimum DL, it provides an absolute measure of performance that is useful in comparisons with psychophysical data. The 10-dB dynamic range varies from fiber to fiber between about 20 and 50 dB, with medium-SR fibers having the largest dynamic ranges, and high-SR fibers having the lowest dynamic ranges. These differences are consistent with the fact that the range of levels over which discharge rates increase rapidly with level is greater for medium-SR and low-SR fibers than for high-SR fibers (Schalk and Sachs, 1980; Evans and Palmer, 1980). As the minimum DL, the 10-dB dynamic range shows no obvious trend as a function of CF. Figure 4 shows histograms of the distribution of the level L_{min} at which the minimum DL occurs for each of the three SR groups. For each fiber, L_{min} is expressed relative to the average thresholds of many high-SR fibers having similar CF's (Liberman and Kiang, 1976). Relative level can be considered to be analogous to sensation level in psychophysics. In the frequency range from 1 to 10 kHz, a relative level of 0 dB corresponds roughly to 10 dB SPL. Although the three distributions overlap considerably, the mean L_{min} for low-SR fibers (30 dB) is greater than that for medium-SR fibers (23 dB), which in turn is greater than that for high-SR fibers (12 dB). The total range of L_{min} for the three fiber groups exceeds 50 dB. This distribution resembles that of the relative thresholds* of auditory-nerve fibers (Liberman, 1978).

Because the characteristics of DL-intensity functions seem to be homogeneous with respect to CF, it is appropriate to average these functions over fibers with different CF's. Separate averages were computed for each of the three SR groups. Figure 5 shows average DL's as a function of relative level for the three SR groups. The average functions have the same general shapes as the functions for individual fibers, and reflect the differences between SR groups observed for individual data. In particular, the minimum of the average DL for medium-SR fibers is lower than that for the other two groups, and the average DL's for low-SR and medium-SR fibers grow more slowly with level beyond the minimum than does the average DL for high-SR fibers.

*If threshold is defined as a constant increment in discharge rate above spontaneous. Other definitions of threshold give somewhat different results (Geisler et al., 1985).

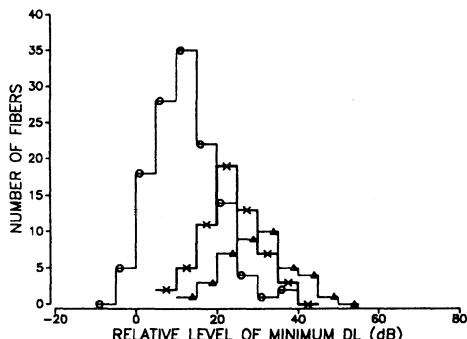


FIGURE 4, Histogram of the distribution of the relative level L_{\min} of the minimum DL for the 3 groups of auditory-nerve fibers. Relative level is stimulus level minus the average threshold of many high-SR fibers in a 0.4-octave band of CF's centered at the stimulus frequency (Fig. 3 in Liberman and Kiang, 1976). The number of fibers in the three histograms are 129, 60, and 40 for the high-SR, medium-SR, and low-SR groups respectively.

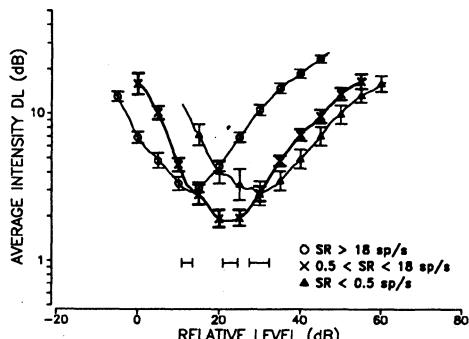


FIGURE 5, Average intensity DL's against relative level for the same groups of auditory-nerve fibers as in Fig. 4. Prior to averaging, the DL-intensity function of each fiber was translated horizontally in order to make its L_{\min} coincide with the average L_{\min} for the corresponding SR group. Vertical and horizontal bars indicate 95% confidence intervals (± 2 standard errors) for the average DL's and the average L_{\min} respectively.

In summary, intensity DL's of single auditory-nerve fibers show a minimum in the intensity range where discharge rate increases rapidly with stimulus level, with poorer performance in the subthreshold and saturation regions. For intensities in the vicinity of the minimum DL, single-fiber DL's are always within an order of magnitude of psychophysical DL's (about 0.7 to 1 dB) for tone bursts of the same duration, and approach psychophysical DL's for certain fibers, particularly those from the medium-SR group. Thus, our measurements are consistent with conclusions drawn from simple statistical models that spike-count information from a small number of unsaturated auditory-nerve fibers suffices to account for psychophysical performance (Colburn, 1981; Viemeister, 1983; Winslow, 1985). On the other hand, the relatively good correspondence between single-fiber DL's and psychophysical DL's applies only to a narrow (20-50 dB) range of levels around the minimum DL. Therefore, accounting for psychophysical performance over a broad range of levels requires combining intensity information from several fibers with different sensitivities (L_{\min}). This is achieved in the model of intensity discrimination described in the next section.

II. MODEL OF INTENSITY DISCRIMINATION BASED ON RESPONSES OF AUDITORY-NERVE FIBERS

The present model of intensity discrimination consists of three components:

- (1) A peripheral component providing predictions of the level dependence of single fiber DL's that can be compared with measurements in auditory-nerve fibers.
- (2) A rule for combining intensity information from auditory-nerve fibers that have the same CF, but differ in their sensitivity. Together, the first two components form a "single-channel" stage.
- (3) A rule for combining information across fibers with different CF's. All three components constitute the "multiple-channel" model.

A. Predicting single-fiber intensity DL's from rate-level functions

We have noted in Sec. I that the general dependence of single-fiber intensity DL's on sound level, as well as the differences in this dependence between the three SR groups, were consistent with what is known about rate-level functions of auditory-nerve fibers. However, specifying how the mean discharge rate (or spike count) varies with level for a particular stimulus is not sufficient for predicting intensity DL's: One also needs a statistical characterization of the variability of the spike count over repetitions of the same stimulus. If, for simplicity, spike-count statistics are approximated by Gaussian distributions, the performance of a fiber in distinguishing two stimuli differing in intensity is completely characterized by the detection-theoretic measure d' , which is equal to the difference in the mean spike counts over the standard deviation of the spike count. Because data on spike count variances as a function of level are sparse (Teich and Khanna, 1985; Young and Barta, 1985), and altogether lacking for the brief stimuli used in the DL measurements, a necessary preliminary step to modeling was to obtain spike count data for the same stimuli and the same fibers as those used in the DL measurements.

Means and variances of the spike counts were estimated from 25 repetitions of a 50-ms tone burst at the CF. Figure 6A (symbols) shows average rate-level functions for auditory-nerve fibers in the three SR groups. Continuous lines show fits to the average data by means of a model of rate-level functions with sloping saturation proposed by Sachs et al. (1986). The fits are close for all levels and all three SR groups. Figure 6B shows averages of the spike count variances plotted against averages of the mean spike counts for three groups of fibers. The relation between mean spike count and spike count variance seems to be similar for all three SR groups. For low discharge rates, the variance is approximately proportional to the mean, but the ratio of variance to mean becomes smaller as the mean increases until the variance becomes approximately independent of the mean at high discharge rates. A similar decrease in the variance-to-mean ratio has been observed by Young and Barta (1985), and has been attributed to refractory effects. As in the data of Teich and Khanna (1985) for continuous tones, the variance is below that predicted on the basis of a Poisson model (straight line), even for low discharge rates. The continuous line is a least-squares fit to the data by means of a modification of the function relating mean and variance

in the dead-time model of Teich and Lachs (1979). Although the fitted curve does deviate from the data points, these deviations will not affect the computation of d' by more than a few percent.

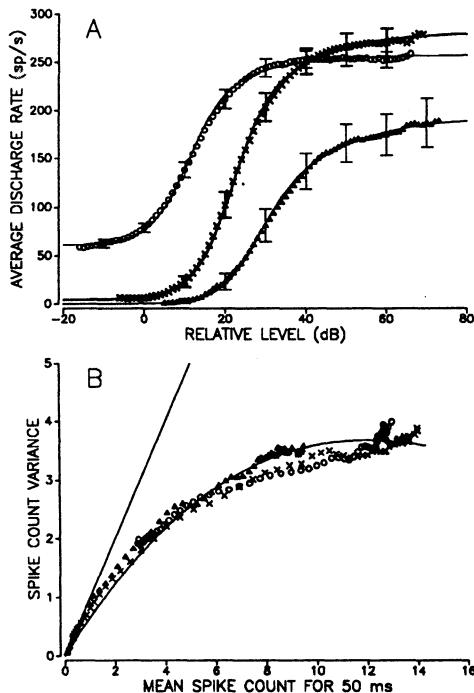


FIGURE 6,

A. Average rate-level functions in response to 50-ms tone-bursts at the CF for the same groups of auditory-nerve fibers as in Fig. 4. Prior to averaging, the rate-level functions were horizontally translated by the same number of dB's as described in Fig. 4. Vertical bars represent ± 2 standard errors. Continuous lines are fits to the average data by means of Sach's et al's (1986) model of rate-level functions.
 B. Averages of the spike-count variances plotted against averages of the mean count in response to 50-ms tone bursts at the CF for the same groups of auditory-nerve fibers as in A. The continuous curve is a least-squares fit to the data using a modification of Teich and Lach's (1979) dead-time model. The straight line shows behaviour expected for a Poisson model.

The smooth curves fitted to the data in Fig. 6 were used to predict intensity difference limens for three "model fibers", each representing one SR group. Specifically, the predicted DL at level L is the increment ΔL such that $d'(L - \Delta L/2, L + \Delta L/2)$ is equal to 1, corresponding to a probability criterion of 0.76 for the two-interval paradigm used in the DL measurements. Before computing d' , a variance-stabilizing transformation (Teich, 1985; Young and Barta, 1985) was applied to the spike counts to correct for the dependence of the variance on the mean.

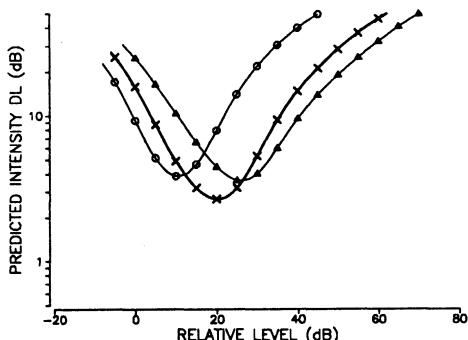


FIGURE 7. Predicted intensity DL's against relative level for 3 model fibers representing the high-SR (circles), medium-SR (crosses), and low-SR (triangles) fibers.

between model predictions and average DL's, the minimum DL's being larger, and the L_{min} lower in the model predictions. As a result of these discrepancies predicted performance is poorer than measured performance at high intensities, particularly for low-SR fibers. Despite these discrepancies, model predictions remain well within the variability of DL-intensity functions in single auditory-nerve fibers, so that the model can be considered sufficient to demonstrate how intensity information from different fibers might be combined. Possible reasons for the discrepancies are discussed in Sec. III.A.

B. Single-channel stage: Combining intensity information from fibers with different sensitivities.

Intensity information from fibers in three SR groups was combined by means of the optimal combination rule of signal-detection theory for independent, Gaussian random variables (Siebert, 1965, 1968; Colburn, 1981; Viemeister, 1986). The independence assumption is consistent with the results of Johnson and Kiang (1976) on recordings from pairs of auditory-nerve fibers. As a further simplification, performance characteristics of all fibers within each SR group were assumed to be identical except for a shift in sensitivity (L_{min}). These performance characteristics are completely specified by the DL-intensity functions of Fig. 7.

A first version of the model was intended to define performance achievable by optimal processing of intensity information available in the entire auditory nerve. Specifically, the population of model fibers consisted of 30,000 elements, of which 60% were high-SR fibers, 25% medium-SR fibers, and 15% low-SR fibers. This distribution

Figure 7 shows predicted DL's as a function of level for the three model fibers. These predictions are comparable to the average DL-intensity functions shown in Fig. 5 because the two measures are based on the same probability criterion and are derived from the same set of auditory-nerve fibers. The general shapes of the DL-intensity functions are similar for the model predictions and the average measurements. The main differences between SR groups that are apparent in the average data are also predicted by the model: The minimum DL for the medium-SR model fiber is lower than the minimal for the other two groups, and the 10-dB dynamic range is wider for the high-SR and low-SR model fibers than for the high-SR fiber. However, for all three SR groups, there are systematic discrepancies

is consistent with data from the cat auditory nerve (Liberman, 1978). The L_{min} distribution for each SR group was set to the corresponding empirical distribution in Fig. 4. At this single-channel stage of the model, we assume in effect that all CF regions are responding alike, so that model predictions are most appropriately compared with psychophysical data for "uniformly exciting" broadband noise. Psychophysical DL's for 50-ms bursts of broadband noise are about 1 dB for sensation levels ranging from 10 dB to at least 80 dB (Miller, 1947; Harris, 1963; Houtsma et al., 1980).* Intensity DL's predicted by this model are shown in Fig. 8 as a function of relative level, separately for each of the three SR groups, and for the entire population of 30,000 fibers. At low and moderate stimulus levels, predicted performance exceeds psychophysical performance by as much as a factor of 40, but model performance deteriorates with increasing level, until predicted performance approaches psychophysical performance near 75 dB. Comparison of the DL-intensity curves for each of the three SR groups with predictions for the entire population show that, even though high-SR fibers are the most numerous, they account for most of the overall performance only at low stimulus levels, while at moderate and high stimulus levels medium-SR and low-SR fibers contribute more strongly to model performance. This general behavior is similar to that of existing single-channel models that include a physiologically-realistic threshold distribution (Colburn, 1981; Winslow, 1985; Viemeister, 1986).

The degradation in performance of the optimum processor model at high intensities implies that, if predicted DL's were to obey Weber's law, the model would have to process more efficiently intensity information from low-SR fibers than information from high-SR and medium-SR fibers. In order to make the notion of processing efficiency more concrete, we varied the distribution of fibers into the 3 SR groups, and thereby tried to obtain a better approximation to Weber's law. Specifically, the model fibers in each of the 3 SR groups were divided into 3 bins along the L_{min} dimension, and the numbers of fibers in the 9 bins were adjusted by an optimization routine to minimize the least-squares deviation from Weber's law.** For each bin, the ratio of the numbers of fibers selected by the optimization routine to the numbers that would be present in the auditory nerve gives a measure of processing efficiency. Figure 9 shows psychophysical DL's and model predictions for a set of 37 fibers whose distribution was selected by the optimization routine. Model predictions are fairly close to psychophysical performance over at least a 60 dB range of low and moderate levels. For levels above 50 dB, predicted performance deteriorates sharply because there are no model fibers whose minimum DL occurs in that range. The fiber distribution selected by the optimizer included 32% of high-SR fibers, 22% of medium-SR fibers, and 46% of low-S fibers. This distribution differs markedly from that found by Liberman (1978) in the cat auditory nerve: The proportion of

* In estimating DL's for 50-ms noise bursts, we used the result that, for tones, a tenfold increase in duration results approximately in a twofold decrease in DL (Florentine, 1986).

** The intervals between bins were 10 dB, and the bins for different SR groups overlapped, so that the total range of L_{min} for the 9 bins was from 0 to 50 dB. Compare with Fig. 4.

low-SR fibers is about 3 times greater in the selected distribution than in the nerve, while the proportion of high-SR is only about half. Among low-SR fibers, the percentage of elements with L_{min} above 40 dB is 6 times greater in the selected distribution than in the empirical distribution of Fig. 4.

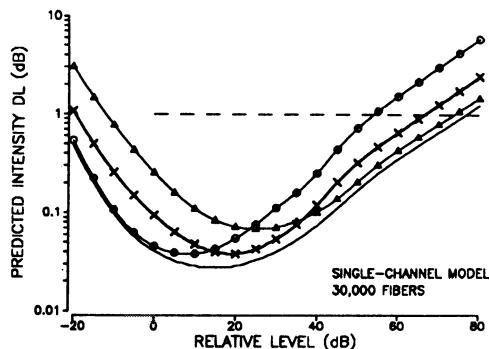


FIGURE 8, Predicted intensity DL against relative level for a 30,000 fiber, single-channel model with realistic distribution of the fibers into the three SR groups (no symbols), and for each of the SR groups separately, using the same symbols as in Fig. 7.

broadband noise bursts. These data were first converted to DL's by means of the formula $DL=10/\delta^l$, then multiplied by 2 to account for the difference in duration between the 50-ms stimuli used in the physiological experiments and the 500-ms stimuli used by Houtsma et al. The horizontal position of the psychophysical data is arbitrary.

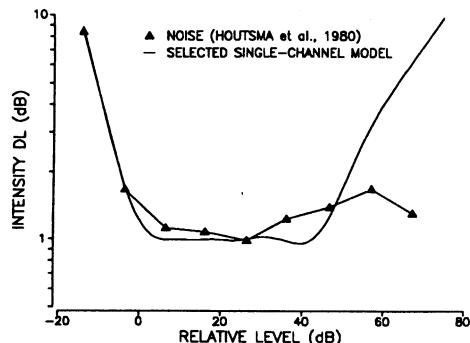


FIGURE 9, Predicted intensity DL's against relative level for a single-channel model in which the distribution of 37 fibers into the 3 SR groups was selected to obtain an optimum fit to Weber's law for relative levels between 5 and 45 dB. Filled triangles are the psychophysical data of Houtsma et al. (1980) for intensity discrimination of

C. Multiple-channel model: Combining intensity information across fibers with different CF's

In order to obtain model predictions for stimuli other than uniformly exciting noise, it is necessary to introduce multiple frequency-selective channels, each one representing the activity of auditory-nerve fibers that innervate a restricted portion of the cochlea. For this purpose, the functional description of rate-level functions (Fig. 6) was preceded in each channel by a linear bandpass filter whose frequency response was obtained by averaging the tuning curves of many auditory-nerve fibers (Liberman, 1978; Delgutte, 1986). Specifically, we used 76 channels whose CF's ranged from 0.1 to 20 kHz, with a spacing between channels corresponding to 1% of the length of the cat cochlea (Liberman, 1982). For combining intensity information across fibers, we used the same optimal rule for independent Gaussian variables that was used in the single-channel

stage. The numbers of fibers per channel, the distributions of fibers into the three SR groups, and the L_{min} distributions in each SR group were assumed to be identical for all channels.

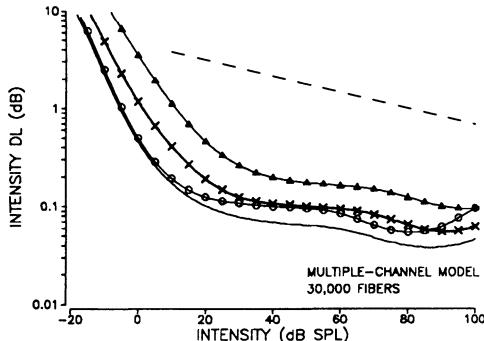


FIGURE 10, Predicted intensity DL's for a 1-kHz tone against relative level for a 30,000 fiber, multiple-channel model with realistic distribution of the fibers into 3 SR groups (no symbols), and for each SR groups separately. The dashed line indicates the approximate level dependence of the psychophysical DL's for a 1 kHz-tone.

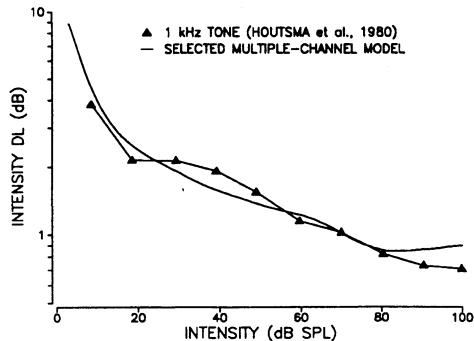


FIGURE 11, Predicted intensity DL's for a 1-kHz tone against stimulus level for a 75 fiber, multiple-channel model with the same distribution of the fibers into the 3 SR groups as in Fig. 9. Filled triangles show psychophysical DL's for 1-kHz tones (Houtsma et al., 1980), corrected for differences in duration as described in Fig. 9.

Figure 10 shows the predicted DL as a function of intensity of a 1-kHz tone for a 30,000-fiber, multiple-channel model with the same physiologically-realistic distribution of fibers into the three SR groups as in the single-channel stage of Fig. 8. Model predictions are shown separately for each of the 3 SR groups, and for the entire population of fibers. Throughout the range of levels, predicted DL's for the 30,000-fiber model are well below psychophysical DL's for 50-ms tones at 1 kHz (dashed lines). In fact, predicted performance for high-SR fibers alone (circles) well exceeds psychophysical performance, and is roughly constant over a wide range of intensities. In no range of intensities do low-SR fibers (triangles) contribute significantly to overall performance. This result, which contrasts with those for the single-channel stage (Fig. 8), is due to the fact that channels whose CF's are increasingly far from 1 kHz begin to respond as level increases, so that at each level there are many unsaturated, high-SR fibers. As expected, the behavior of the multiple-channel model for high-SR fibers resembles that of Siebert's (1968) model.

Figure 11 shows predictions of a 75-fiber* multiple channel for which the distribution of fibers into SR groups was selected to approximate Weber's law in each frequency channel as in Fig. 9. The figure also shows psychophysical DL's for 1-kHz tones (Houtsma et al., 1980). Model predictions are fairly close to psychophysical data throughout the range of levels, and in particular show a systematic improvement in performance as intensity increases. The change from Weber-law behavior in the single-channel model of Fig. 9 to the "near-miss" in the multiple-channel model of Fig. 11 is due to the increase in the number of active channels with intensity, as in the excitation-pattern model of Florentine and Buus (1981).

III. DISCUSSION

A. Relation of the DL measurements to rate-level functions

Pure-tone intensity difference limens were measured in auditory-nerve fibers by means of an adaptive procedure which adjusts discrimination performance to a constant probability criterion. Variations in measured intensity DL's with stimulus level were as expected from the shape of rate-level functions in auditory-nerve fibers: performance was best in the region where discharge rate increases rapidly with stimulus level, and poor both below threshold and in the saturation region. In order to examine quantitatively the relation between measured DL's and rate-level functions, we compared averages DL's for groups of auditory-nerve fibers with predictions from a Gaussian model of spike count statistics derived from the same set of fibers. Although the model correctly predicted the overall shapes of the DL-intensity functions, predicted performance was poorer than measured performance, particularly for high stimulus levels. While we only showed results for averages over fibers, similar discrepancies are found when this analysis is applied to individual auditory-nerve fibers.

One possible explanation for these discrepancies is the assumption that the statistics of the spike counts are Gaussian. Unlike Gaussian distributions, spike count distributions are fundamentally discrete and usually asymmetric (Teich and Khanna, 1985). The validity of the Gaussian approximation was examined by Viemeister (1986). He computed receiver-operating-characteristic (ROC) curves for pairs of spike count distributions measured from auditory-nerve fibers by Teich and Khanna (1985). Viemeister compared the ROC curves for tones at the CF with different intensities with predictions of a Gaussian model in which the means were set to the actual means of the two spike count distributions, and the variance to the geometrical mean of the variances for the two distributions. With the possible exception of low-SR fibers at low stimulus levels, the ROC curves were closely fit by the Gaussian model. Because ROC curves include as a special case the constant-probability criterion used in our DL measurements, Viemeister's analysis suggests that the Gaussian approximation is not

*Because there is only a total number of 75 fibers for 76 frequency channels, we used fractional numbers of fibers in each channel. This is not a problem if the relative numbers of fibers are thought of as measures of processing efficiency.

likely to be responsible for the discrepancies found at high stimulus levels between the predicted and measured DL's.

Another possible reason for the discrepancy between predicted and measured DL's is the difference in temporal parameters of stimulation between the DL paradigm and the estimation of spike-count statistics. It is well known that rate-level functions of auditory-nerve fibers depend on the repetition rate of the stimuli, the order of stimulus levels, and previous stimulation by intense sounds (Kiang et al., 1965; Sachs and Abbas, 1974; Young and Sachs, 1973). One may speculate that prolonged stimulation by intense stimuli during the DL measurements might produce a temporary shift of the fibers' dynamic ranges toward higher intensities. This effect would not be as pronounced for rate-level functions because a DL measurement requires many more tone-burst presentations at each stimulus level than estimation of mean and variance of the spike count. Consistent with this interpretation, discharge rate produced by the reference stimulus in the DL paradigm sometimes started to decrease at high levels, although of course for each reference level (L) the discharge rate produced by the adjusted stimulus ($L + \Delta L$) was always greater than the rate produced by the reference. If this interpretation is correct, one would have to use exactly the same temporal parameters of stimulation in physiological experiments and psychophysical experiments in order to provide comparable characterizations of performance.

B. Comparison of model predictions with psychophysical data

Our model for combining intensity information from populations of auditory-nerve fibers is based on the optimum combination rule of detection theory (Siebert, 1965, 1968). This model is the first to include both multiple frequency channels (Siebert, 1965, 1968; Teich and Lachs, 1979; Lachs et al., 1984), and a realistic distribution of auditory-nerve fibers into groups defined on the basis of spontaneous discharge rate (Colburn, 1981; Winslow, 1985; Viemeister, 1986). A major result is that intensity-discrimination performance predicted by the 30,000-fiber model well exceeds psychophysical performance for both tone and noise stimuli through a wide range of stimulus levels. Before drawing conclusions from this result, it is important to review the limitations of the present study.

One type of limitation is that the model does not simulate the responses of auditory-nerve fibers to tone bursts in anesthetized cats with complete accuracy. For example, it does not simulate the sharp "notches" that occur at stimulus levels above 85 dB SPL in the rate-level functions of many auditory-nerve fibers (Kiang et al., 1969; Liberman and Kiang, 1984). Such rapid variations in discharge rates would be expected to greatly improve performance of the optimum-processor model at high stimulus levels.

Another type of difficulty stems from the fact that predictions of the single-channel stage, which are derived from data for tones at the CF, were compared with psychophysical data for broadband noise. Rate-level functions for broadband noise have shallower slopes and wider dynamic ranges than those for tones at the CF (Ruggero, 1973; Schalk and Sachs, 1980). These differences have been attributed to suppression by components of the noise on either side of the CF

(Schalk and Sachs, 1980). Pilot measurements of single-fiber intensity DL's for broadband noise confirm that the growth of the DL beyond its minimum is shallower for noise than for tones, which would improve performance of the single-channel stage at high intensities.

A third type of limitation is that the model is based on physiological data obtained in cats anesthetized with barbiturates, while psychophysical data were obtained from awake humans. Little is known about variations of intensity DL's with sound level in cats, although DL's at moderate levels are not very different from those of humans (Rosenzweig, 1946; Elliott and McGee, 1965; Oesterreich et al., 1971). Anesthesia might have direct effects on auditory-nerve fiber activity, or might affect responses of auditory-nerve fibers by means of its action on olivocochlear efferents. Although medial olivocochlear efferents are known to be active in cats anesthetized with barbiturates (Liberman and Brown, 1986), their rates of discharge are lower than those required to obtain large effects on auditory-nerve fibers by means of electrical stimulation of the crossed olivocochlear bundle (Wiederhold and Kiang, 1970). To a first approximation, the effect of such stimulation is to shift the rate-level functions of auditory-nerve fibers towards higher intensities (Wiederhold, 1970; Gifford and Guinan, 1983). Winslow (1985) has shown that changes in rate-level functions for tones in noise resulting from electrical stimulation of the efferents considerably improve intensity-discrimination performance predicted by his model.

In reviewing these limitations, it is apparent that a more realistic model, which would include dips in rate-level functions, suppression for noise stimuli, and effects of the olivocochlear efferents, would be expected to perform better at high stimulus levels. Therefore, the conclusion seems inescapable that there is more than enough information in the discharge rates (spike counts) of auditory-nerve fibers to account for psychophysical performance in intensity discrimination. In other words, psychophysical performance is not limited by saturation of auditory-nerve fibers, but by the processing efficiency of more central stages of the auditory system.

Another result of the present study is that the level dependence of the DL predicted by a 30,000-fiber model with a realistic distribution of fibers into the three SR groups clearly differs from psychophysical data for both tone and noise stimuli. Specifically, the model predicts nearly constant DL's for tones (Weber's law) over a wide intensity range (Fig. 10), while psychophysical DL's decrease with increasing intensity. The discrepancy is even more apparent for broadband noise (Fig. 8), where model performance severely degrades with level to approach psychophysical performance at high intensities. These results suggest that the central processor must process information from low-SR fibers more efficiently than that from high-SR and medium-SR fibers. The relative processing efficiencies for each SR group were estimated by adjusting the fiber distribution in order to approximate Weber's law in single frequency channels. Results showed that an appropriate distribution of a fraction of 1% of the available auditory-nerve fibers sufficed to account for psychophysical performance for both tone and noise stimuli over a broad range of stimulus levels. The proportion of low-SR fibers was about 3 times greater in the selected distribution than in the cat auditory nerve, with low-SR fibers with highest L_{min} receiving even greater weights.

In this respect, it is interesting that low-SR auditory-nerve fibers apparently branch more profusely and form more endings in the cochlear nucleus than do high-SR fibers (Fekete et al., 1984). Although branching without changes in the discharge patterns does not alter the performance of the optimum processor, it can affect certain types of suboptimal processors such as one that would sum all the incoming spikes. In general, one should not expect the proportions of fibers in the different SR groups to be optimum for a specialized psychophysical task such as intensity discrimination. The actual proportions are more likely to result from many evolutionary pressures, only one of which may be the need to resolve small changes in intensity.

C. Implications for speech processing

The preceding modeling results can be extended to provide stable representations of the spectra of speech sounds in terms of average discharge rates of auditory-nerve fibers. Key variables in the model of intensity discrimination are the discriminability measures d' for the optimum combinations of fibers in individual frequency channels. By taking the sum from threshold to level L of the d' for small, adjacent intervals ΔL , we define a new variable $D(L)$, the "cumulative d' ", which has the property that equal intervals along this scale are equally discriminable. In other words, $D(L)$ represents the total number of intensity DL 's between absolute threshold and level L for a particular frequency channel. Because d' depends only on the discharge rates (spike counts) of the fibers in one frequency channel, the cumulative d' is also a function of discharge rates. Expressing the cumulative d' as a function of channel CF for a particular stimulus provides a representation of the spectrum of this stimulus.

Figure 12B shows cumulative d' as a function of CF for the synthetic /ɛ/ vowel whose spectrum is shown in Fig. 12A. The cumulative d' is shown for 4 stimulus levels ranging from 30 to 90 dB SPL. This measure was constructed from the version of the model in which the distribution of the fibers into the three SR groups was selected to approximate Weber's law in single frequency channels (Fig. 9). Because the DL is nearly constant over a broad range of levels, the cumulative d' grows linearly with level over that range, and the resulting representations of the vowel spectrum for the lower 3 levels in Fig. 12 are nearly the same except for a vertical translation. At the highest level (90 dB SPL), the representation of the formant frequencies is degraded by the saturation of the cumulative d' , as expected from the rapid increase in DL in Fig. 9.

In summary, the multiple-channel model of intensity discrimination leads naturally to a speech processing scheme that provides a stable representation of the stimulus spectrum over a broad range of stimulus levels. This is a "rate-place" scheme because it expresses as a function of CF a response measure derived from the discharge rates of auditory-nerve fibers that innervate a restricted portion of the cochlea. Although the present scheme does show some degradation of the spectrum at very high stimulus levels, such degradation might be less severe in a more realistic model that would include effects of suppression and efferent feedback to the cochlea. The present results indicate that there may be sufficient information in the discharge rates of auditory-nerve fibers to discriminate between

vowels. This conclusion does not preclude schemes based on fine temporal patterns of discharge (Young and Sachs, 1979; Sinex and Geisler, 1983; Delgutte, 1984; Shamma, 1985) for the auditory processing of speech under certain conditions. However, the present rate-place scheme provides a spectral representation of all classes of speech sounds, while temporal schemes are ineffective for sounds with intense high-frequency components such as fricatives and the bursts of stop consonants (Delgutte and Kiang, 1984). Physiological studies of central auditory pathways and comparisons of the predictions of explicit models with a wide range of psychophysical data will be needed to determine which processing schemes are used under specific conditions.

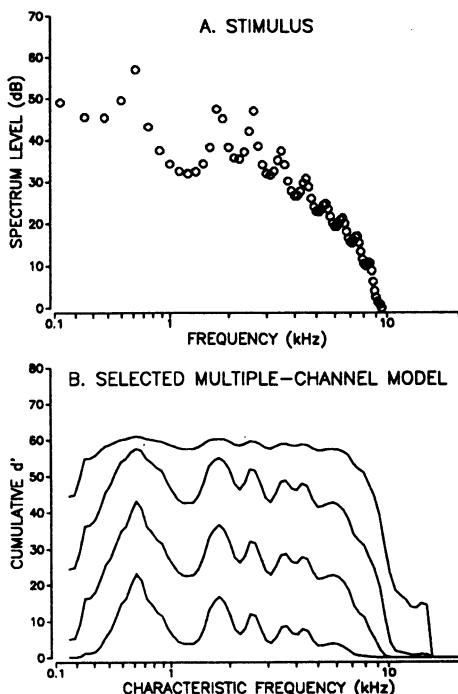


FIGURE 12,

A. Spectrum of a synthetic /ɛ/ vowel with a fundamental frequency of 125 Hz.

B. Cumulative d' in response to the vowel as a function of channel CF for a multiple-channel model in which the distribution of fibers into the three SR groups was selected as in Fig. 9 to approximate Weber's law in individual frequency channels. The cumulative d' is shown for 4 stimulus levels: 30, 50, 70, and 90 dB SPL.

REFERENCES

1. Bos, C.E. and de Boer, E. (1966). Masking and discrimination. *J. Acoust. Soc. Am.*, 39, 708-715.
2. Colburn, H.S. (1981). Intensity perception: Relations of intensity discrimination to auditory-nerve firing patterns. Internal Memorandum, Res. Lab. Electron., M.I.T., Cambridge, MA.
3. Colburn, H.S. (1984). Models of intensity discrimination. *J. Acoust. Soc. Am.*, 76, S5.
4. Delgutte, (1984). Speech coding in the auditory nerve II: Processing schemes for vowel-like sounds. *J. Acoust. Soc. Am.*, 75, 879-886.
5. Delgutte, B. (1986). Analysis of French stop consonants using a model of the peripheral auditory system. In: J.S. Perkell and D.H. Klatt (Eds.), Invariance and Variability of Speech Processes, 163-177. Erlbaum, Hillsdale, NJ.
6. Delgutte, B. and Kiang, N.Y.S. (1984b). Speech coding in the auditory nerve III. Voiceless fricative consonants. *J. Acoust. Soc. Am.*, 75, 887-896.
7. Elliot, D.N. and McGee, T.M. (1965). Effects of cochlear lesions upon audiograms and intensity discrimination in cats. *Ann. Otol. Rhinol. Laryngol.*, 74, 386-408.
8. Evans, E.F. and Palmer, A.R. (1980). Relationship between the dynamic range of cochlear nerve fibers and their spontaneous activity. *Exp. Brain Res.*, 40, 115-118.
9. Fekete, D.M., Rouiller, E.M., Liberman, M.C., and Ryugo, D.K. (1984). The central projections of intracellularly labeled auditory nerve fibers in cats. *J. Comp. Neurol.*, 229, 432-450.
10. Florentine, M. (1986). Level discrimination of tones as a function of duration. *J. Acoust. Soc. Am.*, 79, 792-798.
11. Florentine, M. and Buus, S. (1981). An excitation-pattern model for intensity discrimination. *J. Acoust. Soc. Am.*, 70, 1646-1654.
12. Geisler, C.D., Deng, L., and Greenberg, S.R. (1985). Thresholds for primary auditory fibers using statistically defined criteria. *J. Acoust. Soc. Am.*, 77, 1102-1109.
13. Gifford, M.L. and Guinan, J.J., Jr. (1983). Effects of crossed-olivocochlear-bundle stimulation on cat auditory nerve fiber responses to tones. *J. Acoust. Soc. Am.*, 74, 115-123.
14. Harris, J.D. (1963). Loudness discrimination. *J. Speech Hear. Disord. Mon. Suppl.*, 11, 1-63.
15. Houtsma, A.J.M., Durlach, N.I., and Braida, L.D. (1980). Intensity perception XI. Experimental results on the relation of intensity resolution to loudness matching. *J. Acoust. Soc. Am.*, 68, 807-813.
16. Jesteadt, W., Wier, C.C., and Green, D.M. (1977). Intensity discrimination as a function of frequency and sensation level. *J. Acoust. Soc. Am.*, 61, 160-177.
17. Johnson, D.H. and Kiang, N.Y.S. (1976). Analysis of discharges recorded simultaneously from pairs of auditory-nerve fibers. *Biophys. J.*, 16, 719-734.
18. Kiang, N.Y.S., Baer, T., Marr, E.M., and Demont, D. (1969). Discharge rates of single auditory-nerve fibers as a function of level. *J. Acoust. Soc. Am.*, 46, S106.
19. Kiang, N.Y.S., Watanabe, T., Thomas, E.C., and Clark, L.F. (1965). Discharge patterns of single fibers in the cat's auditory nerve. Research Monograph #35. (MIT Press, Cambridge, MA).

20. Lachs, G., Al-Shaik, R., Bi, Q., Saia, R. and Teich, M.C. (1984). A neural-counting model based on physiological characteristics of the peripheral auditory system V. Application to loudness estimation and intensity discrimination. *IEEE Trans. Syst. Man Cyber.* SMC-14, 819- 836.
21. Liberman, M.C. (1978). Auditory-nerve response from cats raised in a low-noise chamber. *J. Acoust. Soc. Am.*, 63, 442-455.
22. Liberman, M.C. (1982). The cochlear frequency map for the cat: Labeling auditory-nerve fibers of known characteristic frequency. *J. Acoust. Soc. Am.*, 72, 1441-1449.
23. Liberman, M.C. and Brown, M.C. (1986). Physiology and anatomy of single olivocochlear neurons in the cat. *Hear Res.*, 24, 17-36.
24. Liberman, M.C. and Kiang, N.Y.S. (1976). Acoustic trauma in cats: Cochlear pathology and auditory-nerve activity. *Acta Otolaryngol. Suppl.* 358.
25. Liberman, M.C. and Kiang, N.Y.S. (1984). Single-neuron labeling and cochlear pathology IV: Stereocilia damage and alterations in rate- and phase-level functions. *Hearing Res.*, 16, 75-90.
26. McGill, W.J. and Goldberg, J.P. (1968). Pure-tone intensity discrimination and energy detection. *J. Acoust. Soc. Am.*, 44, 576-581.
27. Miller, G.A. (1947). Sensitivity to changes in the intensity of white noise and its relation to masking and loudness. *J. Acoust. Soc. Am.*, 19, 609-619.
28. Miller, M.I. and Sachs, M.B. (1983). Representation of stop consonants in the discharge patterns of auditory-nerve fibers. *J. Acoust. Soc. Am.*, 74, 502-517.
29. Moore, B.C.J. and Raab, D.H. (1974). Pure-tone intensity discrimination: Some experiments relating to the near-miss to Weber's law. *J. Acoust. Soc. Am.*, 55, 1049-1054.
30. Oesterreich, R.E., Strominger, N.L., and Neff, W.D. (1971). Neural structures mediating differential sound intensity discrimination in the cat. *Brain Res.*, 27, 251-270.
31. Penner, M.J. and Viemeister, N.F. (1973). Intensity discrimination of clicks: The effects of click bandwidth and background noise. *J. Acoust. Soc. Am.*, 54, 1184-1188.
32. Raab, D.H. and Goldberg, I.A. (1975). Auditory intensity discrimination with bursts of reproducible noise. *J. Acoust. Soc. Am.*, 57, 437-447.
33. Rabinowitz, W.M., Lim, J.S., Braida, L.D., and Durlach, N.L. (1976). Intensity perception VI. Summary of recent data on deviations from Weber's law for 1000-Hz tone pulses. *J. Acoust. Soc. Am.*, 59, 1506- 1509.
34. Rosenzweig, M. (1946). Discrimination of auditory intensities. *Am. J. Psychol.*, 59, 127-136.
35. Ruggero, M. (1973). Response to noise of auditory-nerve fibers in the squirrel monkey. *J. Neurophysiol.*, 36, 569-587.
36. Sachs, M.B. and Abbas, P.J. (1976). Rate versus level functions of auditory-nerve fibers in cats: Tone-burst stimuli. *J. Acoust. Soc. Am.*, 56, 1835-1847.
37. Sachs, M.B. and Young, E.D. (1979). Encoding of steady-state vowels in the discharge patterns of auditory-nerve fibers: Representation in terms of discharge rate. *J. Acoust. Soc. Am.*, 66, 1381-1403.
38. Sachs, M.B., Voigt, H.F., and Young, E.D. (1983). Auditory nerve representation of vowels in background noise. *J. Neurophysiol.* 50, 27-45.

39. Sachs, M.B., Winslow, R.L., and Kozikowski, J.G. (1986). Model for auditory-nerve rate-level functions with sloping saturation. Abstr. 9th Midwinter Res. Meet. Assoc. Res. Otolaryngol., 62.
40. Sanderson, A.C. (1975). Discrimination of neural coding parameters in the auditory system. IEEE Trans. Syst. Man Cyber., SMC-5, 533-342.
41. Schalk, T.B. and Sachs, M.B. (1980). Nonlinearities in auditory-nerve fiber responses to bandlimited noise. J. Acoust. Soc. Am., 67, 903- 913.
42. Shamma, S.A. (1985). Speech processing in the auditory system II: Lateral inhibition and the central processing of speech evoked activity in the auditory nerve. J. Acoust. Soc. Am., 78, 1622-1632.
43. Siebert, W.M. (1965). Some implications of the stochastic behavior of primary auditory neurons. Kybernetik, 2, 206-215.
44. Siebert, W.M. (1968). Stimulus transformations in the peripheral auditory system. In: P.A. Kollers and M. Eden (Eds.), Recognizing Patterns, 104-133. MIT Press, Cambridge, MA.
45. Sinex, D.G. and Geisler, C.D. (1983). Responses of auditory-nerve fibers to consonant-vowel syllables. J. Acoust. Soc. Am., 73, 602-615.
46. Taylor, M.M. and Creelman, C.D. (1967). PEST: Efficient estimates of probability functions. J. Acoust. Soc. Am., 41, 782-787.
47. Teich, M.C. (1985). Normalizing transformations for dead-time modified Poisson counting distributions. Biol. Cybern., 53, 121-124.
48. Teich, M.C. and Khanna, S.M. (1985). Pulse-number distribution for the neural spike train in the cat's auditory nerve. J. Acoust. Soc. Am., 77, 1110-1128.
49. Teich, M.C. and Lachs, G. (1979). A neural-counting model incorporating refractoriness and spread of excitation. I. Application to intensity discrimination. J. Acoust. Soc. Am., 66, 1738-1749.
50. Viemeister, N.F. (1972). Intensity discrimination of pulsed sinusoids: Effects of filtered noise. J. Acoust. Soc. Am., 51, 1265-1269.
51. Viemeister, N.F. (1983). Auditory intensity discrimination at high frequencies in the presence of noise. Science, 221, 1206-1208.
52. Viemeister, N.F. (1986). Psychophysical aspects of auditory intensity coding. To appear in Functions of the Auditory System.
53. Wiederhold, M.L. (1970). Variations in the effects of electric stimulation of the crossed olivocochlear bundle on the cat single auditory-nerve fiber responses to tone bursts. J. Acoust. Soc. Am., 48, 966-977.
54. Wiederhold, M.L. and Kiang, N.Y.S. (1970). Effects of electric stimulation of the crossed olivocochlear bundle on single auditory- nerve fibers in the cat. J. Acoust. Soc. Am., 48, 950-965.
55. Winslow, R.W. (1985). A quantitative analysis of rate coding in the auditory nerve. Ph.D. Thesis, Johns Hopkins U.
56. Winslow, R.W. and Sachs, M.B. (1985a). Intensity discrimination based on auditory-nerve fiber rate-level functions. Abstr. 8th Midwinter Res. Meet. Assoc. Res. Otolaryngol., 53.
57. Young, E.D. and Barta, P.E. (1985). Rate responses of auditory-nerve fibers to tones in noise near masked threshold. J. Acoust. Soc. Am., 79, 426-442.
58. Young, E.D. and Sachs, M.B. (1973). Recovery from noise exposure in auditory-nerve fibers. J. Acoust. Soc. Am., 54, 1535-1543.

59. Young, E.D. and Sachs, M.B. (1979). Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers. J. Acoust. Soc. Am., 66, 1381- 1403.

ORGANIZATION OF THE COCHLEAR NUCLEUS FOR INFORMATION PROCESSING*

Eric D. Young
The Johns Hopkins University
Baltimore, Maryland, USA

PERIPHERAL REPRESENTATIONS OF COMPLEX STIMULI

The representation of complex stimuli has been studied in detail at the level of the auditory nerve. Although many aspects of stimulus coding in the auditory nerve remain to be worked out, the general outlines seem clear (see Sachs, 1984, for a review). It is generally accepted that the basic organizing principle of the auditory system is tonotopicity (Imig et al., 1982; Merzenich et al., 1982), so the representation of a complex stimulus can be considered as a profile of neural activity against a tonotopic axis. The tonotopic axis is created by the frequency analysis of the basilar membrane and is represented neurally by the tuning of neural elements. The peripheral tonotopic axis is preserved in the central nervous system, in that each central auditory nucleus contains an orderly tonotopic array in which the best frequencies of cells are laid out in a monotonic fashion (Merzenich et al., 1982).

One question that is not resolved concerns the nature of the neural response on which the tonotopic profile is based. Figure 1 shows a summary of three types of neural profiles obtained from auditory nerve fibers. These illustrate the differences between response profiles based on different response measures. The stimuli are steady state, synthetic vowels: /ɛ/ in the left column and /a/ in the right column.

The dashed lines show profiles based on the average discharge rates of spontaneously active (≥ 1 spike/sec.) fibers. Except at low sound levels, these give a relatively poor representation of spectral details like formant peaks and the troughs between formants. For the examples shown in Figure 1, the dashed rate profiles contain clear representations of the formant peaks of the vowels only at the lowest levels shown (38 and 37 dB); at higher levels, the profiles saturate and clear peaks corresponding to the formant frequencies are not seen (Sachs and Young, 1979). A similar flattening of rate profiles occurs in the presence of background noise (Sachs et al., 1983; Delgutte and Kiang, 1984). Despite the poor representation of formant peaks in the spontaneously active fibers' rate profiles, their overall shapes still contain sufficient information to allow vowel identification using

*The work reported in this paper was done by or in collaboration with Carol C. Blackburn, Jeanne-Marie E. Robert, Murray B. Sachs, William P. Shofner, and John A. White. Preparation of this paper was supported by NIH grants R01-NS12524 and RO1-NS12112.

template matching procedures (Winslow, 1985). In Figure 1, there are clear differences between the dashed profiles for /ɛ/ and /a/, even at the highest level, in that the profiles for /ɛ/ extend to higher frequencies.

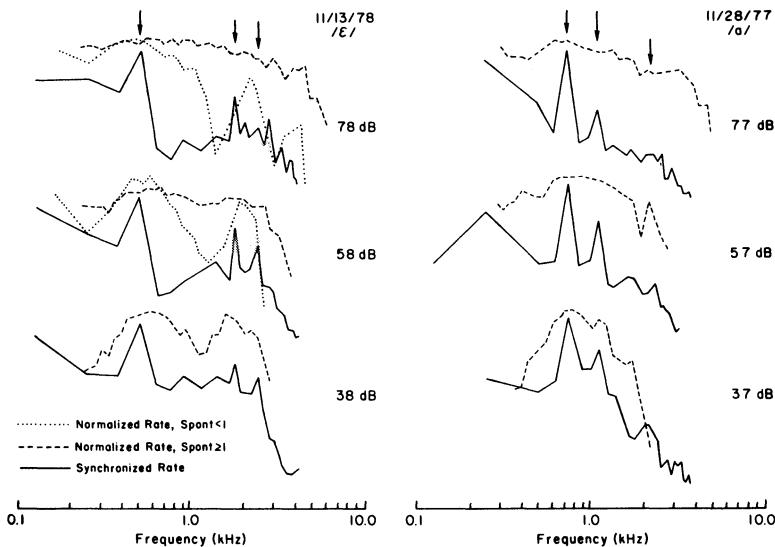


FIGURE 1. Neural response profiles based on normalized discharge rate (dashed and dotted lines) and on a temporal response measure (solid lines). Rate profiles shown with dashed lines computed from high spontaneous rate units (≥ 1 spike/sec.). Rate profiles shown with dotted lines computed from low spontaneous rate units (< 1 spike/sec.). Each curve shows logarithm of response plotted against best frequency. Stimulus is steady state, synthetic /ɛ/ in left column and /a/ in right column. Arrows point to formant frequencies. Response profiles for three stimulus levels are shown. Details of computation of rate and temporal profiles can be found in Sachs and Young (1979) and Young and Sachs (1979).

The dotted lines in the 58 and 78 dB plots for /ɛ/ show rate profiles for low spontaneous rate (< 1 spike/sec.) auditory nerve fibers. A low spontaneous rate profile is not shown at 38 dB because there was little response in low spontaneous rate fibers at this level. Low spontaneous rate fibers constitute 10-20% of the total population and are characterized by higher thresholds and wider dynamic ranges than high spontaneous rate fibers (Liberman, 1978; Schalk and Sachs, 1980). They do not saturate as readily as high spontaneous rate fibers. The dotted profiles in Figure 1 show that a good indication of the formant frequencies is conveyed by the low spontaneous rate population at levels up to at least 80 dB (Sachs and Young, 1979). A good representation of the spectra of vowels can be obtained across sound levels by using rate profiles in which information from low and high

spontaneous rate fibers is combined, with high spontaneous rate fibers providing information at low sound levels and low spontaneous rate fibers providing information at high levels (Delgutte, 1982).

It is easy to see how conventional mechanisms of synaptic transmission and neural integration can be used to process rate information in the central nervous system. In that sense, rate profiles are the simplest form of neural representation. However, the data in Figure 1 show that rate profiles are sensitive to stimulus conditions, so that central mechanisms that use only rate information would have to compensate for stimulus level by shifting the weighting of different populations of auditory nerve fibers as stimulus conditions change (Delgutte, 1982; Winslow, 1985).

Temporal profiles (solid lines) are based on the strength of phase-locking. Temporal profiles give a better, more detailed, representation of stimulus spectrum than do rate profiles and give an especially clear indication of formant frequencies. The representation is stable over a wide range of sound levels and in the presence of background noise (Young and Sachs, 1979; Delgutte and Kiang, 1984; Sachs et al., 1983). If phase-locking information could be used by the central nervous system, it would provide an excellent basis for complex signal perception. However, the means by which temporal information could be extracted by the central nervous system are not clear.

It is certainly possible that temporal information is used, because binaural processing for sound localization requires a temporal analysis similar to that needed to extract temporal information about stimulus spectrum (Merzenich et al., 1980). The binaural system is capable of comparing the phase of tonal stimuli in the two ears for use in determining binaural time-of-arrival differences. This analysis extends to frequencies of 1.5-2.0 kHz (Mills, 1972; Yost, 1974), which is somewhat below the upper frequency limit needed for the second and third formants of speech. Nevertheless, the existence of precise interaural phase analysis in the superior olive and inferior colliculus (Goldberg and Brown, 1969; Yin and Kuwada, 1983) suggests that it is possible for the central nervous system to perform a similar monaural analysis which would result in information about stimulus spectrum. However, the lowpass characteristics of nerve membrane (see below) and the jitter associated with synaptic transmission limit the ability of the nervous system to transmit temporally-coded information. Thus it seems likely that temporally-coded information would be extracted from auditory nerve discharge patterns and converted to rate-place information at an early level of the central nervous system.

It is clear from the discussion above that there are several possible neural codes for stimulus spectrum at the level of the auditory nerve. One approach to the question of how auditory-nerve information is used by the central nervous system is to extend the analysis of the representation of complex stimuli into the central nervous system. Because this approach ultimately involves the use of animal nervous systems to infer properties of the human nervous system, it is based on the assumption that the auditory systems of mammals share a common set of basic auditory analyses which must be done in the early stages of central processing. One such analysis is binaural stimulus comparison, which is necessary for sound localization

and related perceptual abilities. Other such peripheral analyses can be postulated, such as adaptation to noise backgrounds (Gibson et al., 1985), amplification of responses to small stimulus changes (Frisina et al., 1985), and sharpening the representation of stimulus spectra. The last of these, of course, is relevant to the issue of stimulus coding in the auditory nerve. If a sharpening of stimulus representation can be demonstrated in the CNS and the mechanisms by which the sharpening is achieved can be worked out, then the question of the nature of peripheral auditory stimulus representation will have been answered. If temporal aspects of auditory nerve responses are important, then it should be possible to demonstrate the extraction of temporal information and its conversion into rate-place information early in the central nervous system.

NEURAL SUBSYSTEMS OF THE VENTRAL COCHLEAR NUCLEUS

A problem that arises in the attempt to extend the analysis of stimulus representation into the central nervous system is that one is no longer dealing with a single, homogeneous population of neurons. In the cochlear nucleus (CN), which is the first central auditory processing center, there are at least five separate systems of neurons operating in parallel. These systems consist of neurons with different morphologies, which receive inputs from the auditory nerve and elsewhere via different circuits; they project their axons onto higher auditory centers with different patterns (for recent reviews, see Warr, 1982; Cant and Morest, 1984). The differences between the subsystems making up the CN are sufficient to suggest that each subsystem is specialized to perform particular aspects of the overall auditory perceptual task. Given the plausibility of this hypothesis, it is clear that an analysis of information representation and processing in the CN must be done separately for each of the neural subsystems that it contains.

This paper describes the current basis for the study of information processing in two of the major subsystems of the CN: the bushy cells and the stellate cells of the ventral cochlear nucleus (VCN). Neurons of these two classes make up most of the anterior division and part of the posterior division of the VCN and appear to be specialized for different aspects of auditory stimulus processing.

Figure 2 illustrates the morphology and auditory nerve inputs of stellate and bushy cells. Stellate cells have long dendritic trees. Auditory nerve inputs form small bouton terminals on the dendrites and sometimes also on the somata of stellate cells (Cant and Morest, 1979a; Cant, 1981; Tolbert and Morest, 1982). Each cell receives many of these inputs, probably from many auditory nerve fibers. Stellate cells give off small-diameter axons (not shown), which project to the inferior colliculus, periolivary nuclei, and lateral lemniscus (Adams, 1979; Warr, 1982).

Bushy cells have many unusual properties and appear to be specialized for the auditory system. They have small dendritic trees upon which there are few, if any, terminals (Cant and Morest, 1979b). Auditory nerve fibers form large synaptic terminals, called endbulbs of Held, on the somata of bushy cells (Brawer and Morest, 1975; Ryugo and Fekete, 1982). Each endbulb contains a large number of synaptic contacts (Lenn and Reese, 1966) and provides a powerful synaptic

influence on the bushy cell. A bushy cell receives relatively few endbulbs, from one to four (Cajal, 1909; Lorente de Nò, 1981). Thus, bushy cells receive powerful inputs from a very small number of auditory nerve fibers, whereas stellate cells receive many small inputs from a number of auditory nerve fibers. Bushy cells project medium and large diameter axons (not shown) which are the main input to the principal nuclei of the superior olivary complex (van Noort, 1969; Warr, 1982; Tolbert et al., 1982).

EFFECTS OF THE DENTRITIC TREE

The difference in distribution of inputs on bushy and stellate cells has important implications for the abilities of these cells to transmit information about auditory stimuli using a temporal code. The difference has to do with the process by which excitatory currents spread from their site of injection at a synapse to the site of action potential initiation in the soma. Figure 3A shows the situation schematically. A current $I_o(t)$ is injected at a synaptic site on a dendrite and produces a postsynaptic potential which spreads along the dendritic tree to the soma, resulting in a depolarization $V_s(t)$. It is assumed that $V_s(t)$ is the signal that drives the neuron's axon, in that an action potential is fired when $V_s(t)$ exceeds threshold. The ability of such a neuron to transmit temporally-encoded information will be sensitive to any filtering which occurs between the synaptic site and the soma.

FIGURE 2, Schematic summary of the morphology and innervation patterns of bushy and stellate cells from the cat anteroventral cochlear nucleus (Redrawn from Cant and Morest, 1984). Auditory nerve synapses are stippled. Note large endbulbs of Held on bushy cells.

An idea of the nature of current spread along dendritic membranes can be gained using models of dendritic electrotonus (Rall, 1977; Jack et al., 1975). A dendritic tree like the one sketched in Figure 3A is considered to be a series of electrically coupled cylinders of neural membrane. Under certain conditions (Rall, 1962; Jack et al., 1975) this complex of cylinders can be represented as a single equivalent cylinder, as shown in Figure 3B. The conditions under which this simplification is strictly valid do not always hold, but the equivalent cylinder nevertheless serves as a convenient idealization which allows the effects of dendritic current spread to be estimated. When the conditions for reduction to the equivalent cylinder hold, the V_s produced by current I_o is the same in the equivalent cylinder and the original dendritic tree (Rall, 1962).

The dendritic cylinder model is reduced to an electrical transmission line as shown in Figure 3C. The RC circuit consisting of conductance $g_m \Delta x$ in parallel with capacitance $c_m \Delta x$ represents the impedance to flow of current through the membrane of a piece of cylinder of length Δx . The resistors $r_i \Delta x$ represent the resistance to flow of current axially down the inside of the membrane cylinder between two points separated by Δx . The resistance to extracellular current flow is ignored and all points outside the cylinder are assumed to be at ground potential. The current source $I_0(t)$ represents the current injected by a synapse or population of synchronously-firing synapses at some distance L_1 down the cylinder. This RC transmission line represents the dendritic tree of the cell. It is terminated at the soma end by lumped conductance G_s in parallel with capacitance C_s , which together represent the impedance of the whole somatic membrane. The terminal end of the dendrites is terminated by an open circuit. There are a number of assumptions in this model. Some of these are approximations at best, but the model is likely to give generally correct results for the questions addressed here.

The RC transmission line of Figure 3C is a cascade of lowpass filters, because of the membrane capacitances. Thus the voltage signal at the soma V_s will be a lowpass version of the input current I_0 . For a cell whose primary inputs are on its dendrites, this filtering will intercede between synaptic inputs and the cell's output and should eliminate high frequency fluctuations in the cell's inputs. The results of calculations to support this point are shown in

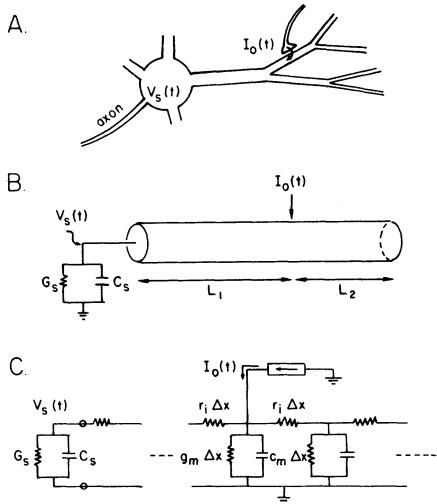


FIGURE 3. A. Sketch of a neuron showing its axon and a dendrite leaving its spherical soma. $V_s(t)$ is the membrane potential in the soma which is assumed isopotential. $I_0(t)$ is current injected into the dendrite by a synapse.

B. Reduction of the dendritic tree to a single equivalent cylinder of membrane terminated by a lumped-parameter model for the soma (G_s and C_s). The current is injected at distance L_1 from the soma and the dendritic cylinder has total length $L_1 + L_2$.

C. Electrical transmission line model for the equivalent cylinder.

Figure 4. This shows the magnitude and phase of the frequency-dependent transfer function from synaptic current I_o to somatic potential V_s . The calculations underlying this figure are described elsewhere (Young et al., 1986). It is assumed that the dendritic tree has a total length (L_1+L_2 in Figure 3B) of 1.5λ where λ is the cylinder's length constant, defined as $\lambda = 1/(g_{mri})^{1/2}$. λ is the distance along the cylinder at which the voltage produced by a steady (D.C.) injected current is reduced to $1/e$ of its value at the site of injection. Two cases are shown: in one, L_1 is 0, so the synapse is at the soma; in the other, $L_1=1.5\lambda$, so the synapse is at the end of the dendritic tree. The parameters of the model have been chosen to be typical of those measured for cat spinal motoneurons (Rall, 1977), because measurements are not available for cells in the CN.

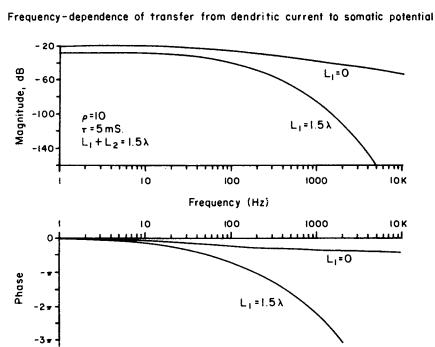


FIGURE 4, Magnitude and phase of the transfer function from I_o to V_s . Magnitude is transfer function multiplied by G_s , in dB re 1. Three free parameters of the model were chosen from data on cat spinal motoneurons (Rall, 1977). p is the dendritic dominance, τ is the membrane time constant, and $(L_1+L_2)/\lambda$ is the length of the dendritic tree relative to the D.C. space constant λ . The most important parameter for these calculations is τ ; plots will scale in frequency in inverse proportion to changes in τ .

Figure 4 shows that the effect of dendritic lowpass filtering is significant when the input is applied at the end of the tree ($L_1=1.5\lambda$). For a 1 kHz input, the signal is attenuated by 60 dB, relative to D.C. components, and undergoes substantial phase lag ($\approx 2\pi$). This is a strong lowpass filter whose cutoff frequency is around 100 Hz. The effects are much smaller when the current is injected near the soma ($L_1=0$), where the predominant effect is due to the input impedance of the soma and dendritic membrane.

It is clear from the calculation illustrated in Figure 4 that dendritic lowpass filtering must be bypassed if precise temporal information is to be transferred through the CN. Bushy cells seem to be specialized for this function. Inputs are applied directly to the somata of bushy cells, so that dendritic filtering is avoided. Moreover, the size of individual inputs is very large because each endbulb contains many synaptic contacts (Lenn and Reese, 1966). As a result, the somatic capacitance is rapidly charged and lowpass filtering due to somatic and dendritic input impedance is also avoided. Thus even the relatively mild lowpass filter of the $L_1=0$ case in Figure 4 is

bypassed. Bushy cells have additional membrane specializations to allow precise input/output timing. These are described by Oertel (1985) and will not be discussed here. As a result of their synaptic and membrane specializations, bushy cells are expected to transfer faithfully the temporal patterns of auditory nerve inputs. This expectation is consistent with physiological evidence, which is discussed below.

Stellate cells seem to be specialized in a different way. These cells integrate the activity of many auditory nerve fibers, each of which contributes a small input. Using the dendritic tree model discussed above, it can be shown that inputs at a distance from the soma along a dendritic tree produce small, slowly rising potentials in the soma, relative to inputs applied near the soma (Jack et al., 1975). Thus in stellate cells, it is likely that action potentials are produced only in response to the summated discharge of several auditory nerve inputs. Because of dendritic lowpass filtering, the temporal patterns of these inputs are likely to be transferred through the cell only at low frequencies. Physiological evidence supports this conclusion as well.

RESPONSE CHARACTERISTICS OF VCN NEURONS

The response properties of neurons in the CN have been studied in some detail. A variety of response patterns is seen and these have been classified in various ways (Pfeiffer, 1966a; Evans and Nelson 1973; Godfrey et al., 1975; van Gisbergen et al., 1975; Bourk, 1976; Young and Voigt, 1982; Shofner and Young, 1985; for a review, see Young, 1984). For the part of the CN under consideration in this paper, a useful classification is based on PST histograms of responses to short tone bursts at best frequency. Examples of common response types are shown in the bottom row of Figure 6. Most units in the anterior VCN fall into the primarylike or chopper categories (left and right columns), or into subclasses of these (Bourk, 1976). The PST histogram in the middle of the bottom row of Figure 6 is representative of a group of responses that share many properties with primarylike units (see below), but have unusual PST histograms that are neither primarylike nor chopper (Shofner and Young, 1985). These units are intermediate between primarylike units and a third class, which gives mainly onset responses to stimuli (Bourk, 1976). There seems to be a continuum of response types between primarylike and onset responses, differentiated in the relative size of the onset and steady state response and in the average discharge rate of the steady state response (Palmer et al., 1986).

An important goal of studies of response properties has been to characterize the response patterns of cells according to their morphology, i.e. to identify the response patterns of bushy and stellate cells. The general pattern of correspondence was originally inferred from indirect evidence; the most direct evidence on this point comes from studies in which cell marking techniques are used to stain and thereby identify the cells from which recordings are made. Using this technique, Rhode and collaborators (1983) and Rouiller and Ryugo (1984) have shown that chopper responses (right column in Figure 6) are recorded from stellate cells and primarylyke responses (left column in Figure 6) are recorded from bushy cells. Rouiller and Ryugo (1984) also showed three examples of onset responses recorded from bushy cells.

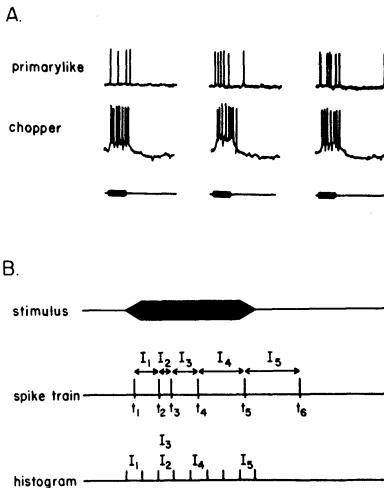


FIGURE 5, A. Spike trains recorded intracellularly from a primarylike unit and a chopper. Bottom row shows stimulus bursts, which were 25 ms in duration (redrawn from Romand, 1978).

B. Schematic explanation of the construction of the time-varying mean and standard deviation plots in Figure 6. Spikes occur at times t_i ; intervals between spikes are I_i . The I_i are placed into 0.1 ms bins according to the bin in which the first spike of each interval lies, as shown in the bottom row. The mean and standard deviation of intervals in each bin are computed after 500 repetitions of the stimulus.

Further evidence for correspondence of response properties and cell types comes from the shape of action potentials. Some cells in the VCN have action potentials with a **prepotential**, a small, usually positive, potential preceding the action potential. An example is shown in the inset in the bottom center PST histogram in Figure 6. The prepotential is indicated by the arrow. Prepotentials are thought to be the discharge of presynaptic endbulb terminals (Pfeiffer, 1966b; Bourk, 1976), and are therefore evidence that a recording is made from a bushy cell. Many units with prepotentials give primarylike responses (Bourk, 1976), but others give unusual responses like that shown in the middle column of Figure 6 (Shofner and Young, 1985). Chopper responses are not recorded from prepotential units. These results are consistent with the conclusion of the previous paragraph that chopper responses are recorded from stellate cells and primarylike, onset and unusual responses are recorded from bushy cells.

It is evident that bushy cells are capable of producing a variety of PST response patterns ranging from primarylike to onset. In many cases, such as the one at bottom center in Figure 6, the PST histograms of prepotential units appear to contain elements of chopping, and this has led to confusion about the differences between stellate cell and bushy cell responses. Another measure of response which clearly differentiates recordings presumed to come from stellate cells (choppers) from those presumed to come from bushy cells (primarylike, onset, and unusual intermediate types) is regularity of discharge. This is illustrated in Figure 5A which shows spike trains from a primarylike unit (top row) and a chopper unit (bottom row). These are typical of the spike discharge patterns displayed by units of the two types in intracellular recordings (Romand, 1978; Rhode et al., 1983). The primarylike unit's discharge is irregular, which means that

the number of spikes and the pattern of discharge vary greatly from burst to burst. The time intervals between successive spikes are also variable, and appear to be quite random. In contrast, the response pattern of the chopper is regular, in that the number and pattern of spikes is repeatable from burst to burst and the intervals between successive action potentials are much less variable than for the primarylike unit.

Figure 5B shows a method of evaluating the regularity of a unit's discharge from the same short tone burst data used to compute PST histograms (Bourk, 1976). The mean and standard deviation of time intervals between successive spikes (the I_i in Figure 5B) are estimated as functions of time through the stimulus burst. This is done by gathering interspike intervals (ISI) into bins according to the bin into which the first spike of an interval falls, as shown on the bottom line of Figure 5B. This process is repeated through 500 repetitions of the stimulus and the mean and standard deviation of ISI length are computed in each bin. This results in time-plots of mean and standard deviation of ISI like those shown in the top row of Figure 6.

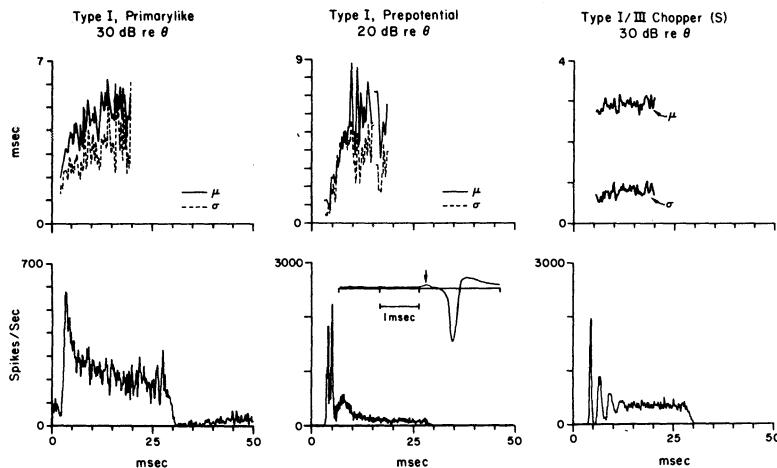


FIGURE 6, Properties of three major groups of units in anteroventral cochlear nucleus. Left column shows primarylike unit; middle column shows prepotential unit with unusual PST histogram; right column shows chopper. Top row shows regularity analysis; bottom row shows PST histograms. Both computed from responses to 25 ms best frequency tone bursts 20-30 dB above threshold. Inset in middle bottom plot shows action potential waveform of unit; arrow points to prepotential.

The data in Figure 6 are typical of primarylike, prepotential and chopper units in that primarylike and prepotential units are irregular and choppers are regular. The standard deviations of ISI (dashed lines) are only slightly below the means (solid lines) for the primarylike and prepotential units, whereas the standard deviations are much smaller for the chopper unit. In a population of cells from the VCN, choppers are separated from primarylike and prepotential units (and auditory nerve fibers) in that choppers have standard deviations less than 50%

of means, whereas primarylike and prepotential units (and auditory nerve fibers) have standard deviations greater than 50% of means. This is illustrated in the population plot in Figure 7.

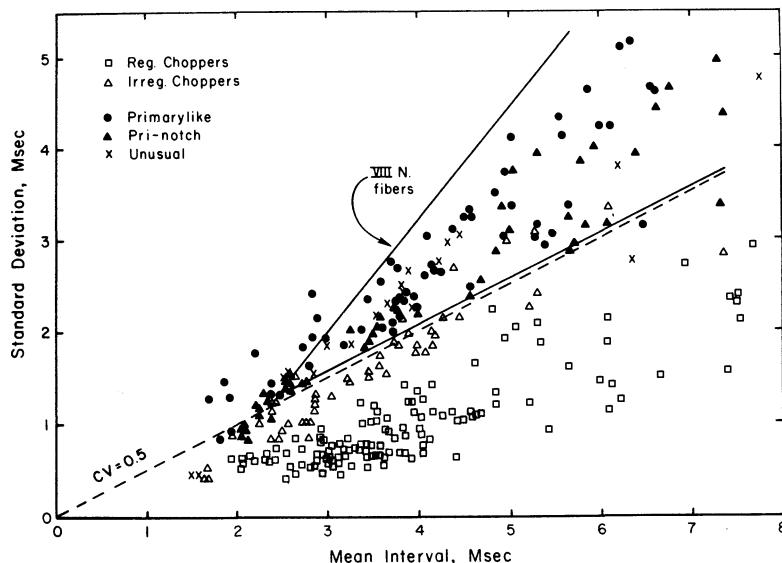


FIGURE 7. Mean interspike interval versus standard deviation for units from the cochlear nucleus. Mean and standard deviation averaged over all bins with latencies between 12 and 20 ms in regularity plots like those in Figure 6. Open symbols are choppers; filled symbols are primarylike units; Xs are units with unusual PSTs. Diagonal line shows where standard deviation equals 0.5 of mean. Region enclosed in solid lines is occupied by auditory nerve fibers (data points not shown). The differences between regular and irregular choppers (square and triangular symbols) will be discussed in a manuscript that is in preparation. Pri-notch units are a subset of primarylike units (Bourk, 1976).

The data on regularity summarized above show that bushy cells as a group are irregular like auditory nerve fibers, regardless of their PST histogram patterns. The irregular discharge of bushy cells can be explained as resulting from the tight synaptic coupling of auditory nerve fibers to bushy cells. A bushy cell fires an action potential in response to each incoming auditory nerve action potential and therefore the bushy cell's spike train is a superposition of the spike trains of the group of auditory nerve fibers forming endbulbs on the bushy cell, perhaps modified by refractoriness. Such a model has been shown to give irregular discharge (Molnar and Pfeiffer, 1968).

The discussion of dendritic filtering above suggests that bushy cells should preserve temporal information in the auditory nerve to higher frequencies than stellate cells. This expectation has been found to be true, in that prepotential units, including primarylike units,

phase-lock to tonal stimuli about as strongly as do auditory nerve fibers. Chopper units, on the other hand, phase-lock well only at lower frequencies (Bourk, 1976). The decrease in phase-locking in chopper units begins around 200 Hz, which is roughly consistent with the model used for Figure 4 if the membrane time constant $\tau = c_m/g_m$ is about 5 ms, which is typical of nerve membrane (Rall, 1977) and seems to be about right for stellate cells, from the data of Oertel (1985).

CN SUBSYSTEMS AND REPRESENTATION OF COMPLEX STIMULI

The morphological and physiological specializations of bushy cells suggest that these cells are designed to transmit precise temporal information from the auditory nerve to the principal cell groups of the superior olive where comparisons of interaural arrival time are done (Goldberg and Brown, 1969; Guinan et al., 1972; Caird and Klinke, 1983). Because bushy cells are specialized to transmit precise temporal information, the temporal representation of complex stimuli in bushy cells should be about the same as it is in the auditory nerve. By the same argument, it is unlikely that any extraction of temporal information or conversion of temporal to rate information occurs in bushy cells.* This expectation has been verified by studies in which the temporal representation of vowel second and third formant frequencies is as good in primarylike and prepotential units as it is in the auditory nerve (Blackburn et al., 1986; Palmer et al., 1986, personal communication).

Stellate cells, on the other hand, integrate the activity of many auditory nerve fibers and produce a very regular discharge. The regularity of discharge of stellate cells (choppers) makes them a precise indicator of stimulus level, in that the variance of their signal is low relative to its mean. Thus stellate cells may serve as a channel conveying accurate information about stimulus intensity, as opposed to information about temporal details of the stimulus.

There is evidence for a dichotomous representation of intensity in stellate cells and timing in bushy cells in the barn owl (Takahashi et al., 1984; Sullivan and Konishi 1984). In this animal, a bushy cell system conveys temporal information for time of arrival decisions to localize sound in azimuth, and a stellate cell system conveys intensity information for interaural intensity decisions to localize sound in the median plane (Knudsen and Konishi, 1979). The situation is more complex in mammals, however, because interaural intensity differences as well as interaural time differences are analyzed in the superior olive (Guinan et al., 1972; Tsuchitani, 1977; Caird and Klinke, 1983), but stellate cells do not seem to project to the principal cells of the olive (van Noort, 1969; Warr, 1982). That is, bushy cells seem to

*This discussion ignores the fact that there is a gradient of complexity in the bushy cell population between spherical and globular bushy cells. The summary given above is most accurate for spherical bushy cells. Globular bushy cells share the property of irregularity with spherical bushy cells, but have somewhat more complicated response properties (Bourk, 1976; Brownell, 1975), including inhibitory sidebands. Studies of complex signal representation in the CN have not yet attempted to analyze spherical and globular cell contributions separately.

support both interaural time and interaural intensity comparisons in mammals. Moreover, the statement that stellate cells are a rate channel coding intensity information ignores their possible contribution to stimulus spectrum coding.

Stellate cells are not expected to faithfully transmit temporal information because they do not phase-lock well at higher frequencies. This expectation has also been verified in that the representation of vowel spectra in chopper temporal response profiles is poor (Blackburn et al., 1986). At the same time, preliminary data suggest that the rate representation of vowel spectra in choppers is quite good and is stable over a wide range of sound levels (Blackburn and Sachs, personal communication). Stellate cells, because of their dendritic trees, offer the possibility of a sophisticated analysis of incoming auditory nerve discharge patterns. This might result in an improved rate-place representation of the spectra of complex stimuli through various means. For example, the membrane potential produced in the soma of a cell can be shown to depend on the spatio-temporal pattern of inputs applied to the dendritic tree of the cell (Rall, 1977). Thus, it is conceivable that dendritic processing could result in an extraction of temporal information from auditory nerve discharge patterns, without resulting in phase-locking. Alternatively, inhibitory interactions could result in a sharpening of rate profiles by a combination of lateral inhibition and the sort of temporal-to-place conversion studied by Shamma (1985). Finally, stellate cells could produce a sharpened and stable rate representation using only rate information from auditory nerve fibers by switching from high spontaneous rate fibers to low spontaneous rate fibers as overall sound level increases (Delgutte, 1982; Winslow, 1985). This takes advantage of the fact that a good rate profile exists in high spontaneous rate fibers at low sound levels and in low spontaneous rate fibers at high sound levels (Figure 1; Sachs and Young, 1979). Studies of stellate cells that are currently underway will help to answer these questions.

REFERENCES

1. Adams, J.C. (1979). Ascending projections to the inferior colliculus. *J. Comp. Neurol.*, 183, 519-538.
2. Blackburn, C.C., Shofner, W.P. and Sachs M.B. (1986). The representation of the steady-state vowel sound /e/ in the temporal discharge patterns of cat anteroventral cochlear nucleus neurons. *Abst. 9th Midwinter Res.Mtg., Assoc. Res. in Otol.*, 9, 130.
3. Bourk, T.R. (1976). Electrical Responses of Neural Units in the Anteroventral Cochlear Nuclues of the Cat. Doctoral dissertation, Massachusetts Institute of Technology, Cambridge.
4. Brawer, J.R. and Morest D.K. (1975). Relations between auditory nerve endings and cell types in the cat's anteroventral cochlear nucleus seen with the Golgi method and Nomarski optics. *J. Comp. Neurol.*, 160, 491-506
5. Brownell, W.E. (1975). Organization of the cat trapezoid body and the discharge characteristics of its fibers. *Brain Research*, 94, 413-433.

6. Caird, D. and Klinke, R. (1983). Cat superior olivary complex (SOC): The basis of binaural information processing. In: R. Klinke and R. Hartmann (Eds.) Hearing--Physiological Bases and Psychophysics, 216- 223. Springer Verlag, Berlin.
7. Cajal, S.R., (1909-1911). Histologie du Systeme Nerveux de l'Homme et des Vertebres. Maloine, Paris.
8. Cant, N.B. (1981). The fine structure of two types of stellate cells in the anterior division of the anteroventral cochlear nucleus of the cat. Neuroscience, 6, 2643-2655.
9. Cant, N.B. and Morest D.K. (1979a). Organization of the neurons in the anterior division of the anteroventral cochlear nucleus of the cat: Light microscopic observations. Neuroscience, 4, 1909-1923.
10. Cant, N.B. and Morest, D.K. (1979b). The bushy cells in the anteroventral cochlear nucleus of the cat. A study with the electron microscope. Neuroscience, 4, 1925-1945.
11. Cant, N.B. and Morest, D.K. (1984). The structural basis for stimulus coding in the cochlear nucleus of the cat. In: C.I. Berlin (Ed.) Hearing Science, Recent Advances, 371-421. College Hill Press, San Diego.
12. Delgutte, B. (1982). Some correlates of phonetic distinctions at the level of the auditory nerve. In: R. Carlson and B. Granstrom (Eds.) The Representation of Speech in the Peripheral Auditory System, 131- 150. Elsevier, Amsterdam.
13. Delgutte, B. and Kiang, N.Y.-S. (1984). Speech coding in the auditory nerve: V. Vowels in background noise. Jour. Acoust. Soc. Am., 75, 908-918.
14. Evans, E.F. and Nelson P.G. (1973). The responses of single neurones in the cochlear nucleus of the cat as a function of their location and the anaesthetic state. Exp. Brain Res., 17, 402-427.
15. Frisina, R.D., Smith, R.L., and Chamberlain, S.C. (1985). Differential encoding of rapid changes in sound amplitude by second- order auditory neurons. Exp. Brain. Res., 60, 417-422.
16. Gibson, D.J., Young, E.D., and Costalupes, J.A. (1985). Similarity of dynamic range adjustment in auditory nerve and cochlear nuclei. J. Neurophysiol., 53, 940-958.
17. Gisbergen, J.A.M. van, Grashuis, J.L., Johannesma, P.I.M., and Vendrik, A.J.H. (1975). Spectral and temporal characteristics of activation and suppression of units in the cochlear nuclei of the anaesthetized cat. Exp. Brain Res., 23, 367-386.
18. Godfrey, D.A., Kiang N.Y.S., and Norris, B.E. (1975). Single unit activity in the posteroventral cochlear nucleus of the cat. J. Comp. Neurol., 162, 247-268.
19. Goldberg, J.M. and Brown, P.B. (1969). Response of binaural neurons of dog superior olivary complex to dichotic tonal stimuli: Some physiological mechanisms of sound localization. J. Neurophysiol., 32, 613-636.
20. Guinan, J.J., Guinan, S.S., and Norris, B.E. (1972). Single auditory units in the superior olivary complex II: Locations of unit categories and tonotopic organization. Intern. Neurosci., 4, 147-166.
21. Imig, T.J., Reale, R.A. and Brugge, J.F. (1982). The auditory cortex: Patterns of corticocortical projections related to physiological maps in the cat. In: C.N. Woolsey (Ed.) Cortical Sensory Organization, Vol. 3, 1-42. Humana Press, Clifton, N.J.

22. Jack, J.J.B., Noble, D., and Tsien, R.W. (1975). Electric Current Flow in Excitable Cells. Clarendon Press, Oxford.
23. Knudsen, E.I. and Konishi, M. (1979). Mechanisms of sound localization in the barn owl. J. Comp. Physiol., 133, 13-21.
24. Lenn, N.J. and Reese, T.S. (1966). Fine structure of nerve endings in the nucleus of the trapezoid body and the ventral cochlear nucleus. Amer. J. Anat., 118, 375-390.
25. Liberman, M.C. (1978). Auditory-nerve response from cats raised in a low-noise chamber. J. Acoust. Soc. Am., 63, 422-455.
26. Lorente de Nò, R. (1981). The primary Acoustic Nuclei. Raven, New York.
27. Merzenich, M.M., Loeb, G.E., and White, M.W. (1980). Extraction of spectral information in auditory brainstem nuclei; hypothesis and experimental observations. J. Acoust. Soc. Am., 68, S19 (Abst.).
28. Merzenich, M.M., Colwell, S.A., and Andersen, R.A. (1982). Auditory forebrain organization: Thalamocortical and corticothalamic connections in the cat. In: C.N. Woolsey (Ed.) Cortical Sensory Organization, Vol. 3, 43-58. Humana Press, Clifton, N.J.
29. Mills, A.W. (1972). Auditory localization. In: J.V. Tobias (Ed.) Foundations of Modern Auditory Theory, Vol. II, 301-348. Academic Press, New York.
30. Molnar, C.E. and Pfeiffer, R.R. (1968). Interpretation of spontaneous spike discharge patterns of cochlear nucleus neurons. Proc. IEEE, 56, 993-1002.
31. Noort, van J. (1969). The Structure and Connections of the Inferior Colliculus, van Gorcum and Co., Assen, The Netherlands.
32. Oertel, D. (1985). Use of brain slices in the study of the auditory system: Spatial and temporal summation of synaptic inputs in cells in the anteroventral cochlear nucleus of the mouse. J. Acoust. Soc. Am., 78, 328-333.
33. Palmer, A.R., Winter, I.M. and, Darwin, C.J. (1986). The representation of steady-state vowel sounds in the temporal discharge patterns of the guinea pig cochlear nerve and primarylike cochlear nucleus neurons. J. Acoust. Soc. Am., 79, 100-113.
34. Pfeiffer, R.R. (1966a). Classification of response patterns of spike discharges for units in the cochlear nucleus: Tone burst stimulation. Exp. Brain Res., 1, 220-235.
35. Pfeiffer, R.R. (1966b). Anteroventral cochlear nucleus: Wave forms of extracellularly recorded spike potentials. Science, 134, 667-668.
36. Rall, W. (1962). Theory of physiological properties of dendrites. Ann. N.Y. Acad. Sci., 96, 1071-1092.
37. Rall, W. (1977). Core conductor theory and cable properties of neurons. In: E.R. Kandel (Ed.) Handbook of Physiology, Section I: The Nervous System. Volume I: Cellular Biology of Neurons, 39-97. American Physiological Society, Bethesda.
38. Rhode, W.S., Oertel, D., and Smith, P.H. (1983). Physiological response properties of cells labeled intracellularly with horseradish peroxidase in the cat ventral cochlear nucleus. J. Comp. Neurol., 213, 448-463.
39. Romand, R. (1978). Survey of intracellular recording in the cochlear nucleus of the cat. Brain Res., 148, 43-65.
40. Rouiller, E.M. and Ryugo, D.K. (1984). Intracellular marking of physiologically characterized cells in the ventral cochlear nucleus of the cat. J. Comp. Neurol., 225, 167-186.

41. Ryugo, D.K. and Fekete, D.M. (1982). Morphology of primary axosomatic endings in the anteroventral cochlear nucleus of the cat: A study of the endbulbs of Held. *J. Comp. Neurol.*, **210**, 239-257.
42. Sachs, M.B. (1984). Speech encoding in the auditory nerve. In: C.I. Berlin, (Ed.) *Hearing Science, Recent Advances*, 263-307. College Hill Press, San Diego.
43. Sachs, M.B. and Young, E.D. (1979). Encoding of steady-state vowels in the auditory nerve: Representation in terms of discharge rate. *J. Acoust. Soc. Am.*, **66**, 470-479.
44. Sachs, M.B. and Young, E.D. (1980). Effects of nonlinearities on speech encoding in the auditory nerve. *J. Acoust. Soc. Am.*, **68**, 858- 875.
45. Sachs, M.B., Voigt, H.F.,and Young, E.D. (1983). Auditory nerve representation of vowels in background noise. *J. Neurophysiol.*, **50**, 27-45.
46. Schalk, T.B. and Sachs, M.B. (1980). Nonlinearities in auditory-nerve fiber responses to bandlimited noise. *J. Acoust. Soc. Am.*, **67**, 903- 913.
47. Shamma, S.A. (1985). Speech processing in the auditory system II: Lateral inhibition and the central processing of speech-evoked activity in the auditory nerve. *J. Acoust. Soc. Am.*, **78**, 1622-1632.
48. Shofner, W.P. and Young, E.D. (1985). Excitatory/inhibitory response types in the cochlear nucleus: Relationships to discharge patterns and responses to electrical stimulation of the auditory nerve. *J. Neurophysiol.*, **54**, 917-939.
49. Sullivan, W.E. and Konishi, M. (1984). Segregation of stimulus phase and intensity coding in the cochlear nucleus of the barn owl. *J. Neurosci.*, **4**, 1787-1799.
50. Takahashi, T., Moiseff, A., and Konishi, M. (1984). Time and intensity cues are processed independently in the auditory system of the owl. *J. Neurosci.*, **4**, 1781-1786.
51. Tolbert, L.P. and Morest, D.K. (1982). The neuronal architecture of the anteroventral cochlear nucleus of the cat in the region of the cochlear nerve root: Electron microscopy. *Neuroscience*, **7**, 3053-3067.
52. Tolbert, L.P., Morest, D.K., and Yurgelun-Todd, D.A. (1982). The neuronal architecture of the anteroventral cochlear nucleus of the cat in the region of the cochlear nerve root: Horseradish peroxidase labelling of identified cell types. *Neuroscience*, **7**, 3031-3052.
53. Tsuchitani, C. (1977). Functional organization of lateral cell groups of cat superior olivary complex. *J. Neurophysiol.*, **40**, 296-318.
54. Warr, W.B. (1982). Parallel ascending pathways from the cochlear nucleus: Neuroanatomical evidence of functional specialization. *Contributions to Sensory Physiology*, **7**, 1-38.
55. Winslow, R.L. (1985). *A Quantitative Analysis of Rate-Coding in the Auditory-Nerve*. Doctoral dissertation, The Johns Hopkins Univ., Baltimore.
56. Yin, T.C.T. and Kuwada, S. (1983). Binaural interaction in low-frequency neurons in inferior colliculus of the cat. II. Effects of changing rate and direction of interaural phase. *J. Neurophysiol.*, **50**, 1000-1019.
57. Yin, T.C.T., Kuwada S., and Sujaku, Y. (1984). Interaural time sensitivity of high frequency neurons in the inferior colliculus. *J. Acoust. Soc. Am.*, **76**, 1401-1410.

58. Yost, W.A. (1974). Discriminations of interaural phase differences. J. Acoust. Soc. Am., 55, 1299-1303.
59. Young, E.D. (1984). Response characteristics of neurons of the cochlear nuclei. In: C.I. Berlin (Ed.) Hearing Science, Recent Advances, 423-460. College Hill Press, San Diego.
60. Young, E.D. and Sachs, M.B. (1979). Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers. J. Acoust. Soc. Am., 66, 1381- 1403.
61. Young, E.D. and Voigt H.F. (1982). Response properties of type II and type III units in the dorsal cochlear nucleus. Hearing Res. 6, 153- 169.
62. Young, E.D., Shofner, W.P., White, J.A., Robert, J.M., and Voigt, H.F. (1986). Response properties of cochlear nucleus neurons in relationship to physiological mechanisms. In: S. Hassler (Ed.) Functions of the Auditory System. John Wiley & Sons, New York.

CHANGES IN THE PHONEMIC QUALITY AND NEURAL REPRESENTATION OF A VOWEL BY ALTERATION OF THE RELATIVE PHASE OF HARMONICS NEAR F1*

A.R. Palmer*, I.M. Winter, R.B. Gardner, and C.J. Darwin,

Laboratory of Experimental Psychology, University of Sussex,
Brighton BN1 9QG, England. *Current address: MRC Institute of
Hearing Research, University of Nottingham, University Park,
Nottingham NG 2RD, England

INTRODUCTION

Darwin and Gardner (1986), in a study of the contribution of mistuned harmonics to the percept of a vowel, discovered that mistunings of only 1-2Hz in harmonics close to the first formant (F1) caused a lowering of the phoneme boundary in a categorization task, indicating that the F1 was perceived as higher in frequency. This occurred whether the mistuned harmonic was just above or just below F1. At larger mistunings (40Hz) the mistuned harmonic is partially excluded from judgement of the vowel quality and the perceived F1 approaches that heard when the mistuned harmonic is physically absent. The changed percept due to small mistuning of the harmonic was explained by Darwin and Gardner as an effect of the altered relative phase of the harmonic.

The present study extends these results in two ways. First we have made more observations of the effect of harmonic phase on vowel quality and show that, compared to other phase shifts, a 90° phase shift of a harmonic above F1 produces the maximum increase in perceived F1 frequency; this is perceptually equivalent to an increase in the level of the normal harmonic by 4dB. Second, we have used similar vowel continua to investigate the way in which the first formant of a vowel is represented in the discharge of auditory neurones. Phase shifting a single component of a vowel produced reliable changes in the neural representation and suggested that neither phase-locking alone nor mean discharge rate alone, but a combined rate/time/place code (see Young and Sachs, 1979) may be important for transmission of information relating to the perceived value of F1.

PSYCHOPHYSICAL METHODS

The paradigm developed by Darwin, (see Darwin 1984, and Darwin & Gardner, 1986, for details) was used to assess the extent to which a particular harmonic is perceptually integrated into the vowel. Briefly, vowels were generated using additive sine-wave synthesis based on Klatt's (1980) cascade synthesizer. Evaluation of the transfer

*ARP was supported by the Royal Society and the MRC, IMW by an SERC studentship and CJD and RBG by the SERC. Thanks to Drs. Carlyon, Rees and Summerfield for helpful criticism of the manuscript.

function at 125 Hz intervals enabled addition of sine-waves of appropriate amplitude and phase to produce the complete vowel. Steady-state continua were produced consisting of 9 sounds of 200 ms duration and 16 ms rise-fall times, with F1 varying in 21 Hz steps from 375-543 Hz (bandwidth 90 Hz), giving /I/ like vowels at low F1 and /e/-like vowels at high F1. The second to fifth formants were 2300, 2900, 3800, and 4800 Hz with bandwidths of 110, 170, 1000, and 1000 Hz respectively. We have used a 'zero-phase' F1 continuum in which all components were at phases given by the Klatt transfer function, and also F1 continua in which the phase of the 500 Hz component was shifted in 45° increments. Additional experiments used continua in which the amplitude of the 500 Hz component was varied in 3dB steps. Subjects listened monotonically over Sennheiser headphones to the vowels presented on-line from the computer at a sampling frequency of 10 kHz, low-pass filtered at 4.5 kHz and 38 dB/octave. Subjects labelled the sounds as either /I/ or /e/. The phoneme boundary for each condition was taken as the 50% point on a probit function expressed in terms of the F1 value used to program the synthesizer.

PHYSIOLOGICAL METHODS

Details of the methods used in the physiological experiments may be found in Palmer et al. (1986). Briefly, a series of synthetic vowels was used with a fundamental of 100Hz and with F1s of 417, 438, 459, and 480 Hz, other details are as above. For each of these vowels three stimuli were presented: a zero phase condition in which all components were at phases indicated by the transfer function (see above), and two further conditions in which the phase of either the 500 Hz or the 400 Hz component (the harmonics immediately above and below the F1) were phase shifted by 90°. 400 ms bursts of the vowels at 1/s were presented at 75 dB SPL to anaesthetized guinea-pigs in a closed sound field. Responses to these vowels of fibres in the auditory nerve and of 'primarylike' cells in the cochlear nucleus were determined. Period histograms locked to the pitch period of the vowel were constructed and analysed by Fourier transformation. Such data from cochlear nucleus neurones tuned to frequencies below 1000 Hz have enabled us to construct functions showing the mean rate distribution and the rate of synchronized discharge to each of the vowel harmonics in neurones tuned to each harmonic (the Average Localized Synchronized Rate (ALSR) function, see Young & Sachs, 1979) and also to determine the component of the vowel which is most effective in driving each neurone (the dominant component, see Delgutte, 1984).

PSYCHOPHYSICAL RESULTS

Figure 1 shows the position of the phoneme boundary, averaged across subjects, between /I/ and /e/ (plotted as the value of the F1 used in synthesis) as a function of the phase of the 500 Hz component. These data confirm, with greater resolution, the effects of phase shifts reported by Darwin & Gardner, 1986. Phase-shifting the 500 Hz component by 90° produced the greatest change in the position of the phoneme boundary: a downward shift by just over 20 Hz. A downward shift in the boundary is equivalent to a increase in the perceived first formant frequency. Similar shifts in the position of the boundary could be achieved by varying the amplitude of the 500 Hz

component: a downward shift of 20 Hz is produced by increasing the 500 Hz component by 4 dB.

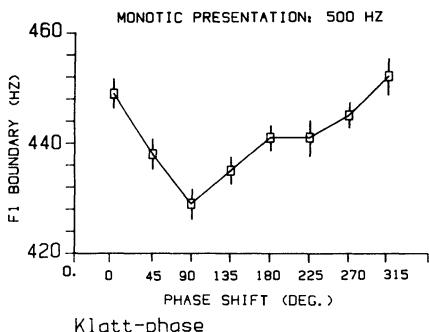


FIGURE 1, Phoneme boundaries of vowel continua as a function of the phase shift of the 500Hz fourth harmonic.

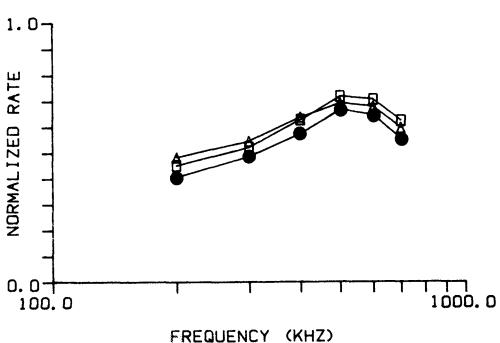


FIGURE 2, Distribution of the normalized mean discharge rate in a population of cochlear nucleus cells. Points represent means from cells whose CF fell within a ± 0.5 octave frequency window around each harmonic. Filled circles: zero phase condition, open squares: 500Hz component shifted by 90°, open triangles: 400Hz component shifted by 90°. F1=480Hz

PHYSIOLOGICAL RESULTS

We have recorded the responses of only a few cochlear nerve fibres to these stimuli; their responses were qualitatively similar to those described here for 'primarylike' cochlear nucleus cells. We could not find any effects of the phase shifts on the responses of neurones tuned to frequencies above about 700 Hz and thus have limited our recordings and analyses to neurones tuned at and below 700 Hz. The largest effects of the phase shifts were found for an F1 of 480 Hz.

(1) Mean Discharge Rate

Figure 2 shows the distribution of mean discharge rates (normalized as in Sachs & Young 1979) in a population of cochlear nucleus cells in response to a vowel with an F1 of 480 Hz. There were no statistically significant differences between the mean-rate functions for the various phase conditions for any of the F1 values. The peak estimated by a quadratic fit to the mean-rate functions varied from 498-537 Hz in the various conditions (table I), which in all cases was greater than the actual F1 and showed no consistent increase with increase in F1. In no case was there any significant change in the estimated peak before and after phase-shifting the 400 or 500 Hz component.

(2) ALSR function

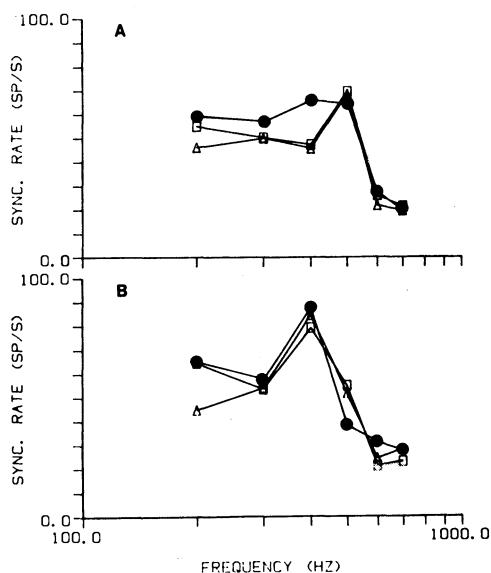


FIGURE 3, Distribution of the discharge rate synchronized to each harmonic. Frequency window and symbols as Figure 2. F1=480Hz in A and 417Hz in B.

ALSR shape whether it is at the 400 or 500 Hz component, but that the response is always greatest at the harmonic closest to the actual F1. Estimates of the position of the peak of the ALSR function by quadratic fits varied from 388 to 484 Hz, with higher F1 frequency resulting in higher frequency peaks. Introducing a phase-shift resulted in all cases in a significant upward shift in the estimated peak by up to 40 Hz (table I).

Table I

F1 Hz	Estimated Peak Frequency					
	0 DEGREES		500 Hz -90°		400 Hz -90°	
	ALSR	MR	ALSR	MR	ALSR	MR
417	388	503	401	498	398	518
438	399	531	423	531	436	518
459	426	533	465	531	466	515
480	460	534	483	537	484	529

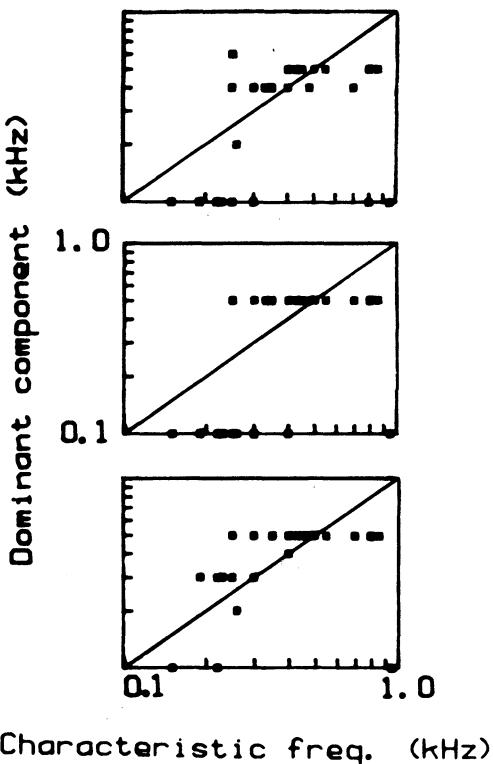
(3) Dominant Components

Figure 4 shows the frequency of the harmonic dominating discharge of each neurone as a function of the characteristic frequency (CF) of that neurone. In the zero phase condition the 400 and 500 Hz harmonics are about equally effective; the 400 Hz

Figure 3 shows the discharge rate synchronized to each of the harmonics of the vowel (the ALSR function). The functions for both phase-shifted conditions lie well below the control condition at the lower harmonics. This reduction in synchronized discharge to the lower harmonics is consistent and quite severe: the discharge synchronized to 400 Hz is reduced from 65/s to 47/s (28%). This same pattern, although somewhat less marked, was found for an F1 of 454 Hz. For both the 417 Hz and the 438 Hz data the change in shape of the ALSR function was somewhat different. In these cases the curve maximum was at 400 Hz, with small reductions in synchronized discharge to the low harmonics and a boosting of the synchronized rate to the 500 Hz component compared to the control by 26% from 34 to 46 per second as shown for an F1 of 417 Hz in Fig. 3b. Notice that in Fig. 3 the phase shift produces the same change in

component dominating neurones of CF 400 Hz and below, while the 500 Hz dominates CFs of 400 Hz and above. A 90° phase shift of either the 400 or 500 Hz components removes the domination by the 400 Hz component; for the 500 Hz phase-shift this leaves only the 500 Hz dominance. Phase shifting the 400 Hz component allows the 500 Hz domination to spread through the 400 Hz region, but CFs below this are now dominated by lower harmonics. It should be remembered that synchronization to 400 Hz is still present (as shown in Fig.3) and indeed quite strong, but is no longer the largest component in the neural discharge.

FIGURE 4, Frequency of the component of the vowel dominating each neurones discharge as a function of the neural CF. Top: zero phase condition, middle: 500Hz component shifted by 90°, bottom: 400Hz component shifted by 90°. Dashed line indicates domination by components at the CF.



DISCUSSION

The psychophysical results shown here confirm the earlier data in demonstrating that the perception of vowel quality depends upon the phase relations of its harmonics. A 90° phase shift of a single harmonic above F1 produces as much elevation of the perceived F1 (about 20 Hz) as increasing the amplitude of the same harmonic by 4 dB.

Merely using a simple quadratic fit to estimate the peak of the neural response functions, we found that vowels synthesised on higher F1s resulted in ALSR functions with higher peaks. We have also demonstrated consistent, statistically significant, changes in the ALSR functions as a result of phase shifting a single component near F1 by

90°. The changes related to the phase shift, which we observed in the ALSR functions, were of the same order of magnitude and in the same direction as the psychophysical observations. We did not find that the peak of the mean rate functions changed systematically with the synthesis F1, nor did we find any consistent phase-shift related increases in the peak frequency. Phase-locking information alone, as indicated in the dominant component analysis, certainly showed phase-shift related changes, but it is unclear how one should suitably weight the data points to obtain an estimate of the change in F1.

There are several pieces of evidence to suggest that a code combining rate, time, and place may be an important means of signalling complex sounds. Such a code is robust, has a wide dynamic range and is resistant to interfering noise (e.g. Young and Sachs, 1979; Sachs et al., 1983; Delgutte, 1984; Delgutte and Kiang, 1984). The present data indicate that this form of code, operating well below levels saturating single fibre discharge, is capable of signalling nuances in the frequency of F1 which are only revealed by sophisticated psychophysical tests.

REFERENCES

1. Darwin, C.J. and Gardner, R.B. (1986). Mistuning a harmonic of a vowel : Grouping and phase effects on vowel quality. *Journal of the Acoustical Society of America*, 79, 838-845.
2. Darwin, C.J. (1984). Perceiving vowels in the presence of another sound: constraints on formant perception. *Journal of the Acoustical Society of America*, 76, 1636-1647.
3. Delgutte, B. (1984). Speech encoding in the auditory nerve: II. Processing schemes for vowel-like sounds. *Journal of the Acoustical Society of America*, 75, 879-886.
4. Delgutte, B. and Kiang, N.Y.S. (1984). Speech coding in the auditory nerve: V. Vowels in background noise. *Journal of the Acoustical Society of America*, 75, 908-918.
5. Klatt, D.H. (1980). Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, 67, 971- 995.
6. Sachs, M.B. and Young, E.D. (1979). Encoding of steady-state vowels in the auditory nerve: representation in terms of discharge rate. *Journal of the Acoustical Society of America*, 66, 470-479.
7. Sachs, M.B., Voigt, H.F. and Young, E.D. (1983). Auditory nerve representation of vowels in background noise. *Journal of Neurophysiology*, 50, 27-45.
8. Palmer, A.R., Winter, I.M. and Darwin, C.J. (1986). The representation of steady-state vowels in the temporal discharge patterns of the guinea pig cochlear nerve and primarylike cochlear nucleus neurons. *Journal of the Acoustical Society of America*, 79, 100-113.
9. Young, E.D. and Sachs, M.B. (1979). Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory nerve fibers. *Journal of the Acoustical Society of America*, 66, 1381-1403

PHASE VOWELS*

Hartmut Traunmüller

Institutionen för Lingvistik, Stockholms Universitet,
S-106 91 Stockholm, Sweden.

INTRODUCTION

Physiological data appear to show that the time-averaged neural discharge rate over tonotopical place at the level of the auditory nerve represents the spectra of speech sounds in a very crude fashion at moderate to high levels of stimulus presentation (Sachs and Young, 1979). Although it has been shown that formant information can be recovered from the pattern of neural activity by computing an "average localized synchronized rate" (Young and Sachs, 1979; Delgutte and Kiang, 1984), this procedure appears unlikely to be neurophysiologically feasible. On the other hand, amplitude modulation of tones ($f_{mod} < 500$ Hz) has been shown to be preserved faithfully in the activity of cochlear nucleus units over the whole dynamic range of auditory perception (Möller, 1974). It should be noted that the accurate representation of amplitude modulation may be relevant to the perception of formants. The phases of the subsequent partials of a voice source signal are shifted by 180 degrees at each formant. As a consequence, the amplitude modulation sensed by a neuron with a center frequency at the formant frequency will be delayed, by a considerable fraction of the fundamental period, as compared with that sensed by neurons centered at frequencies above or below that formant. Wave form and depth of modulation will also be affected. Thus, even if all time-averaged amplitude information is neutralized in a speech sound such as a vowel - which appears to be close to what actually happens in the sense of hearing at high intensity levels - formant information is still preserved as a local perturbation in the time vs place pattern of neural activity.

Is it, then, possible to distinguish vowel qualities if all amplitude information is already neutralized in the acoustic signal?

METHOD

Synthetic "vowels" with a close to natural intonation contour and with phase spectra similar to those of natural vowels, but all with the same peakless envelope of their amplitude spectra, were generated by partial synthesis, performed digitally by means of a block diagram

* This work was supported, in part, by a grant from FRN, the Swedish Council for Planning and Coordination of Research. I am grateful to FM Diana Krull for her assistance in running the experiments.

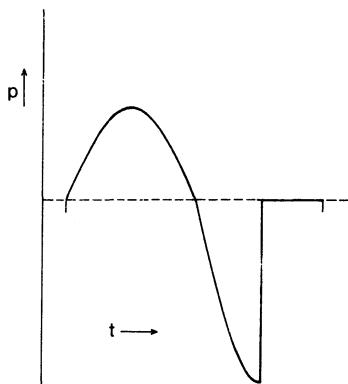


FIGURE 1, Shape of one period of the voice source signal, imitating a glottal pressure pulse, according to a model by G. Fant (1979).

sounding intonation was obtained. The "phase formants" were kept stationary in frequency.

Stimuli were generated with mean fundamental frequencies of 71, 100, 141, 200, and 283 Hz (1/2-octave intervals). The frequencies of the phase formants were those of the four lowest formants of the nine long Swedish vowels (= letter names), as measured by Fant (1959), with the following slight modifications:

- $f_0 = 71$ Hz: male values reduced by 3%;
- $f_0 = 100$ Hz: male values;
- $f_0 = 141$ Hz: mean of male values weighted 0.8 and female ones weighted 0.2;
- $f_0 = 200$ Hz: mean of female values weighted 0.9 and male ones weighted 0.1;
- $f_0 = 283$ Hz: female values increased by 3% plus 30 Hz.

Each of the nine "phase vowels" was generated at each of the five fundamental frequencies using the partials below approximately 6 kHz. Sampling frequency was 16 kHz and a low-pass filter cutting off at 5 kHz was used. Each vowel had a duration of 0.5 s and was repeated with an interval of 1.5 s. The 45 different pairs of stimuli were presented twice in quasi-random order, avoiding immediate repetition of the same f_0 or the same intended vowel. Each stimulus pair was separated from the following different one by a pause of 4.0 s.

25 listeners, mainly employees and students at the Institute of Linguistics, participated as subjects. 17 of them had Swedish as a first language, six had a different first language (English, Estonian, Finnish, German, Portuguese), mostly with good proficiency in Swedish, while

simulating algorithm. As a first step, a Fourier analysis was performed on a glottal pulse described by means of a three-parameter model voice source (Fant 1979). The glottal pulse is shown in Figure 1. Phases and amplitudes of the 80 partials lowest in frequency were calculated. In generating the synthetic stimuli, the calculated amplitude values of each partial were used directly, while the phases were modified as if there had been four vowel formants with an invariable bandwidth of 50 Hz each. Amplitude and period of the glottal pulse were controlled in accordance with the result of an inverse-filter analysis of the vowel [y:] spoken in citation form by a female speaker of Swedish, whereby a naturally

	u	o	ø	æ	e	ø	ø	y	i	*
u	11	33	2	-	4	-	2	-	-	-
o	1	45	6	-	-	-	-	-	-	-
ø	-	-	52	-	-	-	-	-	-	-
æ	-	-	2	40	4	4	-	2	-	-
e	-	-	-	-	33	15	2	2	-	-
ø	-	-	-	-	-	10	42	-	-	-
ø	-	-	-	-	-	-	5	30	17	-
y	-	-	-	-	-	6	4	9	30	3
i	-	-	-	-	-	10	-	4	14	24

$f_0 = 70 \text{ Hz}$

	u	o	ø	æ	e	ø	ø	y	i	*
u	12	11	-	2	8	2	3	5	4	5
o	-	31	1	-	16	-	-	-	-	4
ø	-	-	39	2	8	-	-	-	-	3
æ	-	-	1	35	2	6	2	2	4	-
e	1	5	-	-	25	14	2	3	2	-
ø	-	-	-	-	7	45	-	-	-	-
ø	-	-	-	-	-	2	26	18	4	2
y	1	-	-	-	13	11	6	17	4	-
i	-	-	-	-	14	2	2	6	23	-

$f_0 = 100 \text{ Hz}$

	u	o	ø	æ	e	ø	ø	y	i	*
u	-	8	3	3	16	6	-	6	2	8
o	-	5	4	3	21	3	6	4	4	2
ø	-	6	15	3	13	10	-	-	2	3
æ	-	3	2	12	4	28	3	-	-	-
e	1	3	1	-	12	23	5	1	-	6
ø	-	-	-	2	4	41	2	3	-	-
ø	-	-	-	4	5	2	29	8	2	-
y	-	-	-	4	13	10	10	9	-	6
i	2	4	4	6	18	1	1	2	8	6

$f_0 = 140 \text{ Hz}$

two were native bilinguals (with Danish or Estonian along with Swedish). They listened to the stimuli presented binaurally and at a comfortable intensity level through headphones (Sennheiser HD 414) directly from the computer. The subjects were asked to identify the stimuli and to encircle the corresponding orthographic symbol on an answer sheet containing one preprinted set of the nine vowels for each presented pair of stimuli. The subjects were allowed to encircle two symbols in cases of doubt. One-symbol responses were subsequently counted as two responses for the same symbol. The subjects were also allowed to identify stimuli as non-vowels (writing down an "x").

A set of nine selected different stimuli was presented initially for the purpose of accommodation to the experiment. Only the answers given to the 90 subsequent stimuli were evaluated. The experimental procedure lasted for approximately 15 minutes, including 5 minutes for instructions.

Table 1 (left column),
Confusion matrices of the nine
"phase vowels" presented at F_0
= 71, 100, and 141 Hz.
Responses from the 13 most
efficient listeners.
Vertically: Intended vowels;
Horizontally: Perceived vowel
phonemes. "*" Non-vowel
responses and response
omissions.

RESULTS

Since each stimulus was presented twice during the course of the experiment, it was possible to check how consistently the subjects identified the stimuli. This degree of consistency proved to vary substantially between subjects. Before further evaluation of the results, the subjects were rank-ordered according to the sum of the number of different labels attached consistently to the stimuli at each f_0 . We call this the "efficiency" of the subjects.

At low f_{0s} , most subjects identified the phase vowels quite well, as can be seen in the confusion matrices, Table 1, where the results of the 13 most efficient subjects are displayed for $f_0 = 71, 100,$ and 141 Hz . The intended vowels /u/ and /ə/ were, however, mostly identified as /o/ and /ø/, respectively, even at $f_0 = 71\text{ Hz}$. This kind of "confusion" would probably also be obtained with regular stationary formants with the frequency positions used here. While the Swedish vowels /u:/ and /ə:/ are characteristically labialized towards their end, the formant frequencies had been measured in sustained vowels. At the lowest f_0 , nearly all confusions occurred between vowels that are close to each other in the formant frequency space (e.g. F1 vs F2 vs F3). At $f_0 = 100\text{ Hz}$ the most frequent responses still coincided with those intended, but a bias towards /e/ and /ø/ responses became noticeable. At $f_0 = 141\text{ Hz}$, this bias dominated the picture. Among the other vowels only the intended /ø/ was still most frequently labeled "correctly" as an /ø/. This vowel is the only one that was given the intended label without exception at $f_0 = 71\text{ Hz}$. At the higher f_{0s} , the bias towards /e/ and /ø/ persisted. This kind of bias is probably to be interpreted as an actual bias towards some kind of neutral Schwa vowel [ə]. Most subjects made no use of the option to label stimuli as "non-vowels". In Table 2 the response bias is summarized for all f_{0s} .

Table 2, Distribution of total number of responses, pooled over all f_{0s} , by the 13 most efficient listeners. Within parentheses: Number of "incorrect" responses if /ø/ for /ə/ and /o/ for /u/ is accepted as "correct".

Vowel	e	ø	*	æ	ə	o	y	ɛ	ɔ	i	u
Responses (total)	468	481	153	216	206	177	162	219	140	101	
Responses (incorrect)	367	331	153	118	117	112	94	94	85	74	
				(213)			(57)				

Among the remaining 12 subjects there were some whose performance was not significantly different from random at any f_0 . Some others among those twelve apparently improved their performance during the course of the session. After some training the performance of those subjects would probably have been similar to that of the 13 spontaneously more efficient ones.

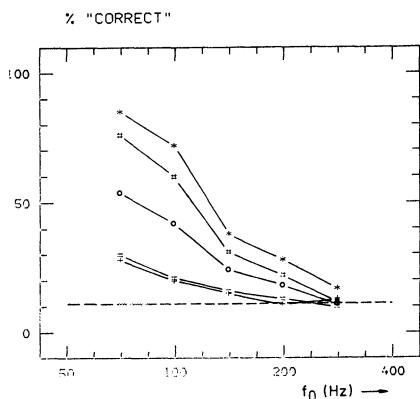


FIGURE 2. Percentage of "correct" identification as a function of f_0 . (*) 6 most efficient listeners; (#) 13 most efficient listeners; (o) All 25 listeners; (=) 12 least efficient listeners; (+) 6 least efficient listeners. Dashed line indicates random performance.

DISCUSSION

The results demonstrate that phase information is sufficient to convey the information necessary to distinguish vowels if f_0 is low enough, i.e. within or below the range typical of adult male speakers. At higher f_0 , this is apparently not the case.

The results agree with the idea that the ear is sensitive to the phase relationships among spectral components within the same critical band (cf. also Rosen, this volume) and capable of a short-time spectral analysis limited by the neural "sampling rate" (at high frequencies) and by the temporal resolution inherent in critical bands (at low frequencies). Phase effects require at least three partials within a critical band in order to affect the modulation wave form. With only two partials any modulation will be sine shaped. However, in phase vowels two partials within the critical band centered at a formant suffice for phase formants to be detectable on the basis of the delay of the amplitude modulation obtained in that critical band as compared with critical bands centered above and below that formant. Two partials within a critical band are also required for the perception of roughness (Terhardt, 1968), and we might say that phase formants can be detected as long as they evoke a sufficient amount of roughness. Table 3 shows the lowest partials for which the frequency distance to the next higher partial is less than one standard critical bandwidth, given the f_0 s used in the present experiment. It can be

Figure 2 summarizes the results in terms of "correct" identifications as a function of f_0 for the most and the least efficient listeners as well as for the whole group of subjects. In this Figure, /ø/- and /o/-responses to intended /u/ and /u/, respectively, have been considered "correct".

The number of "correct" identifications by the subjects with a first language different from Swedish (bilinguals within parentheses) was above average in 5 (+1) cases and below in 1 (+1) case. Experienced phoneticians performed above average in 7 cases and below in 3.

seen that essential formant information is lost with increasing f_0 . F1 (typically between 200 and 800 Hz) is lost in all vowels at $f_0 > 141$ Hz and F2 is lost in back vowels (typically 700 to 800 Hz in [u] and [o]) at $f_0 > 200$ Hz. At $f_0 = 280$ Hz, F2 is also lost in some front vowels. This is compatible with the identification results actually obtained.

Table 3. Frequencies $n f_0$ at which at least two partials can be found within the same standard critical band, specified for the fundamental frequencies f_0 used in the present experiment.

f_0 (Hz)	71	100	141	200	283
partial nr. n	1	2	5	7	7
$n f_0$ (Hz)	71	200	707	1400	2981

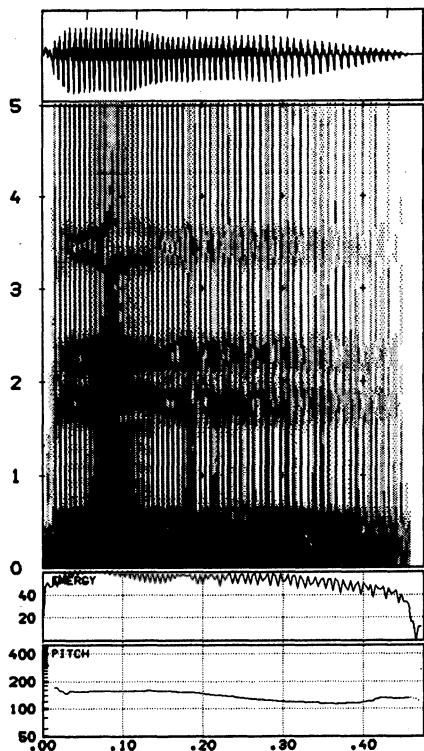


FIGURE 3 (left column), Spectrogram, 0 to 5 kHz, of a phase vowel /ø:/; analysis bandwidth 231 Hz, time window 5 ms. Also shown are oscillogram (above) and fundamental frequency in Hz, logarithmically scaled (bottom). Time scaled in s (bottom).

Phase effects equivalent to those experienced by the sense of hearing appear also on standard sound spectrograms, where phase formants can be seen quite clearly given that a suitable bandwidth of spectral resolution is chosen (see Figure 3). It is hard to see how such phase effects could possibly pass the sense of hearing without being noticed. There is no obvious explanation for the poor performance of some of the present subjects and for the superior performance of non-native speakers of Swedish. Plomp and Steeneken (1969), using non-speech stimuli, also observed significant differences between subjects in susceptibility to phase. They also observed decreased phase effects for increased f_0 . In both respects, the present results appear to be more drastic.

In a preliminary experiment, a provisional one-sample-spoke type of voice source pulse had been used, and the signal was fed through a first order low-pass filter with a limiting frequency equal to f_0 . The phase vowels were identified by one subject. His performance was later found to have been "better" with these stimuli than with their final versions. The use of a more natural type of voice source signal did not improve the performance but only the relevance of the results. During the preliminary experimentation, monaural and binaural presentation through headphones were observed to be equivalent, but presentation through an ordinary "hifi"-loudspeaker in an ordinary room was observed to render the stimuli unintelligible. This was to be expected because of the quite drastic phase distortions introduced in this way.

The recordings of neural activity evoked by vowel sounds in the cat's auditory nerve (Sachs and Young, 1979; Delgutte and Kiang 1984) were obtained presenting vowels with $f_0 = 125$ Hz. The equivalent f_0 of humans would be roughly 60 Hz - considering the critical bandwidths and the upper frequency limit of auditory perception in the two species. This is unusually low but clearly warrants the possibility of short time spectral analysis. The data on neural activity obtained by Young and Sachs (1979) and by Delgutte and Kiang (1984) were, however, not analysed in this way, but phase information was discarded. A representation of the data obtained by Delgutte and Kiang (1984) in the form of a running spectrogram is likely to reveal formant information clearly. This can be traced rudimentarily in their "normalized period histograms" (p.869).

As for the relative importance of phase effects in speech sounds with ordinary amplitude spectra, Plomp and Steeneken (1969) reported comparatively small timbre differences between synthetic vowels consisting either of sine terms or of alternative sine and cosine terms. Carlson, Granström, and Klatt (1979) observed large psychoacoustic but small phonetic changes in timbre when the phases of partials were randomized. Darwin and Gardner (1986) observed a small change in phonetic quality subsequent to a change in the phase of one partial close to F1. In all these studies f_0 was close to 125 Hz. It now appears clear that the choice of f_0 is crucial. Probably, phase is quite important to some listeners when listening to low-pitched speech in a negligibly reverberant environment.

REFERENCES

1. Carlson, R., Granström, B., and Klatt, D. (1980). Vowel perception: the relative perceptual salience of selected acoustic manipulations. STL-QPSR, 3-4/1979, (Royal Inst. Technol., Stockholm), 36-44.
2. Darwin, C.J. and Gardner, R.B. (1986). Mistuning a harmonic of a vowel: Grouping and phase effects on vowel quality. Journal of the Acoustical Society of America, 79, 838-845.
3. Delgutte, B. and Kiang, N.Y.S. (1984). Speech coding in the auditory nerve. Journal of the Acoustical Society of America, 75, 866-886.
4. Fant, G. (1959). Acoustic analysis and synthesis of speech with applications to Swedish. Ericsson Technics 1.
5. Fant, G. (1979). Glottal source and excitation analysis. STL-QPSR 1/ 1979, (Royal Inst. Technol., Stockholm), 85-107.

6. Möller, A. (1974). Dynamic properties of cochlear nucleus units in response to excitatory and inhibitory tones. in E. Zwicker and E. Terhardt (eds.): Facts and Models in Hearing, Berlin, Heidelberg: Springer; 227-240.
7. Plomp, R. and Steeneken, H.J.M. (1969). Effect of phase on the timbre of complex tones. Journal of the Acoustical Society of America, 46, 409-421.
8. Terhardt, E. (1968). über akustische Rauigkeit und Schwankungsstärke. Acustica, 20, 215-224.
9. Sachs, M.B. and Young, E.D. (1979). Encoding of steady-state vowels in the auditory nerve: Representation in terms of discharge rate. Journal of the Acoustical Society of America, 66, 470-479.
10. Young, E.D. and Sachs, M.B. (1979), Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers. Journal of the Acoustical Society of America, 66, 1381-1403.

NONLINEAR RESPONSES IN THE AUDITORY NERVE TO VOWEL-RELATED COMPLEX STIMULI*

J. Wiebe Horst^o, Eric Javel⁺, and Glenn R. Farley⁺

^oInstitute of Audiology
P.O. Box 30 001, Groningen, The Netherlands

⁺Boys Town National Institute
Omaha, Nebraska 68131, USA

INTRODUCTION

Young and Sachs (1979) have shown, in their classical paper on the processing of information on complex vowels in the auditory nerve, how well spectral structure of complex stimuli is retained in temporal aspects of the discharge patterns of populations of single auditory-nerve fibers. They showed that suppression must play an important role in the preservation of the formant structure. In a recent study on the coding of spectral fine structure of complex stimuli in single fibers (Horst et al., 1986a), we were also struck by strong nonlinear aspects of the responses. In order to investigate the mechanisms involved in these nonlinear responses we employed a class of complex stimuli that elicited strongly different nonlinear responses with relatively small changes in the spectral domain. The present paper reports on the responses to one type of stimulus. It is based on a flat spectral envelope. However, the center component is raised in level with respect to the other components, thus simulating a highly simplified formant. We will show that the nonlinear responses agree with Young and Sachs' data and can be explained by the action of a compressive nonlinearity.

METHODS

Responses from single auditory-nerve fibers were obtained from anaesthetized adult cats. Details of the surgical and recording procedures are given by Horst et al. (1986a). After a nerve fiber was encountered, a tuning curve for short-duration tones was taken, which yielded the fiber's characteristic frequency (CF) and approximate discharge-rate threshold. The complex stimuli were multicomponent tones defined by a center frequency F (usually equal to CF) and a spectral spacing factor N. The ratio F/N defined the fundamental frequency F0. Stimuli consisted of successive equal-amplitude harmonics of F0. Spectra were one octave wide, centered geometrically about a fiber's CF. Stimulus levels were expressed in terms of the level of each component, in dB SPL (re 20 μ Pa). The data presented here

*

This work was supported by grant NS-14880 to EJ by NINCDS and a grant to JWH by the Netherlands Organization for the Advancement of Pure Research. P. Hoepfner and W.F. Cleary assisted in collecting and analyzing the data. A. Mensink skillfully prepared the figures.

resulted from a series of experiments where we varied several aspects of the stimuli, including the phase relations between the spectral components and the level ΔL of the center component with respect to the other components, which were taken at equal levels. An example of the stimuli is given in Fig. 1. Here, all components were presented in cosine phase, and the CF component level was raised in steps of 2 dB with respect to the level of the other components. Responses of single nerve fibers were analyzed by compiling period histograms synchronized to the period of the waveform fundamental and determining Fourier transforms of these period histograms.

RESULTS

Period histograms in response to a stimulus with $\Delta L = 2$ dB are shown in Fig. 2. The responses are from a fiber with a CF of 1163 Hz. For N we chose a value of 16, so F0 was 72.7 Hz. At the lowest stimulus levels shown, the period histograms are strongly modulated, in agreement with the stimulus waveform. For increasing stimulus level the shape of the period histogram changes gradually. At 30 dB SPL small peaks arise in the period histogram in addition to the large peaks. As stimulus level increases, the relatively small peaks keep growing. This was often observed in conjunction with a decrease of the largest peaks. Usually, a drastic change of the period histogram was observed at the highest stimulus levels, indicating a redistribution of firing within the period of the fundamental. This can be clearly seen in the 70-dB case in Fig. 2.

In order to determine the response of the nerve fibers to single components in the stimulus spectrum, we employed Fourier analysis of the period histograms. For the data of Fig. 2, the Fourier spectra are shown in Fig. 3. At the lowest stimulus levels the responses are mainly determined by the stimulus spectrum and the filter characteristics of the nerve fiber (see also Horst et al., 1985). Increasing the level results in a gradual suppression of the responses to frequencies other than CF. At the highest stimulus level shown here, the response spectrum is mainly a response to CF, the frequency

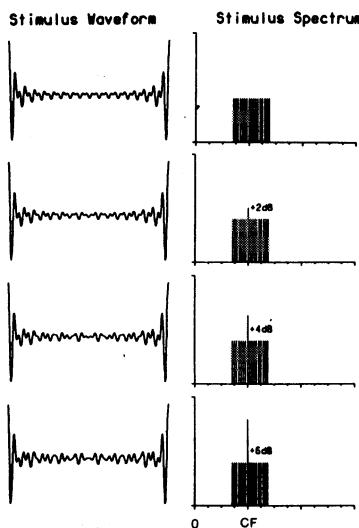


FIGURE 1, Examples of stimulus waveform and stimulus spectra. The stimulus waveform at the top refers to the case that all components had equal level and cosine starting phase. Going from top to bottom, the CF-component level increased in steps of 2 dB. Stimulus spectra were one octave wide and geometrically centered at CF. In this case N=24, i.e. the fundamental frequency F0 equalled CF/24.

of the more intense stimulus component. This agrees of course with the shape of the period histogram, which at this stimulus level closely resembled a pure-tone response.

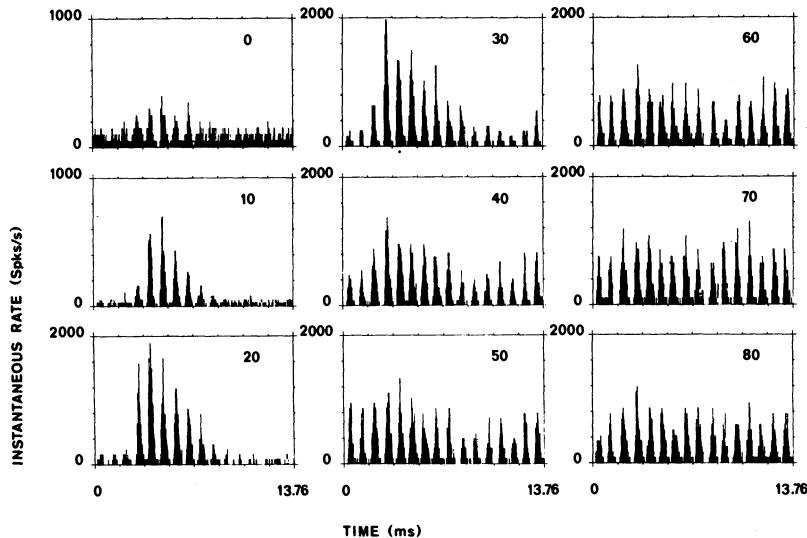


FIGURE 2, Period histograms in response to a stimulus with $CF=1163$ Hz and $F_0=CF/16=72.7$ Hz. Thus, the fundamental period was 13.76 ms. Stimulus level (indicated in each panel) went up in steps of 10 dB. The CF component was in each case 2 dB higher in level, that is ΔL was 2 dB.

The data presented so far suggest that the level-dependent nonlinear behavior of the responses is strongly related to the decrease in modulation of the period histograms with increasing stimulus level. This is in agreement with the action of a compressive nonlinearity. The involvement of a compressive nonlinearity was checked by means of different stimulus types. Clearly, stimulus types with strongly modulated waveforms will be more affected by a compressive nonlinearity than stimulus types with more even waveform envelopes. This was checked by adequate manipulation of the stimulus phase spectra (Horst et al., 1986b). Additionally, our experiments showed that nonlinear behavior was stronger for increased waveform modulation as a consequence of higher values of N . Thus, we focused on stimuli with relatively high N values, in order to get more insight into the nonlinearities involved. This explains the relatively low fundamental frequencies of our stimuli.

Some data from a unit with a more realistic value of F_0 are shown in Fig. 4. The CF of this fiber was 3242 Hz. We stimulated with a signal with $N=32$, so the F_0 was 101.3 Hz. Note that the response spectra show clearly lower values of the synchronization index, as compared to the data in Fig. 3. This is caused by the higher frequencies of the stimulus components. Generally, synchronization for stimuli around 3 kHz is considerably lower than for stimuli near 1 kHz (Johnson, 1980). Nevertheless, it is clear that for higher stimulus levels the response becomes increasingly dominated by the CF

component, which was in this case also 2 dB in level above the neighbouring stimulus components. In this respect, the data of this nerve fiber are in close agreement with the data in Fig. 3.

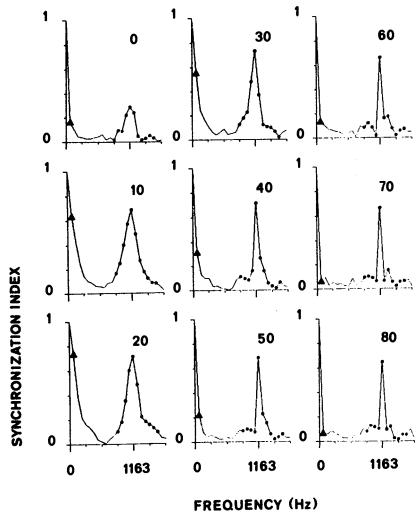


FIGURE 3, Fourier transforms of the period histograms shown in figure 2. The magnitudes of spectral components are expressed in terms of the synchronization index, which is equal to the Fourier coefficient at a given frequency divided by the DC component. Triangles represent synchronization indices at the fundamental frequency F_0 . The stimulus level is indicated in each panel in dB. ΔL was 2 dB.

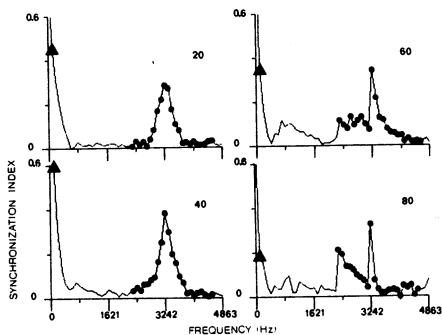


FIGURE 4, Fourier transforms of [period histograms for a nerve fiber that was stimulated with $CF=3242$ Hz and $N=32$. F_0 was 101.3 Hz. Usage of symbols is the same as in figure 3. Again, ΔL was 2 dB.

The domination of the response at CF generally becomes more prominent for stimuli with higher values of ΔL . This is illustrated in Fig. 5, where we have plotted the response ratio of the more intense component and its nearest neighbours, expressed in dB. For this unit ($CF=662$ Hz; $N=24$) we stimulated with ΔL values ranging from 0 to 12 dB. At stimulus levels of 10 and 20 dB the response data are in fair agreement with a linear transduction process. At higher stimulus levels the responses to the more intense component at CF are clearly enhanced with respect to the responses to the neighbouring frequency components. Another conclusion is that the domination of the response gradually shifts to lower stimulus levels as ΔL is increased. This shows the intrinsically nonlinear behavior of single-fiber responses. All data are quite different from the data for $\Delta L=0$. It is very interesting to

note that a difference of only 2 dB in the level of one component induces so strongly nonlinear responses. It is also of relevance that the increased component in the case of Fig. 5 was slightly lower (about 100 Hz) in frequency than the best frequency of this nerve fiber. This indicates that the effect of domination of the response spectrum is not limited to the best frequency of the nerve fiber.

DISCUSSION

The data presented here are in global agreement with single-fiber data from Young and Sachs (1979) for synthetic vowels. Young and Sachs related their data to various mechanisms, including suppression and expansive non-linearities. Hall (1980) showed that a cochlea model incorporating compressive nonlinearities gives rise to enhancement of the formants.

Compressive nonlinearities had been discussed before in cochlea models (Engebretson et al., 1968; Duifhuis, 1976). One possible consequence of a compressive nonlinearity is suppression. Houtgast (1974) has shown in psychophysical experiments that the suppression mechanism may give rise to contrast enhancement for artificial vowels. Our single-fiber data show that strong contrast enhancement may occur for level differences in the stimulus spectrum as small as 2 dB. The frequency range of the present data clearly covers the range of the second and third formants. The data in Fig. 2 exhibiting a strong decrease in modulation of the envelope of the period histogram are in excellent agreement with the action of a compressive nonlinearity. Simulations that we carried out show that the nonlinearity involved is more compressive than in existing cochlea models. This indicates an extra stage of compression at the hair cell-nerve fiber synaps, involving effects of saturation of average and instantaneous spike rate. These effects will be discussed in a future paper.

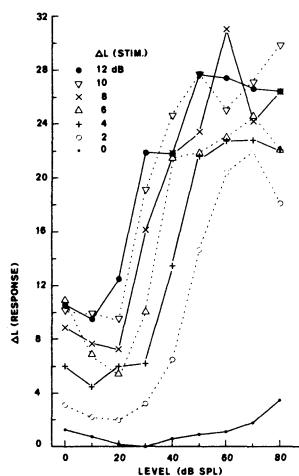


FIGURE 5, Dependence of the amount of domination on the stimulus level. The ordinate gives the ratio (expressed in dB) of the synchronization index to the more intense component and the average synchronization index to the neighbouring components. The center frequency was 662 Hz, N=24, and L of the stimulus ranged from 0 to 12 dB.

The present data show that mechanisms akin to compressive nonlinearities can explain at least partially why the temporal responses at high stimulus levels observed by Young and Sachs (1979) occurred almost exclusively at the formant frequencies. In some fibers, however, they observed responses to only two stimulus components that do not fit into this scheme: at low stimulus levels the responses were dominated by the component nearest to the characteristic frequency of that particular nerve fiber, whereas at high stimulus levels the response was dominated by the center component of the nearest formant. In order to explain that kind of nonlinear behavior, we also have to take into account a change of the bandwidth and/or the characteristic frequency of the nerve fiber (e.g. Müller, 1977; Smoorenburg et al., 1977). Further research is needed to decide to what extent each of the three mechanisms (compressive nonlinearity, bandwidth change, shift of characteristic frequency) is involved in the change with level of single-fiber responses to complex stimuli in general and synthetic vowels in particular.

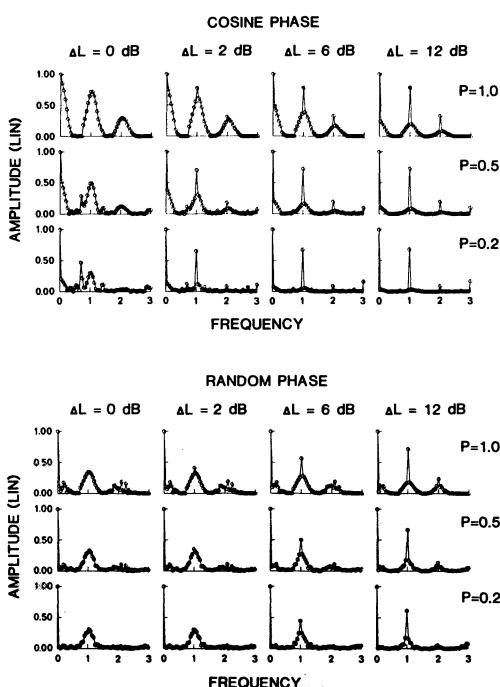


FIGURE 6 (left column), Response spectra from a model consisting of a linear filter (Gaussian shape, $Q_{10}=2$), a nonlinear input-output function $V_0=V_iP$, and a halfwave rectifier. Different columns refer to different format heights. The rows refer to different powers of the nonlinearity. The top part of the figure refers to the stimulus spectrum with all components added in cosine phase relation. The bottom part refers to a randomly chosen phase relation.

Horst et al. (1985, 1986b) have shown that nonlinear effects in single-fiber responses are very dependent on the stimulus phase spectrum. Actually, these effects are greatly reduced for randomly chosen phase spectra. This could easily be simulated with a simple model. The model consists of a linear filter tuned at CF, a halfwave rectifier, and a nonlinear input-output function according to $V_0=V_iP$, where $p>0$. For p we chose various values: $p=1$ refers to the linear case; the smaller p is, the more compressive the nonlinearity is ($p>1$ refers to an expansive nonlinearity, discussed by Horst et al., 1986c). The most important aspects of the single-fiber responses are borne out (Figure 6): when all stimulus components were added in the same phase relation, the

compressive nonlinearity, discussed by Horst et al., 1986c). The most important aspects of the single-fiber responses are borne out (Figure 6): when all stimulus components were added in the same phase relation, the

response spectrum for the case of the most compressive nonlinearity was strongly dominated by the "formant" even when it was only 2 dB high. When the stimulus components were added in a randomly chosen phase relation, no clear influence of the compressive nonlinearity was observed. Obviously, the compressive nonlinearity does not appreciably influence the spectrum of a waveform with a low peak factor.

In the production of synthetic speech it is sometimes desirable to produce waveforms with a low peak factor (see e.g. Schroeder 1985, pp. 296-299). However, this may have consequences for the intelligibility of (the voiced parts of) speech. If enhancement of formants in the temporal aspects of single-fiber responses is actually needed to retain good intelligibility at high stimulus levels, then it may well be that the intelligibility of speech stimuli with low peak factors (not producing enhancement of formants) deteriorates with increasing stimulus level. This aspect of synthetic speech deserves further psychoacoustic investigation.

REFERENCES

1. Duifhuis, H. (1976). Cochlear non-linearity and second filter: Possible mechanism and implications. *J. Acoust. Soc. Am.*, 59, 408-423.
2. Engebretson, A.M. and Eldredge, D.H. (1968). Model for the nonlinear characteristics of cochlear potentials. *J. Acoust. Soc. Am.*, 44, 548-554.
3. Hall, J.L. (1980). Frequency selectivity of the cochlea for formant peaks at high signal levels. *J. Acoust. Soc. Am.*, 68, 480-481.
4. Horst, J.W., Javel, E., and Farley, G.R. (1985). Extraction and enhancement of spectral structure by the cochlea. *J. Acoust. Soc. Am.*, 78, 1898-1901.
5. Horst, J.W., Javel, E., and Farley, G.R. (1986a). Coding of spectral fine structure in the auditory nerve. I. Fourier analysis of period and interspike interval histograms. *J. Acoust. Soc. Am.*, 79, 398-416.
6. Horst, J.W., Javel, E. and Farley, G.R. (1986b). New effects of cochlear nonlinearity in temporal patterns of auditory nerve fiber responses to harmonic complexes. In: J.B. Allen, J.L. Hall, A. Hubbard, S.T. Neely, and A. Tubis (Eds.), Peripheral Auditory Mechanisms, 298-305. Springer-Verlag, Berlin, Heidelberg, New York.
7. Horst, J.W., Javel, E., and Farley, G.R. (1986c). Effects of phase and amplitude spectrum in the nonlinear processing of complex stimuli in single-fibers of the auditory nerve. In: B.C.J. Moore and R.D. Patterson (Eds.) Auditory Frequency Selectivity. Plenum, New York, in press.
8. Houtgast, T. (1974). Auditory analysis of vowel-like sounds. *Acustica*, 31, 320-324.
9. Johnson, D.H. (1980). The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones. *J. Acoust. Soc. Am.*, 68, 1115-1122.
10. Müller, A.R. (1977). Frequency selectivity of single auditory-nerve fibers in response to broadband noise stimuli. *J. Acoust. Soc. Am.*, 62, 135-142.
11. Schroeder, M.R. (1985). Number theory in science and communication. Springer Verlag, Berlin, Heidelberg, New York, Tokyo.

12. Smoorenburg, G.F. and Linschoten, D.H. (1977). A Neurophysiological study on auditory frequency analysis of complex tones. In: E.F. Evans and J.P. Wilson (Eds.), Psychophysics and Physiology of Hearing, 175-183. Academic, London.
13. Young, E.D. and Sachs, M.B. (1979). Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory nerve fibers. J. Acoust. Soc. Am., **66**, 1381-1403.

DISCUSSION OF PHYSIOLOGICAL CORRELATES OF SPEECH PERCEPTION

Guido F. Smoorenburg

Laboratory of Experimental Audiology, Department of
Otorhinolaryngology, University Hospital, Utrecht
and

TNO Institute for Perception, Soesterberg, The Netherlands

INTRODUCTION

Recently there has been a growing interest on the part of auditory physiologists in studying neuronal responses to speech, or speech-like, stimuli. This, in turn, has aroused the interest of speech scientists in auditory physiology. Although this growing mutual interest is stimulating, we should realize that studies of the mechanism of hearing, primarily by electrophysiological means, can contribute only in a limited way to our understanding of speech perception.

A central theme of this workshop is the question of whether (or not) we are able to identify perceptual mechanisms that are specific to the perception of speech as opposed to the perception of other sounds. Since speech perception is a human affair, such a question can, of course, not be answered by electrophysiological experiments in animals. Animal experiments do show, however, how speech stimuli are represented in the peripheral auditory system of mammals and what information about the speech stimuli will be available to the human brain. In addition, studies of signal processing by the auditory systems in animals, using stimuli that are relevant to these animals, may provide insight into neural mechanisms that may also play a part in speech perception by humans. Such mechanisms are, for example, feature extraction and plasticity, the latter referring to the ability of neuronal networks to organize themselves to respond especially to frequently occurring stimuli.

FEATURE EXTRACTION

Probably the most widely known example of feature extraction by the brain derives from the work of the Nobel laureates Hubel and Wiesel (1968, 1974, 1979). These investigators found cells in the visual cortex of cat and monkey that responded selectively to line segments with a certain orientation, while no response was found for other orientations. At more peripheral levels of the visual system this type of coding was not found. Thus, the results suggested that an important feature of visual stimuli is extracted from the evoked neural activity at the cortical level.

Whereas in these mammals there was no evidence of feature extraction at the peripheral level of the visual system, feature extraction at this level was found in more primitive animals. In the frog, Lettvin et al. (1959) found feature extraction at the level of the retina. In their classical paper entitled "What the Frog's Eye Tells the

"Frog's Brain" they concluded on the basis of the electrophysiological response that in the frog there are four types of optic nerve fibers. The different fibers responded to edges (sustained contrast detectors), to the movement of edges (moving-edge detectors), to the curvature of an object (convexity detectors), and to a sudden reduction of illumination (net dimming detectors), respectively. Lettyin and his collaborators felt that these responses had "much more the flavor of perception than of sensation".

The study by Lettyin et al. motivated Frishkopf et al. (1968) to look for similar mechanisms in the peripheral hearing of the bullfrog. They found that fibers innervating the basilar papilla respond to frequencies primarily in the range from 1000 to 2000 Hz, while fibers innervating another part of the frog's peripheral auditory system, the amphibian papilla, respond to frequencies in the range from about 200 to 500 Hz. The fibers in the first population responded in a simple manner without any signs of lateral inhibition and similarly to cochlear nerve fibers. The fibers in the second population showed a more complex response pattern; the response to low-frequency stimuli could be inhibited by additional sound energy in the range from 500 to 1000 Hz. It was striking to note that vocalizations of the bullfrog, especially the mating call, have two spectral peaks; one between 200 and 400 Hz and the other centered around 1400 to 1600 Hz, the exact frequency depending on the individual. These two peaks (formants) perfectly coincide with the two frequency regions to which the bullfrog ear is most sensitive. Moreover, between these two peaks the mating call of the bullfrog shows a pronounced minimum in the spectral energy distribution, suggesting that the mating call will not activate the inhibitory system of the amphibian papilla. Thus, these results suggest that the peripheral auditory system of the bullfrog is optimally equipped to respond to the mating call of this species. Further behavioral experiments in which synthetic mating calls were used showed that this response is highly selective. In particular, reducing the pronounced formant structure by adding sound energy in the region of the spectral minimum (in the inhibitory area) reduced the behavioral response markedly.

Having concluded that the peripheral auditory system of the bullfrog is optimally equipped to detect the mating call, it is obvious that subsequently Frishkopf et al. tried to find neurons at a more central level of the auditory system that would respond specifically to this call. In this attempt they were not successful. It is, of course, possible that such neurons were not found although they are present. However, one also may question the idea that the perception of significant sounds should be related to activity in specific neurons. A specific percept could also be related to a certain pattern of activity across a neural population, the pattern possibly changing with time.

In search of feature extractors in the auditory system of animals it seems only logical to use stimuli that are relevant to these animals. Nevertheless, some investigators have looked also for specific reactions to human speech stimuli. Keidel (1974), for example, reported a consonant-detector neuron at the level of the medial geniculate in the cat that responded to the consonant /f/ in 'fein' and not to 'mein' and 'dein'. The neuron did not respond to simple sinusoidal stimuli nor to frequency-modulated auditory stimuli. It seems to me, however, that it is rather speculative to label such a neuron a detector of a

particular consonant. The neuron might have been selective of a less specific set of complex stimuli all containing high-frequency energy. A similar study in which speech stimuli were presented to birds which are able to imitate human voices was reported by Langner et al. (1981). They concluded that some units responded in a categorical fashion to a restricted set of speech stimuli. Another member of this group, however, reported less specific responses to typical calls of the bird itself (Bonke et al., 1981). The frequency ranges of these calls overlapped, and the neuronal activity evoked by these calls could be explained from the tonotopic responsiveness of the 'cortical' area investigated.

SELF-ORGANIZING SYSTEMS

The previous section presented an example of specialized signal coding at the peripheral level of the auditory system. The examples from vision research given before suggest that specialized signal coding at the peripheral level is likely to be found only in lower species which have to deal with only a limited set of relevant stimuli. In most mammals, for example, we don't find specialized frequency regions at the peripheral level of the auditory system. (The highly specialized ears of bats are, of course, an exception.) Following the example set in vision research by Hubel and Wiesel, we may now look for feature extraction at more central levels of the auditory pathway. We have, however, no *a priori* knowledge of what kind of neural activity would correspond to a certain auditory feature. A certain feature might be coded in the response of a particular cell but also, as stated before, it might be coded in a specific response pattern across a population of neurons, the pattern possibly changing with time. (See for example the suggestion by Traunmüller in this workshop.) In order to find physiological correlates of perception a different aspect of neural information processing may therefore prove to be more informative, namely the ability of the neural system to organize itself in reaction to stimuli. If we trained animals to discriminate certain stimuli and studied the neurophysiological response to these stimuli of trained animals in relation to those of untrained animals or even animals deprived of these stimuli, the differences in response could shed some light upon the question of what neural activity pertains to certain percepts.

An impressive example of the plasticity of cortical coding was given by Merzenich (1983) for the somatosensory system. He showed that the representations of the skin surfaces of monkeys mapped in the cortex are dynamically maintained, and are alterable in adults. He further suggested that anatomical and physiological mapping studies in the auditory system strongly indicate that detailed central auditory neural representation of sound location information is also achieved by an experience-dependent process. We thus arrive at a concept of the auditory system in which there is on the one hand an inborn, hard-wired, system of frequency analysis, the tonotopical representation stretching all the way from the cochlea up to the cortex, and on the other hand a flexible mapping system of spatial localization of the sound source.

Speech perception may be coded in a similarly flexible way. It is, after all, an acquired ability. It would therefore be interesting to extend this type of research to sets of animal vocalizations. It would

be interesting to see whether such patterns develop during exposure to species-relevant stimuli. The neurophysiological measurements should then not be limited to single neurons, but should also include response patterns across neuronal populations. In this respect the experiments by Eggermont et al. should be mentioned. They have started simultaneous recordings from a number of neighboring units (Eggermont et al., 1983; Epping et al., 1984).

Summarizing we may conclude that there is physiological evidence of feature extraction in sensory systems, particularly in the visual system. However, it is only a preconceived notion that physiological correlates of the perception of certain features are to be found in the response of specific cells. A more general approach to find these physiological correlates could be based on the development of neural response patterns while animals are being trained to respond behaviorally to certain relevant stimuli. Responses in untrained or sensorily deprived animals can then serve as a control.

PERIPHERAL CODING OF SPEECH STIMULI

Signal coding in the peripheral system of mammals is essentially a process of separating stimuli into their frequency components. This process is quite uniform across a certain frequency range. The coding process is studied most adequately by means of analytic signals such as sinusoids and clicks. There is no specific coding of speech stimuli at this level of the mammalian auditory system that would justify the use of speech or speech-like stimuli. However, the occurrence of combination tones (Goldstein, 1967; Goldstein and Kiang, 1968; Smoorenburg, 1972; and Smoorenburg et al., 1974) and of suppression of tone-evoked neuronal activity by other frequency components (Sachs and Kiang, 1968 and Houtgast, 1972) shows that peripheral frequency analysis is a highly nonlinear process. This implies that the responses to complex stimuli are difficult to predict. One may therefore choose to study directly the neuronal responses to speech stimuli in order to get better insight into the information about these stimuli that is lost in the peripheral coding process and the information still available to the brain.

The primary problem in trying to understand the peripheral coding process is the limited dynamic range of the neuronal response. In most nerve fibers the discharge rate changes from the spontaneous rate to the maximum rate over a level range of only 30 to, at most, 50 dB, while the discharge thresholds are restricted to a low-level range of about 30 dB. This implies that the discharge rates of most units are saturated when the stimulus level exceeds about 70 dB SPL. Consequently, much information about spectral details will be lost at high sound levels. In a study on encoding of steady-state vowels in the auditory nerve, Sachs and Young (1979) indeed concluded that peaks in the neuronal rate profiles of several vowels disappear at stimulus levels above about 70 dB SPL. However, there is a small population of nerve fibers with dynamic ranges in excess of 50 dB (Palmer and Evans, 1979) and an even smaller population with thresholds up to 80 dB SPL (Liberman, 1978). The fibers in these populations all have a low spontaneous discharge rate. Although one should reckon with the possibility that these fibers are damaged, present evidence suggests that they are always present in normal ears. Sachs and Young (1979) pointed out that the rate profiles of these low

spontaneous units (< 1 spike per second) may retain spectral peaks at higher levels than 70 dB. However, it is not yet clear how the perceptual finding of almost level-independent sound discrimination over a wide range of stimulus levels can be explained by combining the neuronal firing pattern of this small population with the firing pattern of the larger population showing extensive saturation. Delgutte, for example, shows that the just-noticeable change in stimulus level, found psychoacoustically to be almost constant as a function of absolute level, cannot be explained by combining the response characteristics of different fiber populations (this volume). These conclusions are based on the assumption that the neural activity in the different fiber populations is added on a basis of equality. The low-spontaneous units may have a special function at higher centers of the auditory pathway, but this has not yet been demonstrated.

Since it is questionable whether rate profiles contain sufficient information about speech signals at higher sound levels, the temporal fine structure of fiber discharges was included in the analyses. Rose et al. (1967) showed that discharges of primary nerve fibers are phase-locked to the stimulus up to frequencies of about 4000 Hz. A stimulus may become apparent in the temporal fine structure of the discharge at levels as much as 20 dB below the threshold above which the firing rate increases. This may imply an increase of dynamic range. Moreover, at saturation levels the stimulus frequency is still well represented in the discharge pattern, while the peak in the rate profile is lost. Young points out that the 'bushy' cells in the cochlear nucleus are particularly suited to convey detailed temporal information (this volume). Rose et al. (1974) studied the response of these cells to two-tone frequency stimuli. Already in their study the principal property of temporal profiles, found later for speech stimuli (Young and Sachs, 1979; Delgutte and Kiang, 1984), was evident; responses are subject to a gain control, particularly at saturation levels, that does not affect the relative strength of the frequency components as they are present in the temporal discharge patterns. Thus, the stimulus spectrum is better preserved in temporal profiles than in rate profiles. The nonlinear effects mentioned before complicate this picture to some extent (Horst, this volume).

The temporal discharge pattern is usually measured in relation to a certain phase of the stimulus (e.g. in period histograms), which in essence provides us with a kind of cross-correlation function of stimulus and response. Such a function may be used to analyse the cochlear coding mechanism. With respect to speech perception, however, we should get an appropriate impression of the information available to the brain. Since the external phase reference is not available to the brain it would be more adequate to use an auto-correlation type of function, e.g. interval histograms.

The original work by Rose et al. received some criticism. Some auditory physiologists felt that the temporal fine structure of the neuronal discharges is only a side effect of the hair-cell transducer mechanism and not relevant to hearing. They wondered whether neurons in higher parts of the auditory system would be capable of handling this temporal information. This criticism is not valid when the temporal fine structure is used to unravel the cochlear signal-coding mechanism. It is valid, however, when we try to get an idea of the information about speech stimuli that is available to the

brain. At each synapse information will be lost because of time jitter in the signal transfer. Thus, at a relatively low level of the auditory system one might expect a transformation of the temporal fine structure to a code less sensitive to time jitter. The time-to-place coding in localization is a persuasive example. However, such recoding has not yet been demonstrated for monaural frequency representations. The poor frequency discrimination in deaf people provided with cochlear implants contradicts this idea. The temporal coding of stimulus frequency in the auditory nerve of these patients is very accurate.

Summarizing, we may conclude that saturation of the firing rate of auditory nerve fibers does not allow adequate rate coding of speech sounds at higher stimulus levels. The spectral structure is better preserved in the timing of the neuronal discharges but to the brain this may be irrelevant information. The above conclusions constitute a problem not specific to coding of speech stimuli. Physiological correlates of more simple, basic psychoacoustic measures should be studied. A trend in this direction can be noticed, e.g. correlates of intensity discrimination (Delgutte, this volume) and of masking (Young and Barta, 1986; Smoorenburg and Kloppenburg, 1986). Moreover, more attention should be payed to the question of whether anesthesia has decreased the dynamic range of the nerve fibers by (partly) blocking the efferent system.

REFERENCES

1. Bonke, D., Bonke, B.A., Langner, G., and Scheich, H. (1981). Some aspects of functional organization of the auditory neostriatum (field L) in the guinea fowl. In: J. Syka and L. Aitkin (Eds.) Neuronal Mechanisms of Hearing, 323-327. Plenum Press, New York.
2. Delgutte, B. and Kiang, N.Y.S. (1984). Speech coding in the auditory nerve I. Vowel-like sounds. J. Acoust. Soc. Am., 75, 866-878.
3. Eggermont, J.J., Epping, W.J.M., and Aertsen, A.M.H.J. (1983). Stimulus dependent neural correlations in the auditory midbrain of the grassfrog (*Rana temporaria* L.). Biol. Cybernetics, 47, 103-117.
4. Epping, W., Boogaard, H. van den, Aertsen, A., Eggermont, J.J. and Johannesma, P. (1984). The neurochrome; an identity preserving representation of activity patterns from neural populations. Biol. Cybernetics, 50, 235-240.
5. Frishkopf, L.S., Capranica, R.R., and Goldstein, M.H. (1968). Neural coding in the bullfrog's auditory system, a teleological approach, Proc. IEEE, 56, 969-980.
6. Goldstein, J.L. (1967). Auditory nonlinearity. J. Acoust. Soc. Am., 41, 767-789
7. Goldstein, J.L. and Kiang, N.Y.S. (1968). Neural correlates of the auditory combination tone 2f1-f2. Proc. IEEE, 56, 981-992.
8. Houtgast, T. (1972). Psychophysical evidence for lateral inhibition in hearing. J. Acoust. Soc. Am., 51, 1885-1894.
9. Hubel, D.H. and Wiesel, T.N. (1968). Receptive fields and functional architecture of monkey striate cortex, J. Physiol., 195, 215-243.

10. Hubel, D.H. and Wiesel T.N. (1974). Sequence regularity and geometry of orientation columns in the monkey striate cortex. *J. Comp. Neur.*, 158, 267-294.
11. Hubel, D.H. and Wiesel, T.N. (1977). Functional architecture of macaque visual cortex. *Proc. R. Soc. Lond.*, B 198, 1-59.
12. Keidel, W.D. (1974). Information processing in the higher parts of the auditory pathway. In: E. Zwicker and E. Terhardt (Eds.), Facts and Models in Hearing, 216-226. Springer Verlag, Berlin.
13. Langner, G., Bonke, D., and Scheich, H. (1981). Selectivity of auditory neurons for vowels and consonants in the forebrain of the Mynah bird. In: J. Syka and L. Aitkin (Eds.), Neuronal Mechanisms of Hearing, 323-327. Plenum Press, New York.
14. Lettin, J.Y., Maturana, H.R., McCulloch, W.S., and Pitts, W.H. (1959). What the frog's eye tells the frog's brain. *Proc. IRE*, 47, 1941-1951.
15. Liberman, M.C. (1978). Auditory-nerve response from cats raised in a low-noise chamber. *J. Acoust. Soc. Am.*, 63, 442-455.
16. Merzenich, M.M. and Jenkins, W.M. (1983). Dynamic maintenance and alterability of cortical maps in adults; some implications. In: R. Klinke and R. Hartmann (Eds.), Hearing-Physiological Bases and Psychophysics, 162-168. Springer Verlag, Berlin.
17. Palmer, A.R. and Evans, E.F. (1979). On the peripheral coding of the level of individual frequency components of complex sounds at high sound levels. *Exp. Brain Res., Suppl. II*, 19-26.
18. Rose, J.E., Brugge, J.F., Anderson, D.J., and Hind, J.E. (1967). Phase-locked response to low-frequency tones in single auditory nerve fibers of the squirrel monkey. *J. Neurophysiol.*, 30, 769-793.
19. Rose, J.E., Kitzes, L.M., Gibson, M.M., and Hind, J.E. (1974). Observations on phase-sensitive neurons of anteroventral cochlear nucleus of the cat: nonlinearity of cochlear output. *J. Neurophysiol.*, 37, 218-253.
20. Sachs, M.B. and Kiang, N.Y.S. (1968). Two-tone inhibition in auditory-nerve fibers. *J. Acoust. Soc. Am.*, 43, 1120-1128.
21. Sachs, M.B. and Young, E.D. (1975). Encoding of steady-state vowels in the auditory nerve: representation in terms of discharge rate. *J. Acoust. Soc. Am.*, 66, 470-479.
22. Smoorenburg, G.F. (1972). Combination tones and their origin. *J. Acoust. Soc. Am.*, 52, 615-632.
23. Smoorenburg, G.F., Gibson, M.M., Kitzes, L.M., Rose, J.E., and Hind, J.E. (1976). Correlates of combination tones observed in the response of neurons in the anteroventral cochlear nucleus of the cat. *J. Acoust. Soc. Am.*, 59, 945-962.
24. Smoorenburg, G.F. and Kloppenburg, B.A.M. (1986). Single-neuron tuning curves measured with psychoacoustic masking paradigms. In: B.C.J. Moore and R.D. Patterson (Eds.), Auditory Frequency Selectivity, NATO-ASI Series.
25. Young, E.D. and Sachs, M.B. (1979). Representation of steady-state vowels in the temporal aspects of the discharge patterns of auditory-nerve fibers. *J. Acoust. Soc. Am.*, 66, 1381-1403.
26. Young, E.D. and Barta, E. (1986). Rate responses of auditory nerve fibers to tones in noise near masked thresholds. *J. Acoust. Soc. Am.*, 79, 426-442.

GENERAL DISCUSSION OF SESSION 5: PHYSIOLOGICAL CORRELATES OF SPEECH PERCEPTION

Chairman: Guido F. Smoorenburg

The major topics of the general discussion were the relevance of physiological experiments in animals to human speech perception, the problem of the limited dynamic range of neuronal discharge rates, and the issue of rate-place coding versus time-place coding. The topics of feature extraction by neuronal systems and plasticity of the brain, both mentioned in the discussion paper presented by the chairman, were given less priority in the general discussion, because they were not covered by the other papers in this session. However, there was firm interest in the latter topics. In particular with respect to feature extraction the question was asked whether the high accuracy found neurophysiologically in spectral and temporal coding should be taken into account when defining speech boundaries in, for example, formant transition and voice-onset-time studies. The general feeling was that this should be done. The three major topics are reported below.

HUMAN SPEECH PERCEPTION VERSUS ANIMAL PHYSIOLOGY

There was general agreement that at the level of the cochlea and the eighth nerve there is little difference among the various mammals used as experimental animals (excluding highly specialized mammals such as bats and whales). This holds with respect to the anatomy, the physiology and to some extent with respect to psycho-acoustic behavior. The behavioral studies show that sound intensity discrimination in animals is close to that of humans, frequency resolution is also close, but frequency discrimination may in some species differ by a factor of 10. In cat, however, both frequency resolution in terms of critical bandwidth and frequency discrimination are about twice as poor. If we allow for a species-dependent scaling factor in the frequency domain the results may be very similar. The aim of studying neurophysiological responses to speech stimuli is, of course, not the precise mapping of speech stimuli but trying to understand how, for example, the ear separates speech from ambient noise or interfering voices, and how input patterns are represented and might be matched to stored representations of sounds. Although neurophysiological recordings at peripheral levels of the auditory system may provide us with a fair picture of the coding of speech stimuli, we should realize that this is not necessarily so at higher levels of the system. As research proceeds, higher levels of the auditory system will inevitably be investigated and we should ask ourselves, well ahead of this development, at what level of the auditory system responses become animal-specific and what the physiological correlates of perception may be. Training animals to

respond behaviorally to human speech sounds may become an important aspect of these physiological studies.

DYNAMIC RANGE OF NEURONAL DISCHARGES

From psycho-acoustic studies it is clear that there is no sudden decrease in speech discrimination above some critical level, say 70 dB SPL. There is some decrease in performance with increasing level which can be related to upward spread of masking (upward along the frequency scale). We should take into account, however, that in awake humans the acoustic reflex may improve speech discrimination at the highest sound levels. In patients with acoustic-reflex defects degradation of vowel discrimination was found at 90-100 dB SPL. By and large, the psycho-acoustic results strongly suggest that the peripheral auditory system is capable of maintaining frequency resolution up to, at least, 80 dB SPL. Thus, detailed spectral information must somehow be present in the response of the eighth nerve up to, at least, that level. In this respect, the role of the low-spontaneous eighth-nerve fibers in relaying detailed spectral information is interesting. It is also interesting to focus on intensity discrimination at low levels. The neurophysiological data suggest that psycho-acoustic performance could be better than actually measured. We should always reckon with the possibility that the central processor is inefficient.

RATE VERSUS TIME CODING

A question frequently asked is whether detailed information about the spectrum of the stimulus must be coded in the timing of the neuronal discharges, because in most eighth-nerve fibers the dynamic range of discharge rate is too limited, or whether the small number of low-spontaneous fibers showing higher thresholds and larger dynamic ranges relay sufficiently detailed spectral information up to the highest stimulus levels. Present data suggest that there may be adequate rate-place information in the low-spontaneous units. The timing of the neuronal discharges is important for the sensation of beats, roughness, rattle-pitch, and, of course, lateralization. Recent data from cochlear implant patients also bear on the question of the relevance of the time pattern of neuronal discharges for pitch perception. Frequency discrimination in these patients is about an order of magnitude worse than in normals, although timing acuity of the neuronal discharges in these patients is probably at least as good and perhaps even better than in normals. However, no conclusive evidence can be drawn from these observations. Poorer frequency discrimination may be due to the degeneration of a considerable number of the nerve fibers. Secondly, the distribution of the neuronal discharges across the population of nerve fibers is abnormal. The differences in phase between discharges of different fibers in the normal situation are absent in case of electrostimulation, when all fibers discharge at the same time. This may negatively affect central processing of the temporal information. Thirdly, decoding of temporal information may be restricted to time intervals corresponding to about the characteristic frequency of the unit. Thus, poor frequency discrimination found for electrostimulation with high frequencies may be due to absence of fibers innervating the high-frequency region and

low-frequency fibers responding to these high stimulus frequencies but lacking a high-frequency temporal decoder. Finally, the high temporal acuity found for electrostimulation of the eighth nerve in animal experiments is usually measured at higher levels than the stimulus levels used by implant patients. Temporal acuity should be studied as a function of stimulus level.

IN CONCLUSION

The chairman of this session wondered whether we should study the neurophysiological response to speech stimuli. At the peripheral level there is no specific response to speech stimuli. Therefore, the point was raised that it might be better to determine the responses to basic, analytic stimuli, and to use the results to formulate models which can then be used to predict the response to arbitrary stimuli such as speech sounds. However, the general feeling was that studying directly the neurophysiological response in animals to speech stimuli is highly relevant to gain insight into speech perception. Signal processing by the peripheral auditory system is very complex and non-linear. Therefore, we have to study the response to complex signals and we might as well use speech stimuli. These stimuli are well characterized and of enormous interest to everybody.

Chapter 6

PRIMARY SPEECH PERCEPTS

ENGLISH AND FRENCH SPEECH PROCESSING: SOME PSYCHOLINGUISTIC INVESTIGATIONS*

Jacques Mehler
Centre de Sciences Cognitives et
Psycholinguistique
C.N.R.S. & E.H.E.S.S., Paris,
France

Juan Segui
Laboratoire de Psychologie
Expérimentale
Université René Descartes,
E.P.H.E., U.A. 316 C.N.R.S.

Syllables are identified faster than features or phonemes, at any rate in initial position of the target-item (Savin & Bever, 1970). This demonstration has led many authors to suggest that the syllable is the basic segment in speech perception and that phonemes can only be derived from the analysis of the perceptually primary segment, namely, the syllable. Considerable disagreement remains as to how this observation ought to be interpreted. Savin & Bever considered, that the phoneme had linguistic rather than psychological reality. Their interpretation, however, came up against considerable criticism from several authors including McNeil & Lindig (1973), Healy & Cutting (1976), Foss & Swinney (1973).

Foss & Swinney argued that perceptual processes should be distinguished from the processes that are required to identify a target. Although smaller units may be identified by fractioning larger ones, this does not mean that larger units intervene earlier or singularly during perceptual processing.

Mehler (1981) reviews some of the methods and data leading to the rejection of Savin & Bever's conclusions and suggests that many of the initial findings had been replicated after revision for the various criticisms that had been made. For instance, lexical access is related to number of syllables (rather than to real time or number of phonetic segments). Indeed, the length of the word that precedes the target predicts latencies, provided length is measured in terms of number of syllables and not of number of phonemes (see Mehler, Segui & Carey 1978). Moreover, monosyllabic items are not processed like bi- or polysyllabic items. Monosyllabic items give rise to a lexical superiority effect (LSE), namely, Ss' latencies are shorter for detecting the initial phoneme in words rather than in similar legal pseudowords (Rubin, Turvey & van Gelder, 1976; Segui, 1984; Cutler, Mehler, Norris & Segui, in press). No lexical superiority effect has been found for polysyllabic items, (see Segui et al., 1981; Foss & Blank, 1980; etc.).

The results mentioned above have led investigators to evaluate the hypothesis that the syllable is the basic on-line unit of speech

*We would like to thank Peter Jusczyk for his help and comments on earlier versions of this chapter. The work described therein was carried out with the help of CNET (Convention 837BD28 00790 9245 LAA/TSS/CMC), CNRS (ATP "Aspects Cognitifs et Neurobiologiques du Langage) and the European Science Foundation (TW 86/17)

processing. The syllable, according to such speculation, is the basic unit of speech processing because it is used, both to access the lexicon and to analyze the signal into component segments and features. A fairly extensive review of this work has already been presented (Mehler, 1981, and Segui 1984). In this paper more data are presented that buttress the syllable's role in speech processing.

The view that the syllable is a basic unit in speech processing is compatible with the dual-code hypothesis as presented by Foss & Blank (1980) and Cutler & Norris (1979). The dual code hypothesis posits that Ss can process and recognize the acoustic signal on the basis of either the phonetic (or prelexical) or the phonological (i.e., postlexical) codes. The phonetic code relies on the acoustic-phonetic properties of the incoming signal. The phonological code contains information about the phonological structure of words as they are represented in the lexicon; this code becomes available only after the lexicon has been accessed. As it was initially presented, the dual code hypothesis suggests that the prelexical code is rarely used since Ss can rely, in general, on the postlexical code. However, Foss & Gernsbacher (1983) have shown that Ss tend to rely on the prelexical code when listening to speech.

Mehler (1981) and Segui (1984) come close to accepting the dual code hypothesis. However, the syllabic view does not rest on the dual code hypothesis nor on any serial processing model. Indeed, models like those presented by McClelland & Ellman (1986), Dell (1985) accommodate most of the results of speech processing investigations. Even if speech processing is driven by segments that contain more than single phoneme information, parallel distributed processing (PDP) models easily accommodate this fact.

We believe that syllables play a critical role in the perception of words and they believe that the syllable can be bypassed under very special circumstances in speech perception. Using the phoneme monitoring technique to infer speech perception suggests that some monitoring responses are triggered from the lexical representation rather than from the syllabic one. According to segment type, lexical influence may or may not determine latency. This does not require our accepting a race model. Ss may rely on alternative codes or they are perhaps prone to lexical influence when responding to monosyllables and less so when they have to react to polysyllabic items. Mehler and Segui acknowledge the claims made by Fodor (1983) that

the operations of the input systems appear to be, in this respect, inflexibly insensitive to the character of one's utilities. You can't hear speech as noise even if you would prefer to. **The modularity of Mind, p. 53,**

and by Marslen-Wilson & Tyler (1981) that

the analysis of the input will always be developed as far as it can be, and as rapidly as it can be. **Central Processes in Speech Understanding, p. 114**

Indeed, as is argued in the above quotations, processing of any stimulus is carried to the highest level possible. Thus, truly monosyllabic items result in lexical access whenever items have a lexical

representation. For example, the syllable dog results obligatorily in lexical access, but the syllable dag is only processed up to the level of its acoustic-phonetic representation. Syllables that are part of larger items are used for lexical access but also yield their phonetic-acoustic representation prior to the establishment of a lexical representation. Thus, the first syllable of the word driving results in the acoustic-phonetic representation of the syllable dry and it is an empirical issue whether the item dry in this instance is also lexically accessed. The syllabic model presented by Mehler & Segui postulates the existence of strong coarticulatory cues that insure the automatic selection of processing levels for every segment. Aside from the data by Dommergues, Segui & Mehler (in press), results reported by Foss & Swinney (1973) and by Mills (1980) support this claim. Ss seem to have a previously unsuspected sensitivity to aspects of the acoustic signal which deserve much more attention. Our view is close to Marslen-Wilson & Tyler's (1980) cohort model, although we view the elimination of candidates in terms of syllables rather than of phonemes. Tyler (1984) has recently reported data in favor of a similar view since she reports that Ss require more than the mere information contained in the first 150 ms, before they can establish the initial cohort.

In their attempt to evaluate the data in favor of and against the syllable as the basic unit in speech perception Mehler et al. (1981) showed that speech segments are detected faster in the initial position of an item when they correspond to the first syllable of the item than when the segment incorporates more or less information than that contained in the first syllable. For instance, the target /pa/ is detected faster in pa#lace than in pal#mier, while the target /pal/ is detected faster in pal#mier than in pa#lace. Mehler et al. speculated that syllabification might be the reflection of a processing universal, but more recent data induces us to be far more cautious. Below I present data that bear on the issue.

SCOPE OF SYLLABIFICATION ROUTINES

The results presented in the previous section suggest that in the course of speech processing Ss compute syllables. As a matter of fact, syllabification has been viewed as the most basic routine used in speech perception. However, many reasons can be invoked for cautiousness in so far as generalisation of this routine is concerned. To ascertain whether syllabification arises as a consequence of universal speech processing routines, Segui and myself carried out several experiments in collaboration with Cutler & Norris.

Cutler, Mehler, Norris & Segui, (1986) stress the fact that some languages have uniform syllabic structures while others tolerate more or less variation in structure. Also, the way syllabic boundaries are signalled is language-dependent. For instance, in some languages, e.g. English, some words, e.g. [pa/l/ace] have segments like /l/ which belong to two syllables at once. These segments are said to be ambisyllabic. In stressed languages, intervocalic consonants preceding an unstressed vowel are frequently ambisyllabic. French mostly contains words which have no ambisyllability, at least not between the first and the second syllable. The same is the case in Spanish and Italian. This lack of homogeneity across languages raises the possibility that syllabification may vary from language to language. In

fact, syllabification may be much less reliable for coping with languages that tolerate high rates of ambisyllabicity as compared with languages like French and Spanish that do not. To explore these issues, Cutler et al. investigated the way in which French and English speakers cope with speech when they have to detect a segment in a list of words or non-words, i.e., words in a language other than the Ss' native one.

In a first experiment they tested pairs of items that had the same initial CVC segment, but where one element of the pair incorporated the segment in a CVC structure while the other incorporated it into a CV[C] structure. English Ss were informed that they had to press a key as soon as they detected a sequence in a list of words. The results in Table 1 show that words with ambisyllabic consonants elicit faster monitoring responses than words with clear syllabic boundaries. This difference was significant, but there was absolutely no trace of syllabification in English-speaking Ss. Syllabification should yield faster latencies to targets that coincide with the target's first syllable. The interaction of CV versus CVC targets with the initial syllable structure of a word that Mehler et al. found with French materials, failed to be replicated with English materials. Could these results mean that Mehler et al.'s findings with French were due to a subtle acoustic difference between the different tokens in each pair? Probably not. Had this been the case, we would have expected that English Ss listening to the French tape would have provided similar results to those obtained by Mehler et al. However, in an experiment carried out to test this prediction it was found, see Table 2, that English Ss' responses were almost identical on the English and the French tapes. All in all, the English-speaking Ss show no sign of the syllabification effect observed with French-speaking Ss.

Table 1, Mean RT (ms) in Experiment 1 (English subjects, English words)

	CV[C] words (e.g. <u>balance</u>)	CVC words (e.g. <u>balcony</u>)	M
CV targets (e.g. <u>ba-</u>)	456	502	479
CVC targets (e.g. <u>bal-</u>)	448	514	481
M	452	508	

Table 2, Mean RT (ms) in Experiment 2 (English subjects, French words)

	CV words (e.g. <u>balance</u>)	CVC words (e.g. <u>balcony</u>)	M
CV targets (e.g. <u>ba-</u>)	431	471	451
CVC targets (e.g. <u>bal-</u>)	419	500	459
M	425	485	

Before drawing some general conclusions, we asked ourselves why the CVC words received faster responses than the CV[C] ones irrespective of target type, and more specifically whether this might be the result of a post-lexical effect. The most direct way to evaluate whether post-lexical factors influenced the Ss' response was to run a similar experiment using accepted pseudowords and English speaking Ss. The results obtained see (Table 3) are almost identical to those observed in the two previous experiments. Thus, we have to infer that whatever Ss were doing in the first experiment they continued doing the same in the pseudowords' experiment.

Table 3. Mean RT (ms) in Experiment 3 (English subjects, "English" nonwords)

	CV[C] nonwords (e.g. <u>balic</u>)	CVC nonwords (e.g. <u>balgart</u>)	
CV targets (e.g. <u>ba-</u>)	391	399	395
CVC targets (e.g. <u>bal-</u>)	370	393	381
	380	396	

In another experiment, French Ss responded to English words and showed that segmentation is language specific. The French Ss responded rapidly to CVC targets in CVC words and slowly to CV targets in the same words, see Table 4.

Table 4. Mean RT (ms) in Experiment 4 (French subjects, English words)

	CV[C] words (e.g. <u>balance</u>)	CVC words (e.g. <u>balcony</u>)	
CV targets (e.g. <u>ba-</u>)	448	467	458
CVC targets (e.g. <u>bal-</u>)	457	440	448
	453	454	

This result is predictable, since Ss are familiar with the words where syllabification might apply. However, when the French Ss were asked to respond to the CV[C] words that are incompatible with French phonology they identified the CV marginally faster than the CVC targets. The interaction was significant, although we found no significant effect with CV[C] words. Thus syllabification seems effective on material that affords proper acoustic clues. Interestingly, both kinds of items (items compatible with French phonology and ambisyllabic ones) on the average elicit equally rapid responses.

The results presented above show that in a series of experiments designed to asses their speech processing routines, French and English speakers behave differently. These two languages have different phonological and presumably different acoustic structures.

French is an oxytonic language with minimal ambisyllabicity, in particular, between the first and second syllables of polysyllabic words. In contrast, English incorporates a sizeable proportion of polysyllabic items that demonstrate ambisyllabicity between the first and second syllables. Furthermore, English, in contrast to French, has contrastive stress and no tonic accent. It therefore seems hardly surprising that British Ss show no evidence of syllabification, while French Ss always tend to analyze the signal in terms of syllabic segments. Closer analysis of our results suggests that French speakers rely upon the first syllable of polysyllabic items if acoustic cues unambiguously indicate the boundary. In the absence of such unambiguous cues to the syllabic boundary, French Ss revert to alternative structures. Notice however, that the absence of syllabic boundary cues does not result in increased latencies in detecting multiphonetic targets; Ss are equally fast with CVC and CV[C] words. If Ss always analyse the signal in terms of syllables their latencies should be longer when they are faced with ambiguous syllable boundaries. Since Ss are not slowed down by words with ambisyllabic first segments, we must assume that syllabification is one of a number of cues that can be employed during speech processing. Unfortunately, little can be said about the other cues that are active during the first perceptual strategies.

Before drawing some general conclusions on these investigations we would like to point out a problem that makes us uneasy about our results. In the study by Mehler et al. (1981), we found latencies of roughly 360 ms, all conditions included. These results were obtained with French Ss responding to French words. With English Ss responding to English words we found latencies of roughly 480 ms. Surprisingly, English Ss were slightly faster with French than with English words; their latency was, roughly, 455 ms. The average latency of English Ss responding to accepted pseudowords was 485 ms. This was the only condition for which the English Ss were almost as fast as the French. Lastly, let me point out that the French Ss responding to the English list had a mean latency of 450 ms. Needless to say, Ss in all groups were different, and it is dangerous to give too much importance to between-group differences. However, if the mental chronometry paradigm is taken seriously, it becomes quite difficult to justify accepting a 20 ms difference as significant while ignoring, or even worse, not trying to understand, a difference in latency of roughly 100 ms or more.*

One possible hypothesis about the described latencies is that English Ss also compute an intermediary segment like the first syllable of a word. However, since accessing the lexicon by means of first syllables in languages that use contrastive stress is not very informative, Ss may have learned to wait until they discover whether the first syllable is or is not stressed. To obtain that information we would need to compare the amplitudes of the first and second vocalic nucleus, where more stimulus information is needed to effect a response. The additional time may mask any evidence of syllabification due to a ceiling effect.

*We know that statisticians will consider this point ludicrous. Tant pis.

The results presented above suggest that syllabification may be language-specific. Syllabification seems to be employed mainly by native speakers of languages with a regular syllabic structure. Of course, it seems unfair to assert that speakers of other languages do not syllabify, but it must be said that if they do, this behavior is not nearly as obvious. Finally, let me add that Bradley, Sanchez-Casas, and Garcia Albea (personal communication) have found that in languages like Spanish syllabification intervenes during speech perception. This replication with a language other than French that incorporates clear boundaries between the first two syllables of multisyllabic words is important for the position I have espoused above. However, much more research on languages with varying phonetic structures will be necessary before such a position can be acceptable.

In the above review we have assumed that phoneme and syllable monitoring times provide data from which inferences about bottom-up speech comprehension processes can be drawn. Many investigators question such an assumption and bring forward excellent data to buttress their claim. In the section below I will try to argue that it is possible to cope with all the arguments and accommodate all the data that critics have mustered against phoneme detection tasks and finally land with a rather coherent account of lexical access and speech perception.

LEXICALITY EFFECTS IN ON-LINE TASKS

Lexical superiority effects (LSE), that is, the more efficient detection of a sublexical segment in a word rather than in a legal pseudoword, have been reported by experimenters using phoneme or syllable monitoring tasks, (Foss & Blank, 1980; Rubin, Turvey & van Gelder, 1976; Segui et al., 1981). LSE is important since it allows further exploration of the impact of lexical parameters on the perception of speech. Furthermore, given that the arguments supporting a syllabification strategy in French rely upon detection latencies, it seems critical to show that the results reviewed above cannot be viewed as entirely due to lexical effects. Recently, Foss & Gernsbacher (1983) presented data supporting a unitary model of phoneme identification which suggests that LSE are always the result of an artifact. If so, then many of the results used to support syllabification in French must be misconstruals due to artifacts. Before accepting such a major reinterpretation of data, however, it would seem worthwhile to review the evidence once again.

Rubin, Turvey & van Gelder (1976) presented Ss with a sequence of monosyllabic words and pseudowords. The Ss' task was to press a button when they heard /b/ or /s/. Rubin et al. reported much shorter latencies for phonemes in words than in non-words and suggested that phoneme detection generally reflects postlexical responding. However, Foss & Blank (1980) found that when lists of sentences were used the lexical status of the target bearing item did not affect latency. There were many methodological differences between the experiments of Rubin et al. and Foss & Blank. However, Segui, Frauenfelder & Mehler (1981) suggested that the contrasting results can be explained by the differences in the target materials used by these experimenters. Rubin et al. used only monosyllables, whereas Foss et al. used mostly polysyllabic items. If, as I asserted above, the syllable is the smallest bottom-up segment that can be

extracted from the signal, then when the target is a mono-syllabic word, responses will be triggered from the post-lexical code. This is so because the smallest segment available, namely the syllable, coincides with a word. Another way of stating the same facts is to say that when the target is a monosyllabic word, lexical effects will influence the Ss' response through operation of excitatory links in the network. In contrast, when the syllable is a pseudo-word, responses will originate in the phonetic code, according to the dual code hypothesis, or alternatively, the lexical influences will be minor according to interactive models. At any rate, since Rubin et al. used only monosyllabic words and monosyllabic distractors, LSE was only to be expected if the syllable is indeed a basic segment in speech processing.

LSE will not arise with polysyllabic words and pseudowords since their first syllables are the elementary segments used to access the lexicon. These same segments activate response buffers before lexical recognition, which allows us to talk about a primitive syllabic code. Presumably, experiments that use multisyllabic items fail to show LSE although such-an effect appears in experiments that use monosyllabic items. Assuming, as we did earlier, that processing is mandatory to the highest level of structure, the reported pattern of results becomes intelligible. Monosyllabic words access the lexicon on the basis of automatic and mandatory processing. The way in which monosyllabic pseudowords are represented and detected is compatible with many alternative views. For instance, when a S listens to a monosyllabic pseudoword he/she computes that representation more slowly than the representation of a very similar word because of 1) differences in frequency between the two type items, 2) intrinsic differences in the time to elaborate the two types of code, namely, the lexical and the syllabic code, and 3) differences with which the facilitatory effect of word nodes are felt.

The above interpretation was corroborated by results of an experiment designed to evaluate LSE. Segui et al. asked Ss to detect the first phoneme or the first syllable in bisyllabic words and pseudowords. They found that word initial phonemes or syllables are identified prior to lexical access (or before lexical facilitation can percolate down through the network). The response to the first syllables was the same, regardless of the lexical status of the target bearing item. This finding was expected since, by our hypothesis, the initial syllable of words and pseudo-words informs the response buffer. Ss extract the first syllable of multisyllabic items and latencies are the same irrespective of the lexical status of the target-bearing item. The results of Segui et al.'s experiment are presented in Table 5.

Table 5.

	WORD	NON-WORDS
PHONEMES	347	346
SYLLABLES	285	281

As can be seen from Table 5, both phoneme and syllable monitoring latencies are the same for words and pseudowords. Furthermore, the model presented above also correctly predicted that

monitoring latency would not be affected by the lexical status of the first syllable of a multisyllabic target-bearing item. This finding is important since, presumably, PDP- models predict that the lexical status of the first syllable in multisyllabic items affects the response latency. Of course, given the present results, one could still argue that the spread of activation is slow and that the response can be effected before activation occurs.

There is another potential explanation for the lack of a lexical superiority effect in the Segui et al. data. Namely, Mills (1980) reported that syllables receive different responses when they appear as monosyllabic items, e.g. can or in a bisyllabic item, e.g. candle. In order to ascertain the validity of the claim, Segui (1984) further explored this issue using mono- and bisyllabic items in an experiment in which half of the items were words and the other half pseudowords. In order to understand the experiment, consider the pair of items **CRI** and **CRA**. The former is a word while the latter is a pseudoword, and both appear as the first syllable in bisyllabic items, e.g., **CRItere** and **CRAtere**. These items in conjunction with several other quadruples were presented to Ss in lists containing other mono- and bisyllabic items. Ss were asked to phoneme or to syllable monitor for specific targets. The results presented in Table 6 show that an LSE can be observed for monosyllables, whereas bisyllabic items result in similar latencies irrespective of their lexical status.

Table 6.

	WORDS	NON-WORDS
MONOSYLLABIC	319	351
BISYLLABIC	357	356

These results confirm the ones reported both by Segui et al. and by Rubin et al. Moreover, they corroborate the suggestion made as to the difference in processing of mono- and bisyllabic items. Furthermore, the results suggest that although we might orthographically represent **CRI** in cri and in critere in the same way, we do hear it as different in these contexts. Hence, as Mills (1980) claimed, monosyllabic items are processed differently when they are the first syllable of a larger unit, thus suggesting the existence of acoustic correlates to syllable coarticulation.

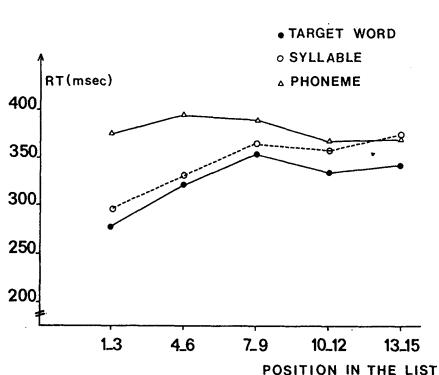
ALTERNATIVES TO LEXICAL SUPERIORITY EFFECTS

In the preceding section we presented data that suggest that monosyllabic items result in a "bona fide" lexical superiority effect. However, before accepting this view, we must again evaluate the claims made by Foss & Gernsbacher (1983) suggesting that all LSE are artifactual. Morton & Long (1976) claimed that phoneme monitoring (and "a fortiori" syllable monitoring) tap postlexical processes, since phoneme monitoring latencies are sensitive to the transition probability of the target bearing items. However, as Mehler (1981) has noted, Morton & Long used mostly mono- syllabic items making it impossible to tease apart the processing of monosyllabic acoustic targets from

their lexical representation (since responses are made after lexical access has taken place or after the effect of lexical nodes on the response buffer has occurred). Of course, this is not necessarily true for polysyllabic items. Indeed, we used bi-syllabic target bearing items in three kinds of transition probability (TP) contexts. In one context the target-bearing word was highly predictable, in another it had a very low TP, while in the third it was intermediate in TP. In all three contexts the target-bearing item had the same initial phoneme while only the first two items had the same initial syllable, e.g., Bibelot, Biberon and Bonbon. Our results are presented in Table 7 and in the figure.

Table 7. Phoneme Monitoring in Sentence Contexts and Lists

	HS (directeur)	LS (dirigeant)	LPh (délégué)
IN CONTEXT	322	347	387
OUT OF CONTEXT	395	406	390



The TP of the target bearing item did not determine the phoneme or syllable monitoring latencies. Indeed, items with identical initial syllables led to similar latencies irrespective of their TPs. In contrast, items sharing the same initial phoneme but different syllable segments had longer latencies than items which shared the same initial syllable. Table 7 makes it possible to compare the latencies for sentence and list presentation of the same items. As can be seen,

the items with identical initial syllables yield faster latencies in sentences than in lists. No difference was observed for the items that share only the initial phoneme. The figure also illustrates another interesting aspect of the data, namely, the evolution of latencies over the experiment. At first, latencies for targets whose initial syllables are identical to the initial syllable of a high TP item are responded to very rapidly. The same is not true of targets that share only the first phonetic segment with high TP items. However, by the end of the experiment, the overall reaction times are slower. This increase in overall reaction time is apparently due to slower responses to targets on trials in which the initial syllable is shared with a high TP item. At that point, and only at that point, does TP seem to play a role. This suggests a switch from a purely bottom-up strategy to one based on lexical responding, as the experiment comes to a close. Thus, our results show that phoneme and syllable monitoring need not tap postlexical processes, contrary to what Morton & Long suggested.

Instead, our results suggest that the first syllable is part of the bottom-up information used to access the lexicon.

Let us now turn our attention to Foss & Gernsbacher's (1983) claims in favor of a unitary theory of speech processing. According to Foss & Gernsbacher, monitoring latencies are mainly determined by pre-lexical factors; the lexical code plays a minor role, perhaps even a negligible role in processing. Indeed, these authors report that vowel duration, among other acoustical properties of the syllable, is almost entirely responsible for the phoneme monitoring latencies observed across all experiments. In an ingenious control, Foss & Gernsbacher show that Morton & Long's results are also observed when the target bearing items are presented in lists rather than in sentences, making the notion of TP no longer relevant:

Phoneme targets beginning words that had been presented by Morton & Long in high predictability contexts were responded to much more rapidly (508 milliseconds) than phoneme targets beginning words that had been presented in low predictability contexts (560 milliseconds), a mean difference in reaction time of 52 milliseconds.

These latency differences, according to Foss & Gernsbacher, arise because the "predictable" items have shorter vowels and fewer initial consonants than the "unpredictable" target items. Given this observation, any experimenter claiming an LSE must reckon with Foss & Gernsbacher's criticism before he/she deserves to be taken seriously. In order to assess if differences in latencies in our own experiment might be attributable to such factors, we explored Ss' responses to target bearing items when they were presented out of sentence context, namely, in lists. We noted that the facilitation effect for initial syllables disappears, see Table 7. Thus, the effects we noted in sentence contexts cannot be attributed to differences in vowel length. This result suggests that the LSE can be observed, at least when experiments are carried out in French.

COMPARING LSE IN ENGLISH AND FRENCH

Given the importance of the issue reviewed in the section above, Cutler, Norris, Segui & Mehler (submitted for publication) decided to examine LSE in greater detail. They first explored the behavior of Ss listening to monosyllabic experimental items of identical vowel length. Half of the experimental monosyllabic items were words and the others pseudo-words; fillers were bisyllabic items. Ss were asked to phoneme monitor for a stop consonant in item-initial position. The items were presented in lists. Five pairs of words and pseudowords were CV, 10 pairs were CVC, and another 10 pairs were CCV in structure. The results presented in Table 8 show a significant lexicality effect for each one of the structures and for all the items lumped together. Irrespective of syllable structure, latencies are shorter for words than for similar pseudowords. This LSE was obtained despite the fact that vowel length was stringently controlled.

Table 8, Mean RT (ms) per condition for Experiment 1

	Words	Non-words	
CV	437	463	450
CVC	439	483	461
CCV	451	492	472
	444	483	

Other acoustic factors may, nonetheless, partially determine latencies. Since this experiment was run with French items and French Ss it was easy to assess whether the LSE observed was a legitimate one or not. Indeed, Cutler et al. tested English Ss using the French tape. The English Ss were not French speakers and therefore could not be expected to show an LSE on this French language tape. Interestingly, we only replicated the French results for syllabic structure (CV was fastest and CCV slowest), but found absolutely no trace of an LSE. This result reinforces our belief that the French displayed a true LSE. When Ss know the words, an LSE can be observed, but this is not the case when Ss are unaware of the difference between the words and pseudowords because they do not speak the language. This finding is discrepant from that of Foss & Gernsbacher. However, before we explain the difference between their results and ours, there are more data to be considered.

Cutler et al. also replicated Foss & Gernsbacher, using only CVC targets corresponding to English words or acceptable pseudowords. We found, like Foss & Gernsbacher, that there was no LSE. However, when the French experiment was replicated in all its details (with the exception of the stimuli which were English words or pseudowords) an LSE once again occurred. At this point, we conjectured that the contrasting nature of results obtained with English and French might be due to the nature of the fillers used in each case. Foss & Gernsbacher used mostly monosyllabic fillers, whereas in the French experiments fillers were mostly bisyllabic. To ascertain whether this might have been the decisive factor, Cutler et al. ran a control removing all the bisyllabic fillers from the French tape and replaced them with monosyllabic ones. The LSE vanished. Hence, it appears that Foss & Gernsbacher cancelled the LSE by relying on a methodology that inhibited lexical influences on the response. Such inhibition may arise due to the monotony of long homogeneous lists of monosyllabic items presented at regular rates. Ss tend to perceive such lists as tedious and meaningless. However, whenever homogeneity disappears, or regularity is disrupted, Ss tend to display an LSE, at least for monosyllabic items.

In support of the above view, Segui et al. found that when Ss have to evaluate the proportion of words vs nonwords in lists where either all items are monosyllabic or where some items are monosyllabic and others bisyllabic, Ss tend to overestimate the proportion of pseudowords in mono-syllabic lists.

In conclusion, the syllable appears to be a natural segment computed during mandatory speech processing routines. Furthermore, the consideration of the nature of syllabification provides a unified account of data thought to be attributable to postlexical access (e.g. Rubin et al., Morton & Long) or different kinds of acoustical artifacts (Foss & Gernsbacher). Indeed, we have been able to show that post-lexical effects are observed for monosyllabic items and that purely phonetic-code effects are functional with polysyllabic items.

REFERENCES

1. Cutler, A. and Norris, D. (1979). Monitoring sentence comprehension. In: W.E. Cooper and E.C.T. Walker, (Eds.) (Sentence Processing: Psycholinguistic Studies Presented to Merrill Garret). LEA. N.J.
2. Cutler, A., Mehler, J., Norris, D., and Segui, J.A. (1983). Language-specific comprehension strategy. Nature, 304, 159-160.
3. Cutler, A., Mehler, J., Norris, D., and Segui, J.A. (1986). The syllable's differing role in segmentation of French and English. Journal of Memory and Language, 25.
4. Cutler, A., Mehler, J., Norris, D., and Segui, J.A. (1986). Phoneme identification and the lexicon. Cognitive Psychology, in press.
5. Dell, G.S. (1985). Positive feedback in hierarchical connectionist models: Application to language production. Cognitive Psychology, 1985, 3-23.
6. Dommergues, J.Y., Segui, J., and Mehler, J. (1986). Context effects in lexical access. Submitted to Journal of Language & Memory.
7. Fodor, J.A. (1983). The Modularity of Mind. MIT Press, Cambridge.
8. Foss, D.J. and Blank, M.A. (1980). Identifying the speech codes. Cognitive Psychology, 12, 1-31.
9. Foss, D.J. and Swinney, D. (1973). On the psychological reality of the phoneme. Perception, identification and consciousness. Journal of Verbal Learning and Verbal Behavior, 12, 246-257.
10. Foss, D.J. and Gernsbacher, M.A. (1983). Cracking the dual code: Towards a unitary model of phoneme identification. Journal of Verbal Learning and Verbal Behavior, 22, 609-632.
11. Healy, A. and Cutting, J.E. (1976). Units of speech perception: Phoneme and syllable. Journal of Verbal Learning and Verbal Behavior, 15, 73-83.
12. Marlsen-Wilson, W. and Tyler, L.K. (1981). Central processes in speech understanding. Philosophical Transactions of the Royal Society, B295, 317-332.
13. McNeil, D. and Lindig, K. (1973). The perceptual reality of phonemes, syllables, words, and sentences. Journal of Verbal Learning and Verbal Behavior, 12, 419-430.
14. McClelland, J.L. and Elman J.L. (1986). The TRACE model of speech perception. Cognitive Psychology, 18, 1-86.
15. Mehler, J. (1981). The role of syllables in speech processing: infant and adult data. Philosophical Transactions of the Royal Society, B295, 333-352.
16. Mehler, J., Segui, J., and Carey, P. (1978). Tails of words: monitoring ambiguity. Journal of Verbal Learning and Verbal Behavior, 17, 29-35.

17. Mehler, J., Dommergues, J.Y., Frauenfelder, U., and Segui, J. (1981). The syllable's role in speech segmentation. Journal of Verbal Learning and Verbal Behavior, 20, 298-305.
18. Mills, C.A. (1980). Effect of the match between listener expectancies and coarticulatory cues on the perception of speech. Journal of Experimental Psychology, E.E.P. 6(3), 528-535.
19. Morton, J. and Long, J. (1976). Effect of word transitional probability on phoneme identification. Journal of Verbal Learning and Verbal Behavior, 15, 43-51.
20. Rubin, P., Turvey, M.T., and van Gelder, P. (1976). Initial phonemes are detected faster in spoken words than in spoken nonwords. Perception and Psychophysics, 19(5), 394-398.
21. Savin, H.B. and Bever, T.G. (1970). The nonperceptual reality of the phoneme. Journal of Verbal Learning and Verbal Behavior, 9, 295-302.
22. Segui, J. (1984). The syllable: A basic perceptual unit in speech processing? In: H. Bouma and D.G. Bouwhuis (Eds), Attention and Performance X. Lawrence Erlbaum Ass., Hillsdale, N.J.
23. Segui, J., Frauenfelder, U., and Mehler, J. (1981). Phoneme monitoring, syllable monitoring, and lexical access. British Journal of Psychology, 72, 471-477.
24. Segui, J., Dommergues, J.Y., Frauenfelder, U. and Mehler, J. (1982). The perceptual integration of sentences: syntactic and semantic aspects. In: J.F. Le Ny and W. Kintsch (Eds.), Language and Comprehension, North Holland, Amsterdam.
25. Tyler, L.K. (1984). The structure of the initial cohort: Evidence from gating. Perception and Psychophysics, 36, 417-427.

UNITS OF ORGANIZATION AND ANALYSIS IN THE PERCEPTION OF SPEECH*

Robert E. Remez

Department of Psychology, Barnard College, 3009 Broadway,
New York, New York 10027, U.S.A.

The message uttered by a talker and the message the perceiver hears possess a common linguistic structure--phones, syllables, words, phrases, and sentences. However, the acoustic connection between the talker and listener is more easily described nonlinguistically, as an assortment of acoustic components. Research on the perceptual links in the speech chain has revealed the listener's impressive accomplishment in recognizing the linguistic properties of such acoustic patterns, inasmuch as linguistic units are not conveyed by an isomorphic set of acoustic units. The distribution of acoustic ingredients in a speech signal is immediately revealing on this point: The junctures between the syllables, words, and even phrases that the perceiver hears do not correspond to the acoustic junctures, which are more numerous and more varied than their linguistic counterparts (Fant, 1962). In consequence, the paradigm problem for research in speech perception is to explain how the listener accomplishes the reduction of acoustic variation to linguistic significance.

To reformulate the theorist's burden in accordance with the psychophysical study of information processing, we may ask: What are the units of perceptual organization and analysis for speech signals? Organization in perception establishes the integrity of coherent sound sources, and in speech perception it is commonly described as a "cocktail party" problem (Cherry, 1953). In circumstances when the signals of simultaneous talkers combine to compose the stimulation, by virtue of perceptual organization the perceiver may follow just the components relevant to the particular talker of interest, segregating that stream of elements from an acoustic background originating from other sound sources. The more familiar function of analysis in perception involves the identification of attributes of the utterance, and has naturally been viewed as a matter of determining the sequence of linguistic elements. Perceptual analysis presupposes perceptual organization, or, to put it more baldly, the analysis of the signal depends on the availability of a coherent whole, ready for analysis.

The focus of this paper is perceptual organization, in specific the limiting case of a single speech signal without an interfering acoustic background. Superficially, this case seems to present no problem at all for explaining perceptual organization, especially when organization is seen as the separation of signal from background. But, even a single speech signal comprises diverse acoustic elements which

*This work was supported by grant NS-22096 from the National Institute of Neurological and Communicative Disorders and Stroke.

despite their variety fuse into a single perceptual stream. It is this fusion of dissimilar elements that presents the problem. While the cocktail party problem offers little to illuminate this aspect of organization, Hockett's (1955) well-worn Easter egg analogy captures the issues well.

"Imagine a row of Easter eggs carried along a moving belt; the eggs are of various sizes, and variously colored, but not boiled. At a certain point, the belt carries the row of eggs between two rollers of a wringer, which quite effectively smash them and rub them more or less into each other. The flow of eggs before the wringer represents the series of impulses from the phoneme source; the mess that emerges from the wringer represents the output of the speech transmitter. At a subsequent point, we have an inspector whose task it is to examine the passing mess and decide, on the basis of the broken and unbroken yolks, the variously spread-out albumen, and the variously colored bits of shell, the nature of the flow of eggs which previously arrived at the wringer. Notice that he does not have to try to put the eggs together again--a manifest physical impossibility--but only to identify." (Page 210.)

From the perspective of Hockett's analogy, the organizational problem of the single signal may be simply stated: Why do the diverse bits fuse? It seems no less likely for the perceiver to form one stream of yolk, another of white, a third of shell--or one of clicks, another of buzzes, one of murmurs, another of hisses, and another of hums--especially since the physical attributes of the signal are so much more apparent than the underlying phonetic significance. The limiting case of the single signal presents a problem for explaining organization after all.

In venturing an answer to this question about organization, it is natural to begin with the evidence for perceptual units and perceptual organization that has already been proposed in other settings. This way we may evaluate the applicability of existing conceptualizations of units for the present venture. My review will amount to a recommendation to supplement the usual sensory and linguistic approaches to perceptual organization and analysis with a perceptual approach. Then, I will propose that evidence from studies of sinusoidal replicas of speech is consistent with the argument, and appears to suggest a reformulation of the question that may improve it.

It will simplify and abbreviate my exposition to present the premises and conclusions here, at the outset. 1) Studies of speech perception implicate minute acoustic cues in the recognition of linguistic properties; (2) Other studies undermine this characterization, indicating that the perceptual value of a cue is contingent on the local and remote environment in which it occurs; (3) Therefore, the designation of an acoustic element as a cue appears to depend on the signal properties that compose the context. (4) Perceptual organization is an aspect of perception that establishes the coherence of the signal, a precondition for analysis; (5) The unit of organization necessarily incorporates cue-size units of analysis; (6) Therefore, properties of perceptual organization determine the acoustic loci of phonetic information; and, (7) Units of perceptual analysis are nested within units of perceptual organization.

PERCEPTUAL ANALYSIS

A Bit of History. The identification of a minute unit of speech perception is a goal with a distant past. Titchener's generation of psychophysicists offered an empirical defense of the claim that the starting point for phonetic perception was a set of primitive sensory elements that were given, and irreducible. The primary sense quality of vocality named the phonetic dimension in the experience of pure tones, or what were taken to be pure tones (Boring, 1942; Köhler, 1910; Modell & Rich, 1915; Titchener, cited in Boring). In other words, the sensory consequences of simple tones included impressions of pitch, loudness, timbre, and phonetic attributes. Insofar as a complex wave is the sum of simple waves, then impressions of a complex waveform reduce to compound impressions. In the case of speech these impressions were aggregations of the vocalities of the simple components.

We can calibrate those antique siren-stimuli today, and perhaps explain the reported impressions of vocality by noting the similarity between the rather complex spectra of such sounds and the spectra of vocal productions. However we view the early attempt to explain phonetic perception by primary sense qualities, we should recognize the theoretical posture in contemporary attitudes. For example, the phonetic identification of chirps, the isolated frequency transitions of single formants (Nusbaum, 1983); or identification of noise bursts typical of stop consonant release (Cole & Scott, 1972) are intended to supply evidence about the perceptual effects of momentary snippets of the acoustic speech signal. Such studies preserve the perspective that speech perception is essentially a kind of accretion of elementary phonetic impressions. This is one way we understand speech cues: each cue is a kind of speech sound. In this respect, the unit of perception collapses into a hypothetical unit of auditory analysis, and the explanation of perception becomes equivalent to a catalog of phonetic impressions of acoustic details. The elegant simplicity of this view has great appeal.

The Acoustic Cue. The sensory explanation of speech perception; then, has been based on this modification of the notion of the acoustic alphabet, here recast as an inventory of cues. The perceptual effects of particular acoustic cues are studied in many labs, and, ironically, this research has raised many doubts about framing the story of perception in terms of sensory elements. No single acoustic element has proven necessary for perceiving any particular consonant or vowel, and many acoustic elements appear to bear information simultaneously about several different phones. Moreover, listeners tolerate well the lack of unique correspondences between signal elements and phonemes, and seem easily able to recognize speech when the signal is degraded by noise, by clipping, or by the effects of whispering, shouting or rapid articulation. The emphasis on particular cues is undercut by the flexibility of the listener and the absence of perceptual difficulty with novel signals.

In the acoustic mosaic conceptualization of the speech signal, the perceptual system must determine the significance of each ingredient. It is doubtful, though, that piecemeal evaluation of speech cues really occurs, as much research shows (Bailey, Summerfield & Dorman, 1977; Best, Morrongiello & Robson, 1981; Grunke & Pisoni,

1982; Liberman, Isenberg & Rakert, 1981; Liberman, Harris, Kinney & Lane, 1961; Mattingly, Liberman, Syrdal & Halwes, 1971; Pisoni, 1971; Popper, 1972; Rand, 1974; Samuel, 1981; Shattuck & Klatt, 1976). The general finding of these studies is that sensitivity to the acoustic details of a signal is rather poor unless the particular element in question has been acoustically isolated. With an intact speech signal, the fine grain of the stimulus may elude the listener, who instead perceives the mechanical properties of the sound source (Schubert, 1974). [This is actually to say no more than to translate the commonplace distinction between the proximal stimulus and the distal object to the speech case. Although the stimulus at the ear is the proximal cause of perception, the object of perception is an impression of a distal source for the pattern in the perceptual medium (Liberman & Mattingly, 1985).] Given a signal that promotes speech perception, its acoustic properties seem less well resolved than its phonetic properties, even in the subcategorical case (Whalen, 1984).

For an example of the relative inaccessibility of acoustic elements, consider Mattingly et al. (1971), who tested discrimination of formant frequency transition cues in the context of an intact syllable pattern, and compared this to discrimination performance for the identical acoustic elements removed from syllable context. Listeners performed poorly in discriminating elements within acoustic patterns that promoted phonetic identification, and relatively better when the elements to be judged were presented without the accompanying acoustic pattern of the syllable. In a variant of this approach, Samuel (1981) showed that the difference between a noiseband and a mixture of noise and formants was not easily detected if the two different stimuli were presented embedded in words, which were otherwise intact. When a discrimination test was performed on the critical pieces of noise or consonant+noise mixture, isolated from the lexical and acoustic context, listeners had less difficulty with the discrimination. In general, perceivers are less able to judge properties of acoustic elements in speech signals when phonetic categorization occurs. This suggests that during normal perception with ordinary speech signals, the listener is not assessing each element of stimulation, since under conditions more favorable for identifying elemental attributes the ability to do so is demonstrably poor. This is not to deny that specially trained subjects differentiate subphonemic variation, (for example, Carney, Widin & Viemeister, 1977), nor that subphonemic variation is available ordinarily (Massaro, in press). But it does suggest that explicit identification of acoustic elements ought to be considerably better than it is if phonetic perception ordinarily depends on that ability.

Phone, syllable, word, phrase. Linguistically motivated approaches to the perceptual unit have differed from the auditory perspective, and find support in a large variety of evidence. In essence, the view alleges that perceptual processes treat the speech signal as if it were composed of linguistic structures. The phonetic segments, consonants and vowels, are the most obvious and popular choices for processing units on these grounds, and the studies of the limits on recognition and discrimination that are rationalized by segmental properties are well known (Liberman, Harris, Hoffman & Griffith, 1958). Listeners perceive graded acoustic variation phonetically--more or less--despite the effects of sensory memory or the presumed natural nonlinearities in auditory resolution (Aslin, Pisoni

& Jusczyk, 1984; Miyawaki, Strange, Verbrugge, Liberman, Jenkins, & Fujimura, 1975; Pisoni, 1971, 1975). Other facts about speech errors, diachrony, reading, and tests of recognition and similarity have been mobilized to support the case for the segment as unit of perception (Fowler, 1986). However, the interpretation of these findings as evidence of a segmental unit of perception has its detractors, and the syllable has often figured as the next preferred unit (Bondarko, 1969; Ladefoged, 1967; Savin & Bever, 1970; Segui, 1984; Studdert-Kennedy, 1980).

Because information for consonants and vowels is seldom found in brief acoustic packets, and is often distributed throughout a syllable (Liberman, 1970), the unitary conception of the syllable has held great attraction. Widening the temporal size of the unit, though, also admits other possibilities. First, syllables are not independent, and the evaluation of the internal acoustic structure of one may be affected by the properties of its neighbors (for example, Summerfield, 1981). Given the auditory fragility of the acoustic cues, it appears unlikely that such effects are cognitive, latterly occurring, or subject to informed guessing. More likely, the effects are perceptual, that is, pertaining to the conversion of the acoustic signal to a linguistic percept.

A second effect of considering larger portions of the stimulus within a single perceptual unit has been noted in experiments by Ganong (1980), Rubin, Turvey & van Gelder (1975), and Remez, Rubin, Katz & Dodelson (1985), among others. These studies concluded that perceptual sensitivity to acoustic signals can be modulated by the lexical status of the utterance. There is apparently a perceptual difference between stimuli that form words and those that do not. The aspect of perception responsible for the effect is obscure. Again, the mischief of these studies is the suggestion that small linguistic units are contingent perceptually on larger linguistic units, the phone contingent on the word, for instance. Evidence of this kind contradicts the ordinary notion of relations among levels of analysis, and is discouraging for element-based hierarchical accretion accounts of phoneme perception.

Adopting the logogen as processing unit hardly solves the problem, though. There is old and durable evidence that a phrasegen may be used for the perception of words if not sentences. (Bever has mentioned a clausagen.) This is a third effect of entertaining larger linguistic constituents as perceptual units. It takes the phrase as an important unit of integration of stimulus information, based on studies that showed perceptual contingencies on syntactic and semantic coherence across many words (Fletcher, 1929; Hunnicut, 1985; Lashley, 1951; Lieberman, 1963; Miller, Heise & Lichten, 1951) and, possibly, sentences (Mehler & Carey, 1967). To summarize Miller et al., subjects do not transcribe test words presented in lists in noise as well as the same words presented in sentences in noise. In other words, despite the presumably constant contribution of fine grained auditory processes, perceptual effects varied with the coarse grained pattern of the stimulation. The usefulness of acoustic cues proved greatest when they corresponded to phones that made words within well formed phrases. A pessimist, viewing these findings, might conclude that nothing less than a grammar, available on-line, would suffice as the motor of perception.

To review the problem of the unit: If we approach the definition from the perspective of auditory sensory resolution, we find that the phonetic value of an acoustic element varies hugely, and cannot be determined by considering elements singly and independently, as should reasonably be true of processing units. Also, the set of elements contributing to perception appears to be open rather than closed. An approach to recognition based on acoustic elements can tolerate even vast inventories of distinctions as long as they are finite in number and unique in correspondence. Neither of these properties seems to be true of phonetic perception. This point will be crucial in evaluating the instance of sinewave replicas in the discussion of organization. If we turn, instead, to an approach that defines processing units linguistically, we find again that the units do not seem monadic, but are leaky. Moreover, this approach reifies linguistic attributes as corpuscles of processing, rather than considering linguistic descriptions the end products of perception, plain and simple. [This seems to be a potential instance of the fallacy called the experience error by Köhler (1947).] In sum, there is evidence for and against both the sensory and linguistic conceptualizations of the analytic unit. It seems that the attempt to identify the unit of perceptual organization from candidates drawn from the proposed units of analysis is frustrated by the dilemmas of perceptual analysis of speech. A direct exposition of perceptual organization, applied specifically to speech signals, is warranted.

PERCEPTUAL ORGANIZATION

Auditory Continuity. In their discussion of speech perception, Bailey & Summerfield (1980) suggest that because the potential set of acoustic elements responsible for phonetic impressions is open, it seems unlikely that an account of perception based on collections of acoustic cues will be adequate. If there is no limit to the number of acoustic elements, then the patterns composed by the elements must be more crucial to perception than the specific acoustic properties of the elements themselves. There are, in fact, a number of portraits of speech perception that echo Bailey & Summerfield (Kewley-Port & Luce, 1984; Liberman & Cooper, 1972; Remez, Rubin, Pisoni & Carrell, 1981; Walley & Carrell, 1983). Given the facts of acoustic variation, the listener's inability to extract fine acoustic details from intact speech signals, and the questionable value of identifying elements anyway, it seems plausible that the perceiver integrates elements before perceptual analysis occurs. This kind of integration is the function of perceptual organization, and it is probably keyed to phonetically appropriate patterns of modulation of the signal rather than to its elemental details. Put simply, the role of organization is to ensure a coherent stimulus pattern for the mechanisms of recognition.

Perceptual organization considered as an auditory process leads us to expect an appropriate characterization in the guidelines discovered in general, nonspeech cases (for example, Bregman, 1981; Bregman & Campbell, 1971; Dannenbring, 1976; Julesz & Hirsh, 1972; van Noorden, 1971; Vicario, 1982). This research extends the principles of the visually based Gestalt laws of form, and shows that groups of elements form along a variety of auditory dimensions. For example, temporal groups form through the device of rhythmic repetition of brief elements, including pulsing in synchrony; spectrally

based groups form when tones fall within a close frequency range of one another; perceptual coherence over time is established for spectra that change smoothly and continuously; subjects may even ignore brief interruptions in smoothly changing stimuli by the principle of closure. Do such principles of grouping suffice in the case of the speech signal? Organization by principles such as "interpolate small gaps," or "elements close in frequency belong together," or "smooth continuation defines a stream" would lead the listener to hear a single speech signal as multiple streams, three for the asynchronously changing formants, a fourth stream for the fundamental, a train of bursts over several consonantal releases, and many intermittent pulses: noise pulses when frication or aspiration occurred, murmurs when nasalization occurred. (Figure 1 shows spectral representations of natural utterances that should split into streams by these principles, but which cohere, in fact.) A speech signal organized in this way would differ from the case of ordinary speech perception, though perhaps it would not be too different from perceptual disorganization observed in certain aphasic patients (Albert & Bear, 1974; Goldstein, 1974) or perhaps in neonates (Jusczyk, 1985). Of course, the ordinary listener hears a steady stream of linguistic stuff despite the acoustic discontinuities.

These auditory laws of form are cast in the terms of the proximal stimulus, though the motivation is explicitly the veridical perception of the distal object. As Bregman & Pinker (1978) state: "If two sets of acoustic features have arisen from a single source, they should be fused in perception (i.e., experienced as features of a single stream of sounds); if they arise from two separate sources, they should participate in two separate fusions" (page 20). This describes the objectives of the laws of organization, and such formulae based on physical continuity, spectral similarity and metrical repetition may actually distinguish sound sources when they are instruments of the orchestra; and they may explain the musical device of implied polyphony. However, such characterizations do not describe a source for speech sounds, which is a compound viscoelastic tube. The auditory system and the theorist alike must exploit different grouping principles for phonetic distal objects.

Support for my claim comes from Lackner & Goldstein (1974), who used the paradigm of rapid repetitive presentation to induce fission of speech sounds. They observed a dissociation of a single nonsense speech signal into two streams, one apparently consisting of syllable nuclei, the other of syllable onsets, an outcome consistent with the rules described by Bregman. But, they noted that the stream formation was atypical of ordinary speech perception, in which the listener does not experience such demisyllabic disintegration, or at least fails to notice it if it occurs. Most likely, the use of the repetition paradigm as a test of perceptual organization is inappropriate for studying speech, for it appears to impose the dominance of similarity, cooccurrence, smooth continuity and the like as organizational principles, overriding those that may apply to the ordinary speech signal. It need hardly be remarked that the perceptual conditions that more typically confront the speech perceiver could not be more dissimilar from presentations in this test paradigm, since physically identical, metrical repetition of brief acoustic patterns--to guess recklessly--must be exceedingly rare in utterances.

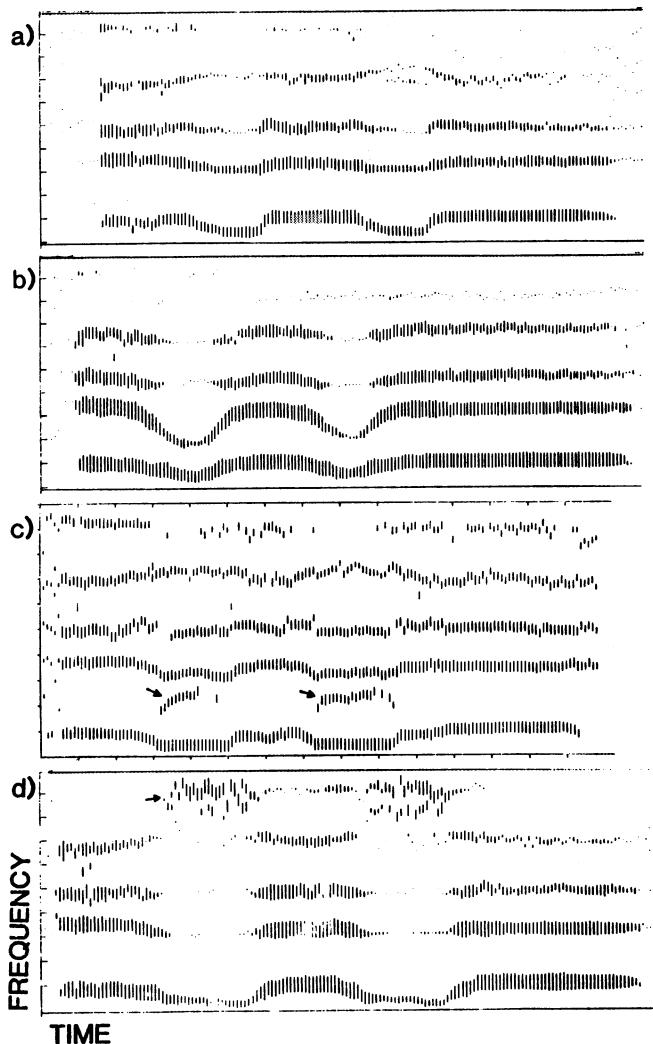


FIGURE 1. These four panels illustrate departures from the simple physical coherence in variations within natural speech signals. (A) Nonsense utterance /elələ/, in which the first formant changes in frequency, though little else changes in the spectrum. (B) Nonsense utterance /swəwəʃ/, exhibiting disproportionate frequency changes in the first and second formants, and negligible change in higher formants. (C) Nonsense utterance /ɛnɛnɛ/ with discontinuous nasal resonances (arrow). (D) Nonsense utterance /ɛzɛzɛ/, in which the first formant alone is continuous, and fricative (aperiodic) excitation may be noted (arrow).

A distinction between the organization of speech and nonspeech can be put succinctly. Speech perception starts with a rapid sequence of acoustic components, as also happens in the general auditory case to which the laws of form or the rules of streaming presumably apply.

However, phonetic perception employs auditory units larger than the individual acoustic elements, and the signals do not generally exhibit rapid repetition of brief elements. These factors block the formation of perceptual streams along simple criteria, allowing the exercise of more complex susceptibilities to the stimulation. Some recent studies in our laboratory implicate these perceptual functions. They point to a perceptual sensitivity to time-varying properties of the signal spectrum, and encourage the claim that this sensitivity is responsible for the organization of the speech signal.

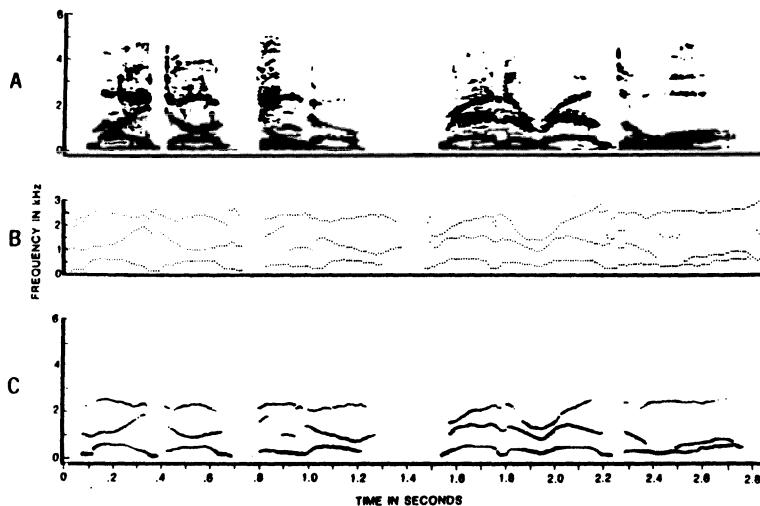


FIGURE 2, Illustration of sinewave replication of a naturally produced utterance. (A) Spectrographic representation of natural speech. (B) Extracted formant tracks at 10 ms intervals. (C) Spectrographic representation of three-tone replica. The sentence is, "My dog Bingo ran around the wall." (From Remez & Rubin, 1983)

Time-varying coherence. Our research on sinewave replicas of speech signals (Remez, Rubin, Nygaard & Howell, in press; Remez et al., 1981; Remez & Rubin, 1983, 1984) has aimed at elaborating the perceptual organization and analysis of speech by normal adult perceivers, specifically testing the reliability of coarse-grained patterns of spectrotemporal variation for providing phonetic information. With the technique of sinewave replication of natural signals, a few sinusoids are used to represent the frequency and amplitude variation of natural signal elements, for example, of the formants, fundamental frequency, bursts, fricative resonances, or aspiration. Such signals lack harmonically related components, broadband resonances and the pulsing structure of natural speech. (See Figure 2.) Our results show that listeners understand the replicated utterance by attending to time-critical properties of the patterned variation of the sinusoidal signal. In other words, speech perception can effectively proceed from the variation of nonspeech tonal elements that preserve the natural pattern of speech signal changes. Although the short-time spectral properties of natural speech are undeniably important in normal speech perception, not least because they are responsible for the impression

of natural vocal quality, they are not a necessary condition of speech perception. The sinusoidal technique separates the perceptual effects due to specific sensory attributes of signal elements from the perceptual effects of the configurations of elements composing the signal.

The perceptibility of sinewave sentences implicates the listener's sensitivity to patterns of signal variation in speech. Most perceivers tolerate this discrepancy of unnatural signal elements undergoing natural signal variation because the tonal complexes exhibit abstractly vocal attributes. Of course, these studies also show rather straightforwardly that attributes of the auditory dimension of timbre are completely different from phonetic attributes. Linguistic properties, including consonant and vowel features, are preserved over the transformation from natural to sinusoidal signals, though the voice quality of sinusoidal signals is unmistakably unnatural.

In speech perception with natural acoustic elements, we may assume that a key component of perceptual organization is the listener's specific susceptibility to the modulation of speech signals, whatever momentary spectral considerations also apply. Essentially, the principles of speech signal variation supplement those appropriate for the general auditory cases to guide perceptual organization. Some of these have been described by Stevens (1972) and Stevens & Blumstein (1981): The alternating rise and fall of the amplitude envelope; the presence of spectral peaks and valleys which change in frequency as the amplitude envelope changes; and, of course, nonuniform spectrum changes brought about over time by quantal acoustic effects of relatively independent articulators. My acoustic-phonetics tutorial here suggests specific additions to this list of attributes characteristic of speech sounds: The abrupt introduction of periodic or aperiodic signal elements, perhaps accompanying the restructuring of the long-term spectrum, respectively brought about by coupling the nasopharynx to the oropharynx or by narrowing the vocal tract to create turbulence. Such principles rationalize the correspondence between the sound source and the sound, however imprecise the current formulation. They therefore may approximate the perceptual principles that detect continuity in the source even when the speech signal changes in ways that are alternatingly discrete, or asynchronous across the elements, or gradual and uniform.

CONCLUSION

Sinewave replication studies hold specific consequences for determining the units of speech perception. Consider the principal finding: the acoustic output of time-varying oscillators, linear emitters atypical of vocally produced sounds, provides linguistic information when the pattern exhibits the long-term properties of a vocal source. This means that phonetic perception incorporates susceptibility to signal dynamics that supersedes the effects of acoustic details. When we are able to describe this susceptibility precisely, we will have a means to identify the information in acoustic signals independent of the barrage of specific elementary acoustic attributes. Moreover, since the principles of coherent vocal spectrum variation determine which acoustic elements participate in the process of linguistic perception, these studies offer the prospect of discovering the units of

organization and analysis rather than assuming them. Because perceptual analysis presupposes an organized signal, the determination of the principles of perceptual organization is required for a complete explanation of the familiar questions about invariance and segmentation in speech perception. The terms of the investigation should change only slightly, however. In the empirical challenge to perceptual accounts that I have sketched, we continue to need a model of the auditory and perceptual functions that register signal variation, to be sure. We will also profit from a model of signal production in time-critical dimensions, in order to assess the information about linguistic properties available to the listener and the means by which it is detected.

REFERENCES

1. Albert, M.L., and Bear, D. (1974). Time to understand: A case study of word deafness with reference to the role of time in auditory comprehension. Brain, 97, 393-394.
2. Aslin, R.N., Pisoni, D.B., and Jusczyk, P.W. (1984). Auditory development and speech perception in infancy. In M.M. Haith and J.J. Campos (Eds.), Infancy and the biology of development. New York: Wiley.
3. Bailey, P.J., and Summerfield, Q. (1980). Information in speech: Observations on the perception of [s]-stop clusters. Journal of Experimental Psychology: Human Perception and Performance, 6, 536-563.
4. Bailey, P.J., Summerfield, A.Q., and Dorman, M. (1977). On the identification of sine-wave analogues of certain speech sounds. Haskins Laboratories Status Report on Speech Research, SR-51/52, 1-25.
5. Best, C.T., Morrongiello, B., and Robson, R. (1981). Perceptual equivalence of acoustic cues in speech and nonspeech perception. Perception and Psychophysics, 29, 191-211.
6. Bondarko, L.V. (1969). The syllable structure of speech and distinctive features of phonemes. Phonetica, 20, 1-40.
7. Boring, E.G. (1942). Sensation and perception in the history of experimental psychology. New York: Appleton-Century.
8. Bregman, A.S. (1981) Asking the "what for" question in auditory perception. In M. Kubovy and J.R. Pomerantz (Eds.), Perceptual organization. Hillsdale, New Jersey: Lawrence Erlbaum, Pp. 99-118.
9. Bregman, A.S., and Campbell, J. (1971). Primary auditory stream segregation and perception of order in rapid sequence of tones. Journal of Experimental Psychology, 89, 244-249.
10. Bregman, A.S., and Pinker, S. (1978). Auditory streaming and the building of timbre. Canadian Journal of Psychology, 32, 19-31.
11. Carney, A.E., Widin, G.E., and Viemeister, N.F. (1977). Noncategorical perception of stop consonants differing in VOT. Journal of the Acoustical Society of America, 62, 961-970.
12. Cherry, C. (1953). Some experiments on the recognition of speech with one and two ears. Journal of the Acoustical Society of America, 25, 975-979.
13. Cole, R.A., and Scott, B. (1972). Toward a theory of speech perception. Psychological Review, 81, 348-374.
14. Dannenbring, G.L. (1976). Perceived auditory continuity with alternately rising and falling frequency transitions. Canadian Journal of Psychology, 30, 99-114.

15. Fant, C.G.M. (1962). Descriptive analysis of the acoustic aspects of speech. Logos, 5, 3-17.
16. Fletcher, H. (1929). Speech and hearing. New York: Van Nostrand.
17. Fowler, C.A. (1986). An event approach to the study of speech perception from a direct-realist perspective. Journal of Phonetics, 14, 3-28.
18. Ganong, W.F. (1980). Phonetic categorization in auditory word perception. Journal of Experimental Psychology: Human Perception and Performance, 6, 110-125.
19. Goldstein, M.V. (1974). Auditory agnosia for speech (pure word deafness); a historical review with current implications. Brain and Language, 1, 195-204.
20. Grunke, M.E., and Pisoni, D.B. (1982). Some experiments on perceptual learning of mirror-image acoustic patterns. Perception & Psychophysics, 31, 210-218.
21. Hockett, C.F. (1955). A manual of phonology. Baltimore: Waverly Press.
22. Hunnicut, S. (1985). Intelligibility versus redundancy--conditions of dependency. Language and Speech, 28, 47-56.
23. Julesz, B., and Hirsh, I.J. (1972). Visual and auditory perception: An essay of comparison. In E.E. David and P.B. Denes (Eds.), Human communication: A unified view. New York: McGraw-Hill, Pp. 283-340.
24. Jusczyk, P.W. (1985). On characterizing the development of speech perception. In J. Mehler and R. Fox (Eds.), Neonate cognition: Beyond the blooming, buzzing confusion. Hillsdale, New Jersey: Lawrence Erlbaum.
25. Kewley-Port, D., and Luce, P.A. (1984). Time-varying features of initial stop consonants in auditory running spectra: A first report. Perception & Psychophysics, 35, 353-360.
26. Köhler, W. (1910). Akustische Untersuchungen, II. Zeitschrift für Psychologie mit Zeitschrift für Angewandte Psychologie und Charakter-kunde, 58, 59-140.
27. Köhler, W. (1947). Gestalt psychology. New York: Liveright.
28. Lackner, J.R., and Goldstein, L.M. (1974). Primary auditory stream segregation of repeated consonant-vowel sequences. Journal of the Acoustical Society of America, 56, 1651-1652.
29. Ladefoged, P. (1967). Three areas of experimental phonetics. London: Oxford University Press.
30. Lashley, K.S. (1951). The problem of serial order in behavior. In L.A. Jeffress (Ed.), Cerebral mechanisms in behavior. New York: Wiley, Pp. 112-136.
31. Liberman, A. M. (1970). The grammars of speech and language. Cognitive Psychology, 1, 301-323.
32. Liberman, A.M., and Cooper, F.S. (1972). In search of the acoustic cues. In A. Valdman (Ed.), Papers in linguistics and phonetics to the memory of Pierre Delattre. The Hague: Mouton, Pp. 329-338.
33. Liberman, A.M., Harris, K.S., Hoffman, H.S., and Griffith, B.C. (1957). The discrimination of speech sounds within and across phoneme boundaries. Journal of Experimental Psychology, 54, 358-368.
34. Liberman, A.M., Harris, K.S., Kinney, J.A., and Lane, H. (1961). The discrimination of relative onset time of the components of certain speech and nonspeech patterns. Journal of Experimental Psychology, 61, 379-388.

35. Liberman, A.M., Isenberg, D., and Rakert, B. (1981). Duplex perception of cues for stop consonants: Evidence for a phonetic mode. Perception and Psychophysics, 30, 133-143.
36. Liberman, A.M., and Mattingly, I.G. (1985). The motor theory of speech perception revised. Cognition, 21, 1-36.
37. Lieberman, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech. Language and Speech, 6, 172-187.
38. Massaro, D.W. Categorical partition: A fuzzy logical model of categorization behavior. In S. Harnad (Ed.), Categorical perception. New York: Cambridge University Press. In press.
39. Mattingly, I.G., Liberman, A.M., Syrdal, A.K., and Halwes, T.G. (1971). Discrimination in speech and nonspeech modes. Cognitive Psychology, 22, 131-157.
40. Mehler, J., and Carey, P. (1967). Role of surface and base structure in the perception of sentences. Journal of Verbal Learning and Verbal Behavior, 6, 335-338.
41. Miller, G.A., Heise, G.A., and Lichten, W. (1951). The intelligibility of speech as a function of the context of the text materials. Journal of Experimental Psychology, 41, 329-335.
42. Miyawaki, K., Strange, W., Verbrugge, R.R., Liberman, A.M., Jenkins, J.J., and Fujimura, O. (1975) An effect of linguistic experience: The discrimination of +rij and +lij by native speakers of Japanese and English. Perception & Psychophysics, 18, 331-340.
43. Modell, J.D., and Rich, G.J. (1915). A preliminary study of vowel qualities. American Journal of Psychology, 26, 453-456.
44. Nusbaum, H.C. (1983). Possible mechanisms of duplex perception: "chirp" identification versus dichotic fusion. Perception & Psychophysics, 35, 94-101.
45. Pisoni, D.B. (1971). On the nature of categorical perception of speech sounds. Supplement to Haskins Laboratories Status Report on Speech Research, SR-27.
46. Pisoni, D.B. (1975). Auditory short-term memory and vowel perception. Memory & Cognition, 3, 7-18.
47. Popper, R.D. (1972). Pair discrimination for a continuum of synthetic voiced stops with and without first and third formants. Journal of Psycholinguistic Research, 11, 205-219.
48. Rand, T.C. (1974). Dichotic release from masking for speech. Journal of the Acoustical Society of America, 55, 678-680.
49. Remez, R.E., and Rubin, P.E. (1983). The stream of speech. Scandinavian Journal of Psychology, 24, 63-66.
50. Remez, R.E., and Rubin, P.E. (1984). On the perception of intonation from sinusoidal sentences. Perception & Psychophysics, 35, 429-440.
51. Remez, R.E., Rubin, P.E., Katz, M., and Dodelson, S. (1985). On the influence of lexical status in phonetic perception. Paper presented at the 26th Annual meeting of the Psychonomic Society, Boston, Massachusetts.
52. Remez, R.E., Rubin, P.E., Nygaard, L.C., and Howell, W.A.. Perceptual normalization of vowels produced by sinusoidal voices. Journal of Experimental Psychology: Human Perception and Performance, in press.
53. Remez, R.E., Rubin, P.E., Pisoni, D.B., and Carrell, T.D. (1981). Speech perception without traditional speech cues. Science, 212, 947-950.

54. Rubin, P.E., Turvey, M.T., and van Gelder, P. (1975). Initial phonemes are detected faster in spoken words than in spoken nonwords. Perception and Psychophysics, 19, 394-398.
55. Samuel, A.G. (1981). The role of bottom-up confirmation in the phonemic restoration illusion. Journal of Experimental Psychology: Human Perception and Performance, 11, 1124-1131.
56. Savin, H., and Bever, T.G. (1970). The nonperceptual reality of the phoneme. Journal of Verbal Learning and Verbal Behavior, 33, 295- 302.
57. Schubert, E.D. (1974). The role of auditory perception in language processing. In D.D. Duane and M.B. Rawson (Eds.), Reading, perception and language. Baltimore: York, Pp. 97-130.
58. Segui, J. (1984). The syllable: A basic perceptual unit in speech processing? In H. Bouma and D.G. Bouwhuis (Eds.), Attention and Performance X: Control of Language Processes. Hillsdale, New Jersey: Lawrence Erlbaum, Pp. 165-181.
59. Shattuck, S.R., and Klatt, D.H. (1976). The perceptual similarity of mirror-image acoustic patterns. Perception & Psychophysics, 20, 470- 474.
60. Stevens, K.N. (1972). The quantal nature of speech: Evidence from articulatory-acoustic data. In E.E. David and P.B. Denes (Eds.), Human communication: A unified view. New York: McGraw-Hill. Pp.51- 66.
61. Stevens, K.N., and Blumstein, S.E. (1981). The search for invariant acoustic correlates of phonetic features. In P.D. Eimas and J.L. Miller (Eds.), Perspectives in the study of speech. Hillsdale, New Jersey: Lawrence Erlbaum, Pp. 1-38.
62. Studdert-Kennedy, M. (1980). Speech perception. Language and Speech, 23, 45-66.
63. Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. Journal of Experimental Psychology: Human Perception and Performance, 77, 1074-1095.
64. Vicario, G.B. (1982). Some observations in the auditory field. In J. Beck (Ed.), Organization and representation in perception. Hillsdale, New Jersey: Lawrence Erlbaum, Pp. 269-283.
65. von Noorden, L.P.A.S. (1971). Rhythmic fission as a function of tone rate. IPO Annual Progress Report, No. 6. Eindhoven, The Netherlands.
66. Walley, A.C., and Carrell, T.D. (1983). Onset spectra and formant transitions in the adult's and child's perception of place of articulation in stop consonants. Journal of the Acoustical Society of America, 73, 1011-1022.
67. Whalen, D.H. (1984). Subcategorical mismatches slow phonetic identification. Perception & Psychophysics, 35, 49-64.

IMPLICATIONS FROM INFANT SPEECH STUDIES ON THE UNIT OF PERCEPTION*

Peter W. Jusczyk
Department of Psychology, University of Oregon,
Eugene, Oregon 97403, USA

One of the great overlooked tasks of language acquisition is how the language learner develops a representation of the sound properties of a word that allows for the recognition of words in fluent speech. Thus, while child phonologists have given a great deal of attention to the nature of the representations that underlie the child's earliest productions of speech (e.g. Ferguson, 1986; Macken, 1980; Menn, 1980; Vihman, 1978), and other child language specialists have focused on the structure of the child's semantic categories, much less is known about what sort of representation of the sound structure of words permits their comprehension by the child (cf. Jusczyk, 1985). Yet, one could argue that learning the essential phonemic characteristics that distinguish between one name and another in one's native language is at least as important as distinguishing the boundaries of potential referents for the names. Clearly, to be a fluent speaker-hearer of a language, one needs to develop the appropriate categories for both the sounds and the meanings of words.

What little information is available about the nature of underlying phonetic representation during the early stages of language acquisition comes from two sources: infant speech perception experiments and word learning studies with children during the second year of life. At first glance, the information from these two sources appears to be contradictory. The studies of infant speech perception (see Aslin, Pisoni & Jusczyk, 1983 for a review) indicate an organism with very fine discriminative capacities, capable of distinguishing contrasts that might occur in any language. In contrast, the studies of early word learning (e.g. Shvachkin, 1973; Garnica, 1973) suggest that the process of acquiring phonemic contrasts occurs a step at a time over the course of many months. Although this discrepancy might simply be attributable to the sensitivity of the tasks employed, a closer inspection suggests that the tasks are tapping different kinds of capacities. A typical infant speech perception experiment, employing a procedure like high amplitude sucking (HAS), is a same-different discrimination task. In order to be credited with discriminating some contrast (even between items that differ by a single phonetic feature) the infant only needs to note that the preshift stimulus differs in some way from the postshift stimulus. The infant need not register the

*The research reported here was supported by a research grant to the author N.I.C.H.D. (#HD-15795). The author also wishes to thank Ranka Bijeljac-Babic, Josiane Bertoni, Patricia Kuhl, Jacques Mehler, Joanne Miller, Deborah Kemler Nelson, Robert Remez and Janet Werker for comments they made on a previous version of this manuscript.

way in which the stimulus pair differ, only that they are not identical. For this purpose, a holistic comparison of the overall similarity of the stimuli may be sufficient for discrimination.

A word-learning type task presents the child with a different type of situation. In such a task, the child is often presented with a number of novel objects, each of which has an unfamiliar name. The child must learn which name goes with which object. Here the child cannot succeed by simply making holistic same-different judgments since the contrast is not simply between a pair of items but among many different ones that differ along a number of dimensions. In order to succeed on such a task, the child must develop some sort of representation of the sound properties of the name that allow it to be picked out and identified from all the other possible names that might be uttered in that situation. Presumably the representation which the child must develop for this purpose is related in some way to the underlying speech perception capacities that allow infants to distinguish speech sounds from one another.¹

In what ways might the infant's speech perception capacities be related to the kinds of representations used to recognize words in fluent speech? At the very least, the basic speech perception capacities provide some sort of psychophysical lower-bound on the kinds of representations that one could form concerning the sound structure of words in a language. But this need not imply that the representations themselves are packed with such fine-grained information about all the potential phonetic distinctions that could arise between words. In fact, given the goal of arriving at a maximally efficient representation of the sound structure of words in a language--one that allows them to be rapidly and accurately identified during fluent speech--it seems likely the ideal representation would be one with only enough detail to distinguish between possible words in a language. This suggests an additional source of constraint upon the lexical representations developed by the language learner, viz. the phonological structure of the target language itself. Knowledge of which sound categories from the inventory of those occurring in a particular language mark out meaningful distinctions, as well as the permissible sequences of these categories (i.e. what phonotactic constraints exist) is useful in developing the most efficient representations of the acoustic characteristics of lexical items. In particular, information about which strings of phonemes constitute legal sequences for forming words in a language provides important information that could be employed in segmenting continuous speech into discrete words. Of course, information about the prosodic patterns, phonemic categories (and the allophones that map into each category) and phonotactic constraints of a particular language is present in the input that the language learner receives. Consequently, such regularities observed in the nature of the input during the course of language acquisition may well influence the way in which speech sounds are represented by infants. This being the case, the infant would be expected to move from a general representation of utterances to a language-specific one.

Thus, in order to understand how the infant becomes a recognizer of fluent speech we need to know something about (1) their underlying sensory capacities, (2) the kinds of perceptual representations they employ, and (3) the way in which knowledge of

the sound structure of a particular language affects the perceptual representations. At present, we have learned quite a bit about the first of these, are beginning to learn about the second, and know almost nothing at all about the third. In part, the reason we know so little about the third is because understanding how language-specific representations develop depends on the second. In other words, what is required is an index of the initial forms these perceptual representations take, i.e. prior to a great deal of input from a particular language. Hence, it is necessary to determine the nature of the young infant's representation of speech sounds. How specific is the information that an infant might store about the sound structure of a particular word in order to be able to recognize it on a subsequent occasion? What is the organization of the information that goes into such representations? What relationship exists between segmental and prosodic information in these early representations? Are segmental and prosodic features differentiated from each other in the representations? If so, how? Are prosodic and segmental features tied together in a tier-like arrangement of the sort described in autosegmental phonology (Goldsmith, 1976) or is there some other type of organization employed?

In order to begin to answer such questions, we have initiated a series of studies aimed at describing the nature of the young infant's representation of speech sounds. This research was conducted in both France and the United States and in collaboration with Jacques Mehler, Josiane Bertoncini, Ranka Bijeljac-Babic, and Lori Kennedy. We began by asking whether the early representations reflect some sort of phonemic description of speech information. In other words, is speech information analyzed and are representations organized in terms of a string of phonemic segments? For example, given a string of syllables like [bi], [si], [li], [mi], do infants detect the presence of a common vowel segment and do their representations of these utterances reflect this commonality? Our investigations focused on two groups of infants: newborns and 2-month olds. Previous studies of infant speech perception have not uncovered any detectable change in discriminative capacity during this age period. In fact, what little evidence has been reported regarding changes in discriminative capacities with linguistic experience has been noted only for infants 9 months of age or older (Werker & Tees 1983). However, given our assumption that linguistic input has its primary impact on the representations employed rather than the underlying perceptual capacities themselves, we thought it useful to probe the representations even during this early period of linguistic exposure.

The young age and limited response repertoire of our subjects limit the kinds of measures that can be employed to investigate the nature of the infant's representation of speech. Our approach was to modify the HAS procedure, which has been used so successfully to examine speech discrimination in infants at this age. The usual procedure extracts a same-different judgment from the infant by rewarding sucking with the repeated presentation of a single sound to listen to until a decline in sucking occurs, and then introducing a new sound as the reward. Discrimination of the pair of sounds is indexed by a significant increase in sucking to the new sound (for a more complete description see Jusczyk, 1985). We modified this procedure in a way that would force the infant to employ some sort of representation in dealing with the speech sounds. In particular, instead

of a single syllable, a set of different syllables presented in random order was used to reward sucking. Following the expected decline in sucking with the repeated presentation of the original set members, a new element was introduced into the set. Because the syllables differed from each other in various ways, the only way in which infants might recognize a new item added to this set would be to encode enough information about each member of the set to distinguish it from the others. In effect, the infants must form representations of the set members in order to succeed on the task.

TABLE 1

MODIFIED HIGH AMPLITUDE SUCKING PROCEDURE

PRESHIFT STIMULUS SET

bi, ba, bo, bɔ (presented in random, rather than fixed, order)

POSTSHIFT STIMULUS SET

Preshift set + a new syllable of one of the following types:

- 1) New instance from a familiar category, (e.g. bu) making the postshift set: bi, ba, bo, bɔ, bu (randomized).
- 2) New instance from a novel category, (e.g. du) making the postshift set: bi, ba, bo, bɔ, du (randomized).
- 3) No new instance, Instead the frequency of one of the preshift set members is increased (e.g. ba) making the postshift set: bi, ba, bo, ba, bɔ, ba (randomized). This is the control group.

To investigate whether infants encode information about individual segments, we varied the characteristics of the stimulus set (see Table 1). The items in the original stimulus sets were chosen so as to share a particular phonetic segment, either a consonant or a vowel. The new syllable added to this set (following the decline in sucking) either shared or did not share the phonetic segment with the original set members. Judging from previous investigations of visual categorization studies with young infants (e.g. Cohen & Strauss, 1979; Bornstein et al., 1976), our expectation was that the addition of an item which shared a phonetic segment with the original set members (i.e. a new instance from a familiar category) would be perceived by the infants as less novel than one which lacked the common phonetic segment (i.e. a new instance from a novel category). The greater novelty of the new instance from the new category was expected to manifest itself either by a greater increase in sucking or in a longer time to show a significant decline in sucking after the addition of the new token.

Our first study examined whether infants display any tendency to represent a stop consonant like [b] when it recurs in a series of different syllables. In other words, do infants perceive different syllables containing the same consonant segment as inherently more similar than ones containing a different segment? To enhance the possibility that the infants might focus on the common phonetic identity of the segments rather than on some common acoustic feature,

we chose the syllables in our stimulus set so as to maximize differences in their formant structure. Thus, the syllables [bi], [bo], [b], and [ba] were selected for the original stimulus set. Infants in all the test conditions heard this set of syllables during the initial (or preshift) phase of the experiment. The presentation sequence of the syllables was completely randomized, the only restriction being that each syllable occurred equally often. At each of the two age levels, there were four test groups (see Table 2). For one of the groups, the new token added to the original set was a new instance from the familiar category - [bu].² For the second group, the item added was a new instance from a novel category - [du]. Because the new token added in the second group included a new vowel as well as a new consonant, it was necessary to include a group for which only new consonantal information was added. Thus, for the third group, the consonantal information for one of the stimuli was changed from [ba] during the preshift phase to [da] during the postshift phase. Finally, the fourth group served as a control condition; the only change that occurred during the postshift period was an increase in the frequency of one of the original set members, [ba].

TABLE 2
DESIGN OF EXPERIMENT 1

PRESHIFT SET	POSTSHIFT SET	
bi, b̄, bo, ba	bi, b̄, bo, ba, bu*	(familiar C)
bi, b̄, bo, ba	bi, b̄, bo, ba, du*	(novel C and V)
bi, b̄, bo, ba	bi, b̄, bo, da*	(novel C only)
bi, b̄, bo, ba	bi, b̄, bo, ba*	(Control group)

*Note that in each case the frequency of occurrence of the new token is equated to the frequency of all the other tokens combined.

The data from this study were examined to see whether there was any indication that the new instances from the novel consonant category led to great increases in postshift sucking or in a longer time to re-habituate to the new stimulus set. The data concerning the postshift sucking increases are presented in Figure 1. Consider first the results for the 2-month-olds. All three test groups showed significantly greater increases in sucking than did the control group. More importantly, there was no indication that infants responded at a lower rate to a new instance from the familiar category (if anything, the results were in the opposite direction). Thus, there is no evidence from this measure that these infants are extracting a common consonant from these syllables. Similarly, an analysis of the time to re-habituate to the postshift test sets indicated no significant differences among the test groups.

The data from the newborns are similar in that there was no indication that the infants responded differentially to the presence of a common [b] segment. Specifically, there was no difference in either postshift sucking or time to re-habituate between the groups in which a new token from a familiar [bu] or novel category [du] was added to the original stimulus set. Both groups showed a significant increase in

sucking relative to the control group during the postshift period. However, there was one important difference in the results for the newborns relative to the 2-month-olds. The former gave no evidence of discriminating the change involving a single consonant (i.e. the substitution of [da] for [ba] in the postshift stimulus set). There are at least two possible explanations for this result. First, newborns may lack the necessary perceptual capacity to discriminate stop consonant contrasts. However, in view of previous results demonstrating that newborns are capable of discriminating such contrasts when the standard HAS procedure is employed (e.g. Mehler, 1985), a more likely explanation seems to lie in the nature of the representations that the infants formed. Specifically, the newborn's representation may be undifferentiated with respect to stop consonants. Some support for this hypothesis comes from follow-up studies which we have been conducting.

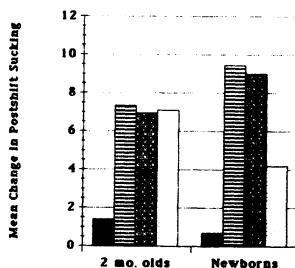


FIGURE 1, Displays the mean change in postshift sucking for the 2-month old and newborn subjects in each of the test conditions for the consonant categorization task. The scores are determined by subtracting the average sucking rates from the last two preshift minutes from the average of the first two postshift minutes.

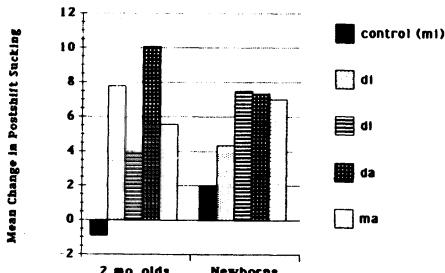


FIGURE 2, Displays the mean change in postshift sucking for the 2-month old and newborn subjects in each of the test conditions for the vowel categorization task. The scores are determined by subtracting the average sucking rates from the last two preshift minutes from the average of the first two postshift minutes.

Although our first study gave no indication that infants' early representations of speech are structured in terms of phonetic segments, it focused only on stop consonant information. Given that stop consonants are known to be subject to considerable influence from the surrounding context (e.g. Liberman, Cooper, Shankweiler & Studdert-Kennedy, 1967) and that they have relatively short durations as compared to other types of phonetic segments, infants might be slower to develop representations for stop consonants than for vowels which tend to be both more stable and of longer duration. To investigate this possibility, we conducted a second experiment in which we presented a stimulus set during the preshift phase consisting of syllables sharing a common vowel ([bi], [si], [li], [mi]). At each age level, there were five different postshift test groups (see Table 3). For

the first group, the new item added to the original stimulus set was a new instance from the familiar vowel category, [di]. The new item added for the second and third groups involved a new instance from a new vowel category. In the case of the second group, the new item [dl] was selected from a vowel category that is very close to the original category in a psychological similarity space, whereas the new item for the third group ([da]) was chosen from a category perceived to be very different from the original vowel category (Shepherd, 1972). The purpose of this manipulation was to determine whether the degree of acoustic similarity of the vowel categories had any effect on their representation by the infants. For the fourth group of infants, the new item added shared a consonant with one of the original set members ([ma]) and thus, differed only in vowel quality. The fifth group was a control in which only the frequency of one of the original set members ([mi]) was increased.

TABLE 3
DESIGN OF EXPERIMENT 2

PRESHIFT SET	POSTSHIFT SET	
bi, si, li, mi	bi, si, li, mi, di*	(familiar V)
bi, si, li, mi	bi, si, li, mi, dl*	(new close V)
bi, si, li, mi	bi, si, li, ma*	(new V only)
bi, si, li, mi	bi, si, li, mi*	(Control group)

*Note that in each case the frequency of occurrence of the new token is equated to the frequency of all the other tokens combined.

The data for the second study were analyzed as per the first one. Again the time to rehabilitation measure proved to be uninformative as there were no differences between any of the test groups at either age level. The data concerning the postshift sucking increases are presented in Figure 2. The results from the 2-month-olds parallel the results of the first experiment. Each of the four test groups showed significant increases in sucking relative to the control group. Again there was no evidence that the infants responded more strongly to a new instance from a new vowel category ([da] or [dl]) than they did to a new instance from the familiar category ([di]). Moreover, there was no evidence that proximity in adult psychological space affected the way in which the 2-month olds represented the vowels. Hence, there was no evidence that these infants were representing common vowel segments in the syllables.

At first glance, the data from the newborn subjects do appear to offer support for the view that infants structure their representations in terms of common vocalic segments. In particular, only the groups who received new instances from new vowel categories showed a significant increase in sucking relative to the control group during the postshift period. The group that heard a new instance ([di]) from the familiar category did not exhibit such an increase. However, further analysis of the data shows that the performance of these subjects was not reliably better than that of those groups who heard a new instance from a new vowel category. Given this result and the

lack of any evidence for a vowel segment representation in the data of the 2-month-olds, the explanation of the newborn data may lie elsewhere than in a tendency on their part to extract common vowels from the syllables. One alternative is suggested from the results of the first experiment with newborns. Recall that the infants failed to show a reliable increase in postshift sucking when only consonantal information was changed from the preshift set. We noted that one possible explanation for this result was that stop consonant information is undifferentiated in these infants' representations. A similar account can be advanced for the second experiment. If the infant has only a global sort of representation of information for stop consonant segments, then the addition of an item like [di] would not be differentiated from an item like [bi] in the original stimulus set.

It is premature to draw any firm conclusions about whether the young infant's representations of speech are structured with respect to phonetic segments or whether they take a more holistic form. Although we favor the latter view, our position is based on arguments about parsimony (not attributing to the infant the baggage of a segmental representation if the data can be explained without it), rather than on definitive evidence ruling out segmental representations. Nevertheless, any argument in favor of segmental representations will have to account for the fact that only in one case out of four (viz. the newborn results with the common vowels) are the data consistent with this position. Presumably any such account should also explain why newborns, but not 2-month olds, give evidence of representations structured according to common vowel segments.

In the end, however, it is clear that this issue will only be resolved with additional data. For this reason, we are in the process of conducting two additional experiments which may help clarify the nature of the young infant's representations. The first of these is a similar investigation to the second experiment, but with a simple rearrangement of the syllables that appear in the preshift and postshift sets. In the new experiment, infants are presented with a series of syllables sharing the same vowel (e.g. [bi], [di], [li], [mi]). Once again a new syllable with an identical vowel is added during the postshift period. However, in this instance, the new token [si] contains a consonant which is highly discriminable from those in the preshift set. Thus, if it were the case that the newborns in the [di] group in the second experiment failed to show an increase in postshift sucking because they recognized the common vocalic segment, then they should respond in a similar manner for this rearranged series. If on the other hand, the newborns responded as they did simply because their representations were not sufficiently detailed to distinguish between the new token [di] and one of the preshift tokens [bi], then the addition of a highly discriminable token like [si] to the postshift set may result in significant increases in postshift sucking. In addition, the question about whether the representation is structured in terms of phonetic segments would be clarified if information existed about alternative sorts of structuring, such as a syllable-based representation. Strong evidence in favor of a syllabic organization in the absence of further evidence for structuring by phonetic segments would tend to argue against the latter. To evaluate such an alternative organization, we are testing whether the infants' representations are structured in terms of some syllable-length units or not. To determine this, it is necessary to examine the infant's representation of

information in multisyllabic utterances. Hence, during the preshift phase, infants are presented with a series of multisyllabic utterances which share a common syllable (e.g. [bima], [bidu], [bisœ], and during the postshift phase a new item is added which either shares (e.g. [bigo]) or does not share (e.g. [tuko]) the common syllable.

Finally, there is one other aspect of the present results that should not be overlooked. There is some suggestion, for the first time, of a developmental effect in speech processing capacities during the first two months of life. Specifically, the newborn infants did not give evidence of distinguishing a new token differing only in its initial consonant from a preshift set of syllables sharing the same vowel. Whatever the representations employed by newborns and 2-month olds, it seems clear that some developments in information processing capacities are taking place during this period. Thus, while the 2-month old is able to detect changes involving a single stop consonant even in the face of a varied habituation set, the newborn may be limited to situations involving contrasts between single syllables. Further work in our laboratories suggests that similar constraints may operate for certain close vowel distinctions (e.g. [a] and [ʌ]) during the first two months of life. It is of great interest to know the extent to which any development of such capacities may be affected by linguistic input even at these very early stages. Thus, this is another area which merits further study.

The results of the two experiments presented here constitute only the first steps towards understanding how the child develops a representation of the sound properties of words that can be used in fluent speech recognition. In addition to the work with young infants, we need to begin examining more closely the way in which linguistic input leads to any re-organization of representations employed by children engaged in the acquisition of the phonology of a particular language. Despite the long road ahead, it seems likely that many important insights about the processes underlying fluent word recognition are at the end of such an undertaking.

Notes

1. This comparison of the word-learning and infant-speech discrimination tasks points out an important consideration which must be kept in mind in attempts to determine the infant's representation of speech sounds. One needs to establish whether a description structured in terms of experimenter-defined features is really the same one on which the infant-perceiver operates (Kemler Nelson, 1984). Hence, although the experimenter may be able to provide an analytic description in terms of phonetic features, this does not mandate that the infant has a phonetic feature based representation of speech sounds. Instead, there may be other descriptions which are compatible with the available data, and which might better characterize the infant's responses. One such alternative, given the speech discrimination results, is that the representations that infants operate on may be holistic ones.

2. Given that the presentation of stimuli is dependent on the infant's sucking behavior, to ensure that infants would have ample opportunity to hear the new item, its frequency of appearance in the set was

adjusted so that it occurred as often as all of the other members combined.

REFERENCES

1. Aslin, R.N., Pisoni, D.B., and Jusczyk, P.W. (1983). Auditory development and speech perception in infancy. In M. Haith and J. Campos (Eds.), Carmichael's handbook of child phonology: Infant development. New York: Wiley & Sons.
2. Bornstein, M.H., Kessen, W., and Weiskopf, S. (1976). Color vision and hue categorization in young infants. Journal of Experimental Psychology: Human Perception and Performance, 2, 115-129.
3. Cohen, L.B. and Strauss, M.S. (1979). Concept acquisition in the human infant. Child development, 50, 419-424.
4. Ferguson, C.A. (1985). Discovering sound units and constructing sound systems: It's child's play. In J.S. Perkell and D.H. Klatt (Eds.), Invariance and Variability in Speech Processes. Hillsdale, NJ: Lawrence Erlbaum Assoc.
5. Garnica, O. (1973) The development of phonemic speech perception. In T.E. Moore (ed.), Cognitive development and the acquisition of language. New York: Academic Press.
6. Goldsmith, J. (1976). An overview of autosegmental phonology. Linguistic Analysis, 2, 23-68.
7. Jusczyk, P.W. (1985). The high amplitude sucking procedure as a methodological tool in speech perception research. In G. Gottlieb and N.A. Krasnegor (Eds.), Infant methodology. Norwood, NJ: Ablex.
8. Jusczyk, P.W. (1985). On characterizing the development of speech perception. In J. Mehler and R. Fox (Eds.), Neonate cognition: Beyond the blooming, buzzing confusion. Hillsdale, NJ: Lawrence Erlbaum Assoc.
9. Kemler Nelson, D.G. (1984). The effect of intention on what concepts are acquired. Journal of Verbal Learning and Verbal Behavior, 23, 734-759.
10. Liberman, A.M., Cooper, F.S., Shankweiler, D.P., and Studdert-Kennedy, M. (1967). Perception of the speech code. Psychological Review, 74, 431-461.
11. Macken, M.A. (1980). Aspects of the acquisition of stop systems: A cross linguistic perspective. In G.H. Yeni-Komshian, J.F. Kavanagh and C.A. Ferguson (Eds.), Child phonology, Vol.1, Production. New York: Academic Press.
12. Mehler, J. (1985). Language related dispositions in early infancy. In J. Mehler and R. Fox (Eds.), Neonate cognition: Beyond the blooming, buzzing confusion. Hillsdale, NJ: Lawrence Erlbaum Assoc.
13. Menn, L. (1980). Phonological theory and child phonology. In G.H. Yeni-Komshian, J.F. Kavanagh and C.A. Ferguson (Eds.). Child phonology, Vol. 1, Production. New York: Academic Press.
14. Shepherd, R.N. (1972). The psychological representation of speech sounds. In E.E. David and P.B. Denes (Eds.), Human communication: A unified view. New York: McGraw-Hill.
15. Shvachkin, N.K. (1973). The development of phonemic speech perception in early childhood. In C.A. Ferguson and D.I. Slobin (Eds.), Studies of child language development. New York: Holt, Rinehart & Winston.

16. Vihman, M.M. (1978). Consonant harmony: its scope and function in child language. In J.H. Greenberg et al. (Eds.), Universals of human language, Vol. 2, Phonology. Stanford, CA: Stanford University Press.
17. Werker, J.F. and Tees R.C. (1983). Changes in categorization of speech sounds. Paper presented at the biennial meeting of the Society for Research in Child Development, Detroit, MI.

GENERAL DISCUSSION OF SESSION 6: PRIMARY SPEECH PERCEPTS

Chairman: A. Cohen

The title of this session suggests the search for perceptual units, which at one time in the past had already been designated as a wild goose chase. It seems clear by now that what we do in processing the incoming acoustic speech signal is highly task dependent. This goes for ordinary speech communication situations and all the more so in laboratory bound tasks.

The main issue at this session was the relation between, on the one hand, perceptual processing units as such, which are primarily stimulus bound, and can be considered to be the result of sensory transduction, i.e. auditory sensations, and, on the other hand, computational units which presuppose patterning of higher order mental constructs, such as are required, e.g., for the easy access of items in the mental lexicon. This apparent dichotomy can also be formulated in terms of proximal and distal objects.

Both aspects can be approached by mediation through the generally accepted notions derived from a linguistic analysis, giving rise to features, phonemic or allophonic segments, syllables, words, and even phrases. The smaller the chunk in focus the more stimulus bound the searcher's stance turns out to be.

The outcome of the discussion was that there are no ready to hand primary percepts, although there seemed to be a consensus that linguistic units certainly help to focus the various components that make up the highly complex speech signal. Strong claims were made for the allophone as an ordering principle particularly from the point of view of those studying the relations between dynamic spectral characteristics and the auditory domain, whereas others claimed the study of word recognition to be the most rewarding approach for getting to grips with the primary task of speech understanding. Thus we can distinguish between an analytical and a more global approach; these go hand in hand with the two main characteristics inherent in the speech signal, one leading to segmentation of the time axis by dint of which linguistically induced contrastive events can be discerned, and the other being a continuous scanning with the emphasis on the identification of the components making for the coherent shape of the speech signal. In the latter sense acoustic similarities help in localising and focusing on the speaking source. It is mainly through this coherence in natural speech, often lacking in synthesised versions, that, in conformity with the laws of form, comparable to Gestalt phenomena in visual processing, an interpretation can be given to the auditory input.

The paper by Jusczyk stressed the importance of studying the ontogenetic buildup of the perceptual discriminating faculty in focusing on minimal consonant and vowel distinction in neonates and very young infants. The object is to determine the relation between the perceptual characteristics of each of these two classes of segments in conjunction with those of the syllable. Moreover, a language specific shaping is not ruled out and constitutes a combined research object with Mehler, who found evidence for a different handling of syllabification between speaker of different languages, such as French and English, in a number of experimental tasks involving reaction time measurements.

Remez, using a more global approach, concentrated on continuities in the speech signal which, as such, help in organising the overall patterning from which segmental information can be removed and replaced by sinusoids without too great detriment to intelligibility.

PROPOSITION

One may wonder, at this juncture, whether a fresh phenomenology of the speech chain is in order. This might help in setting up research objectives for the near future in which auditory sensation as the most peripheral input to the speech processing mechanism is also studied in its more global aspects and made to match better with the quintessential character of connected speech as a continuum. Any subsequent segmentation must be seen for what it is most of the time, i.e. derived from a linguistic analysis. While such a segmentation seems inevitable, it should be acknowledged that the choice of level of analysis is arbitrary or due to the experimenter's taste.

If we are interested in the organ of hearing we are inclined to come up with a different choice of units than if the aim is to go in for high quality synthesis in which perceptual criteria are invoked to obtain optimal intelligibility, naturalness, speaker identity, and variation.

Our conclusion may be that there is no such thing as The psychophysics of speech perception, yet the combined efforts of experts from a variety of disciplines are needed to make the study of speech perception a rewarding pursuit. In this pursuit the chase for perceptual units should be called off for the time being until the speech science community has sufficiently assimilated its awareness of the impact of linguistic propensities. In the meantime, contributions from the linguistic points of view may help in deciding which features in the speech chain are due to universal and which to more language specific properties.

It is my belief, as a linguist by training, that though linguistics in the past may have been a bane to speech studies at times, we should turn that into a boon, if properly dimensioned, in the future.

Chapter 7

PSYCHOPHYSICS AND SPEECH PERCEPTION IN THE HEARING-IMPAIRED

RELATIONSHIP BETWEEN PSYCHOPHYSICAL ABILITIES AND SPEECH PERCEPTION FOR SUBJECTS WITH UNILATERAL AND BILATERAL COCHLEAR HEARING IMPAIRMENTS*

Brian C.J. Moore and Brian R. Glasberg

Department of Experimental Psychology, University of Cambridge,
Downing Street, Cambridge CB2 3EB, England.

INTRODUCTION

Over the past few years several groups have reported studies of the relationship between psychophysical abilities and speech perception in the hearing impaired (e.g., Dreschler and Plomp, 1980, 1985; Festen and Plomp, 1983; Patterson et al., 1982; Tyler et al., 1982). Such studies can have a number of different (non-exclusive) purposes, including: 1) To gain more insight into the difficulties of speech perception experienced by the hearing impaired, especially in noisy situations; 2) To make inferences about the relative importance of different types of information in speech (e.g., spectral versus temporal); 3) to provide guidelines for the design of "signal-processing" hearing aids which are intended to (partially) compensate for one or more of the deficits found in the hearing impaired. Our own work in this area has all three of these purposes, but in this paper we will concentrate on the first two.

Although the rationale for our work is similar to that for the studies referenced above, there are some critical differences in the ways in which the experiments have been conducted. Firstly, most previous studies have used relatively untrained subjects, and the time available for each test has been very limited. This means that performance may sometimes have been limited not by the sensory capacities of the subjects, but by "higher level" factors related more to cognitive abilities. Our subjects took part in an extensive series of carefully conducted tests, usually attending for two hours per week over a period of several months. Sufficient practice was given in each task for performance to stabilise. The subjects can thus be considered as relatively highly trained.

Secondly, most previous studies have used subjects with a wide range of audiometric configurations; both the average degree of hearing loss and the slope of the audiogram have varied considerably across subjects. This causes a number of problems: 1) When testing with speech stimuli, the proportion of the speech energy which is above absolute threshold varies considerably from one subject to another, except at very high sound levels. These variations in the audibility of speech can obscure the effects of other psychoacoustic factors; 2) Steeply sloping audiograms can make it difficult to obtain and interpret psychoacoustical measures of frequency selectivity; 3) It is difficult to know whether to compare subjects at equal sensation

*This work was supported by the Medical Research Council (U.K.).

levels or equal sound pressure levels. We have used subjects with a relatively narrow range of threshold losses (40 - 65 dB HL), at least over the frequency range 500 - 2000 Hz. In addition, most subjects had relatively uniform losses as a function of frequency. We have tested all subjects at relatively high sound levels, chosen so that the stimuli were generally 20 dB or more above absolute threshold.

Finally, in previous studies relationships between psychoacoustic abilities and speech perception may have been obscured by the fact that speech perception depends on cognitive processing as well as sensory processing. The former may vary considerably across subjects, particularly when elderly subjects are used. We have concentrated on testing subjects with unilateral cochlear impairments; performance using the normal ear then acts as a control for performance in the impaired ear. Assuming that cognitive processing will be the same for the two ears of a given subject, we hoped that differences in the speech intelligibility scores for the two ears would give a direct measure of the deleterious effects on speech perception of the deficits in sensory processing of the impaired ear.

EXPERIMENTAL MEASUREMENTS

Subjects

Eight subjects with unilateral cochlear impairments and six subjects with bilateral cochlear impairments were used (aged 43 - 72 years). All subjects were tested to exclude the possibility of conductive or retrocochlear involvement; details of testing may be found in Moore et al. (1985). Table 1 shows the means and standard deviations of the absolute thresholds for the impaired ears and normal ears. Thresholds given in dB SPL were obtained using the two-alternative forced-choice task described below (200-ms tones). Thresholds given in dB HL were obtained using standard audiometry. In general, the hearing losses in the impaired ears of the subjects with unilateral impairments were slightly greater than those for the bilaterally impaired subjects.

Psychoacoustic tests were generally conducted at three centre frequencies, 0.5, 1.0 and 2.0 kHz. Most subjects had absolute thresholds in the range 40 to 60 dB HL at these frequencies, but a few had thresholds as low as 20 dB HL or as high as 70 dB HL at one of the test frequencies. Results will only be given for conditions where the signals were more than 10 dB above absolute threshold. Subjects were paid for their services.

General test method

All psychoacoustic thresholds were measured using an adaptive two-interval two-alternative forced-choice procedure that estimates the 71% point on the psychometric function (Levitt, 1971). After two correct responses the task was made one step harder and after each incorrect response it was made one step easier. Testing continued until 16 turnarounds had occurred (12 for tasks involving the detection of frequency or amplitude modulation), and threshold was taken as the mean of the values at the last 12 (the last eight for modulation detection). Feedback was provided by lights on the response box. Each threshold reported is the mean of at least two runs (usually three or

Table 1. Means and standard deviations (in brackets) of the absolute thresholds for the normal and impaired ears, in dB SPL (Sennheiser HD 414 earphones) and dB HL (TDH 39 earphones).

Frequency, kHz		0.25	0.5	1.0	2.0	4.0	8.0
Impaired ears (bilateral)	dB SPL	49 (14)	46 (10)	48 (12)			
	dB HL 41 (9)	44 (10)	44 (10)	50 (11)	59 (20)	65 (18)	
Impaired ears (unilateral)	dB SPL	59 (15)	58 (10)	50 (8)			
	dB HL 59 (10)	57 (12)	57 (6)	51 (10)	54 (15)	71 (13)	
Normal ears (unilateral)	dB SPL	16 (6)	10 (7)	13 (7)			
	dB HL 16 (6)	14 (3)	11 (8)	12 (10)	29 (13)	42 (20)	

four). Subjects were tested on each task until their performance was stable. Each ear of each subject was tested separately, the non-test ear being masked with pink noise (25 dB spectrum level at 1 kHz) where appropriate. Stimuli were delivered via Sennheiser HD 414 earphones. Unless stated otherwise, all tests were conducted at three centre frequencies, 0.5, 1.0, and 2.0 kHz. The silent interval between the two stimuli in a trial was always 500 ms.

The method used for determining the threshold for speech, in quiet and noise, will be described later.

PSYCHOACOUSTIC TESTS

Detection of a Temporal Gap in a Noise Band

Subjects were required to detect a temporal gap in a gated bandpass noise, presented in a complementary continuous band-reject background. The method is very similar to that described by Shailer and Moore (1983). The bandpass noise was arithmetically centred at 0.5, 1.0 or 2.0 kHz and the bandwidth was one half of the centre frequency. The notch in the background noise was of the same width. The spectrum level of the bandpass noise, in its passband, was 60 dB at 0.5 kHz, 57 dB at 1.0 kHz, and 54 dB at 2.0 kHz, giving an overall level of 84 dB SPL. The spectrum level of the background noise was 20 dB below that of the bandpass noise. The steady-state duration of each burst of noise in a trial was 400 ms, and onsets and offsets were shaped with 10-ms raised-cosine functions. The fall and rise time of the gap were each 0.5 ms (raised-cosine function), giving a minimum gap, measured at the 6-dB down points, of 0.5 ms.

Detection of Amplitude Modulation

Subjects were presented with two successive tone pulses, one of which had a constant amplitude, and the other of which was sinusoidally amplitude modulated at a 4-Hz rate. Subjects were required to detect which tone was modulated. The level of the unmodulated tone was 80 dB SPL. Each tone lasted 1020 ms, including the 10-ms raised-cosine onset and offset ramps. Tests were conducted both in

quiet and in the presence of continuous one-octave wide noise; this noise was designed to mask the upper side of the excitation pattern of the tone, to assess the influence of nonlinearities on the high-frequency side of the excitation pattern. The lower cutoff frequency of the noise (-3 dB point) was twice the signal frequency, slopes were 96 dB/oct, and the overall level of the noise was 77 dB SPL.

Intensity Discrimination of Tone Pulses

Subjects were required to identify which of two tone pulses was more intense. The level of the less intense tone pulse was 80 dB SPL. Each tone pulse had a 200-ms steady state portion and 10-ms raised-cosine onset and offset ramps. Again, tests were conducted both in quiet and in the presence of noise, which was identical to that described for amplitude modulation detection.

Detection of Frequency Modulation

Subjects were presented with two successive tone pulses one of which had a constant frequency, the other of which was sinusoidally frequency modulated at a 4-Hz rate. Other details are the same as for amplitude modulation detection.

Frequency Discrimination of Tone Pulses

Subjects were required to identify which of two tone pulses had the higher frequency. Other details are the same as for the intensity discrimination of tone pulses.

Estimation of the Auditory Filter Shape

This test was intended to estimate the degree of frequency selectivity of the subjects by measuring the shape of the auditory filter using a notched-noise masker (Patterson, 1976). The method used is exactly as described by Glasberg and Moore (1986) and the reader is referred there for details. Briefly, the threshold of a sinusoidal signal was measured as a function of the width of a spectral notch in a noise masker. The notch was positioned both symmetrically and asymmetrically about the signal frequency, in order to assess the asymmetry of the filter. The noise spectrum level was 50 dB. Both the signal and the masker were gated, with a 200-ms steady-state portion and 10-ms raised-cosine onset and offset ramps. Since this procedure takes a considerable time, only a centre frequency of 1.0 kHz was used. Details of the results obtained for this test are given in Glasberg and Moore (1986). In this paper we will make use of only two summary statistics: 1) the equivalent rectangular bandwidth (ERB) of the auditory filter, which is related to the shape of the tip of the filter; 2) the attenuation of the filter, in dB, averaged for the frequencies 0.6 and 1.4 kHz (i.e., 40% below and above the centre frequency), which is related to the steepness of the skirts of the filter. We will refer to this measure as ATT.4. Note that ATT.4 is always a negative number; smaller (more negative) values indicate sharper filters.

SPEECH TESTS

We have conducted two types of speech test. The first is similar to that used by Plomp and his colleagues (Plomp and Mimpens, 1979), and involves the measurement of the speech reception threshold (SRT) for sentences. The SRT is the level of the speech required for 50% intelligibility, and we have measured it both in quiet and in two levels of speech-shaped noise, 60 and 75 dB SPL. The SRT gives a measure of the overall speech communication abilities of the subject.

The other type of test uses the Boothroyd word lists (Boothroyd, 1968). Each list contains 10 monosyllabic words comprised of 30 common phonemes. This test allows a more analytic study of the patterns of errors produced by subjects. Again, tests have been conducted both in quiet and in various types of background noise.

We present here details only of the SRT measurements, since the tests using the Boothroyd lists are still in progress.

Measurement of SRTs in Quiet and in Noise

The stimuli were presented via a Monitor Audio MA4 loudspeaker directly in front of the subject at a distance of 1.3 m. Tests were conducted in a sound-attenuating room with sound absorbing (but not completely anechoic) walls. The non-test ear was plugged with an EAR plug and masked with pink noise. Test materials were tape recordings (Revox A77) of the BKB sentence lists (Bench and Bamford, 1979), each list containing 16 sentences with 3 or 4 key words per sentence. Scoring was by key word. The SRT for a given condition (quiet, 60-dB noise or 75-dB noise) was determined as follows. The first sentence of a list was presented at a level about 5 dB above the value of the SRT determined in practice runs. If the subject scored 2 or 3 key words correct out of 3 (or 3 or 4 out of 4) the speech level was decreased. Otherwise the level was increased (except when the subject scored 2 out of 4, when the level was left unchanged). Initially the level was changed in steps of 5 dB. After three turnarounds the step size was decreased to 2 dB, and the initial estimate of threshold was obtained by averaging the levels visited on all subsequent trials. A small correction was then applied to allow for the fact that the percentage correct on those trials often differed from 50%; see Laurence, Moore and Glasberg (1983) for details of the correction. The standard deviation of the SRT measured in this way is about 1.4 dB. Each SRT was measured at least twice (using different lists) so the average standard error of the SRTs reported is 1 dB or less.

RESULTS

General Features

Some of the results of the individual psychoacoustic tests have been published elsewhere, and we will restrict ourselves here to a brief outline.

In line with previous reports (Fitzgibbons and Wightman, 1982; Tyler et al., 1982) gap thresholds were usually, but not always, larger for the impaired ears than for the normal ears. For our subjects the

differences tended to increase with decreasing frequency. The differences between the impaired and normal ears were reduced, but not always eliminated, when the normal ears were assessed at sensation levels (SLs) equivalent to those used in the impaired ears. This suggests that part of the difference between the normal and impaired ears, when tested at equal sound levels, can be attributed to the lower SLs in the impaired ears. More details can be found in Glasberg, Moore, and Bacon (1986).

Thresholds for the detection of amplitude modulation, and for the intensity discrimination of pulsed tones, were, on average, similar for the normal and impaired ears, although performance did tend to be slightly better for the impaired ears (Moore and Glasberg, 1986a). The bandpass noise tended to impair performance, suggesting that information from the upper side of the excitation pattern is important for intensity discrimination. The noise had a greater effect for modulation detection than for the discrimination of pulsed tones, and its effect was generally greater for the normal ears. This can possibly be attributed to a reduced nonlinearity in the growth of excitation on the high-frequency side of the excitation pattern in the impaired ears.

Thresholds for the detection of frequency modulation were generally larger for the impaired ears than for the normal ears, and the difference was larger at low frequencies. Frequency discrimination of pulsed tones was often considerably worse for the impaired ears, but there was no clear trend with frequency. The bandpass noise generally had little effect on performance, both for the detection of frequency modulation, and for the frequency discrimination of pulsed tones, except for subjects with very little frequency selectivity in their impaired ears. This suggests that information from the high-frequency side of the excitation pattern is not important for frequency discrimination. More information about the results for frequency discrimination may be found in Moore and Glasberg (1986a).

For each subject with a unilateral impairment, the auditory filter was broader in the impaired ear than in the normal ear, but the extent of the difference varied considerably across subjects. A common pattern was that in the impaired ears the auditory filters had very shallow low-frequency skirts, indicating a marked susceptibility to the upward spread of masking. The upper skirt was sometimes less steep than normal and sometimes nearly normal. More details of the results of the filter-shape measurements may be found in Glasberg and Moore (1986).

In summary, cochlear hearing impairment in our subjects was associated with reduced temporal resolution (as measured by gap detection), reduced frequency discrimination, and reduced frequency selectivity. Intensity discrimination was not markedly affected.

The measurements of the SRTs are consistent with previous work in showing higher values for the impaired ears, both in quiet and in noise. The elevation in quiet is inevitable, given the higher absolute thresholds for the impaired ears. However, the level of the 75-dB noise was sufficient to raise the SRTs well above those measured in quiet, and even at this noise level the SRTs were higher for the impaired ears. For the 75-dB noise, the mean SRT for the normal ears (standard deviations in brackets), expressed as the ratio of the peak

level of the speech to the RMS level of the noise in dB, was 1.3 (2.0). For the impaired ears of the subjects with unilateral impairments the mean SRT was 8.7 (2.4). Thus the impaired ears require, on average, a 7.4 dB higher speech-to-noise ratio to achieve 50% intelligibility. For the subjects with bilateral impairments the mean SRT was 3.3 (4.3). The better performance for the bilaterally impaired subjects might be a consequence of the fact that their hearing losses were slightly less than those for the impaired ears of the unilaterally impaired subjects (see Table 1). The bilateral subjects also tended to perform better on the psychoacoustic tasks described above, which may indicate that the effects of the pathology were more severe for those with unilateral impairments. Four of the subjects with unilateral losses had Meniere's disease, which may have particularly severe effects on the ability to understand speech.

Overall, these results suggest that the subjects have difficulty understanding speech in noise even for stimuli well above absolute threshold. We turn now to a consideration of the relationship of this difficulty to the psychoacoustic variables.

RELATIONSHIP BETWEEN SPEECH INTELLIGIBILITY AND PSYCHOACOUSTIC FACTORS

In this section we will consider correlations between the psychoacoustic measurements and the measures of speech intelligibility, particularly in noise. We are especially concerned with finding relationships which are not simply the result of the mutual interdependence of the variables with absolute threshold. To investigate this we have determined partial correlations for which the effect of absolute threshold at the test frequency is "partialed out" or held constant. If the correlation between two variables is reduced by this then we infer that their interdependence arises in part through the agency of variations in absolute threshold; and if the partial correlation is very small, we infer that their interdependence is entirely attributable to that variable. Conversely, if the partial correlation between the two variables is larger than the original correlation, we infer that the variations in absolute threshold were obscuring the stronger connection, or "masking" the correlation (Kendall and Stuart, 1967).

As mentioned in the introduction, the relationships between psychoacoustic measures and speech intelligibility may be obscured by differences in the cognitive abilities of the subjects. Hence, as well as determining correlations between the psychoacoustic measures and the SRTs, as has been done in several previous studies, we have also made use of the differences between the SRTs for the two ears of our subjects with unilateral impairments. The rationale is this: for the normal ears, there is relatively little variation in the psychoacoustic measures from one subject to another. For example, the ERB of the auditory filter at 1 kHz has a mean value of 153 Hz, and a standard deviation of only 11 Hz. Thus, it seems reasonable to assume that variations in the SRTs in noise for the normal ears (which have about the same standard deviation as for the impaired ears) arise primarily from variations in cognitive ability. However, the cognitive ability of a given subject is, presumably, the same for sounds presented to either ear. Thus the difference in SRT for the two ears can be taken as a measure of the specific effect of the deficit in sensory processing of

the impaired ear. Using this rationale, we have determined the correlations between the psychoacoustic measures in the impaired ear and the difference in SRT for the normal and impaired ears for all subjects with unilateral impairments.

Frequency Selectivity and the SRT

Considering the results for all ears (normal and impaired), both the SRTs in quiet and in noise were significantly ($P<0.01$) correlated with the two auditory-filter measures, the ERB and ATT.4; correlations were higher for the latter. However, the correlations were close to zero after partialling out the effects of absolute threshold at 1 kHz, which is not surprising since ATT.4 was also highly correlated with the absolute threshold at 1 kHz, ($r = 0.91$). For the impaired ears only, the correlations were lower, and again became close to zero after partialling. Finally, the difference in SRTs between the normal and impaired ears was not significantly correlated with either of the filter measures in the impaired ear, either before or after partialling. The results for the measure ATT.4 are summarised in table 2 and figure 1.

Table 2. Correlations and partial correlations (in brackets) between the SRTs and the measure of filter sharpness, ATT.4. Correlations significant at the 0.05 level or better are indicated by*.

	SRT in quiet	SRT in 60-dB noise	SRT in 75-dB noise
All ears	0.85*(0.06)	0.79*(0.11)	0.58*(0.11)
Impaired ears	0.61*(0.10)	0.50*(0.03)	0.49*(0.16)
Impaired-normal (unilaterals)	0.26 (0.16)	-0.11(-0.30)	-0.08 (0.01)

Overall, the results make it difficult to rule out the idea that the correlation between frequency selectivity and the intelligibility of speech is mediated by the relationship of both to absolute threshold.

Intensity Discrimination

The measures of amplitude-modulation detection and intensity discrimination of pulsed tones did not show any consistent relationship to the SRTs, so we will not discuss them further here.

Temporal Resolution

The gap thresholds at all three centre frequencies showed positive correlations with the SRTs. The correlations were highest for a centre frequency of 1 kHz, so we will concentrate on them. Table 3 and figure 2 summarise the results. Notice that, for the SRTs in 75-dB noise, the correlations remain significant after partialling. This was also true at 2 kHz but not at 0.5 kHz. Furthermore, the correlations

between the gap thresholds in the impaired ears and the differences between the SRTs for the impaired and normal ears (third row of table) are higher than those for the other two rows, and actually increase slightly after partialling.

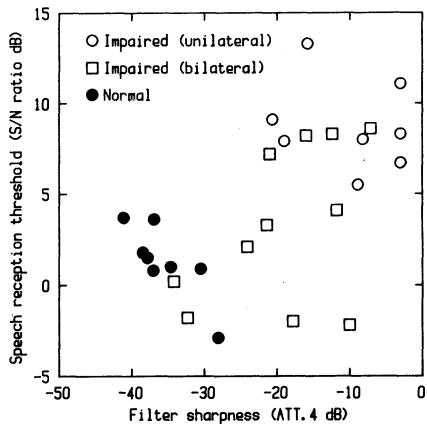


FIGURE 1. The SRT for a noise level of 75 dB SPL is plotted as a function of the measure of filter sharpness, ATT.4. Solid circles show results for the normal ears of the subjects with unilateral impairments. Open circles show results for the impaired ears of those subjects. Open squares show results for subjects with bilateral impairments.

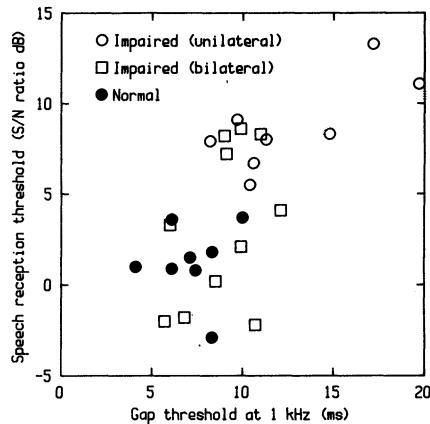


FIGURE 2. The SRT for a noise level of 75 dB SPL is plotted against the gap threshold at a centre frequency of 1 KHz. Symbols have the same meaning as for Figure 1.

These results suggest that temporal resolution is strongly related to the intelligibility of speech, both in quiet and in noise, and that the effect is not mediated by variations in absolute threshold. The higher correlations found by using differences between the SRTs for the normal and impaired ears support our idea that cognitive differences between subjects can obscure relationships between psychoacoustic measures and speech perception.

Table 3. Correlations between the gap thresholds at 1 kHz and the SRTs. Format is the same as for table 2.

	SRT in quiet	SRT in 60-dB noise	SRT in 75-dB noise
All ears	0.59*(0.20)	0.70*(0.49*)	0.67*(0.50*)
Impaired ears	0.55*(0.30)	0.67*(0.51*)	0.64*(0.51*)
Impaired-normal	0.85*(0.90*)	0.83*(0.93*)	0.68*(0.89*)

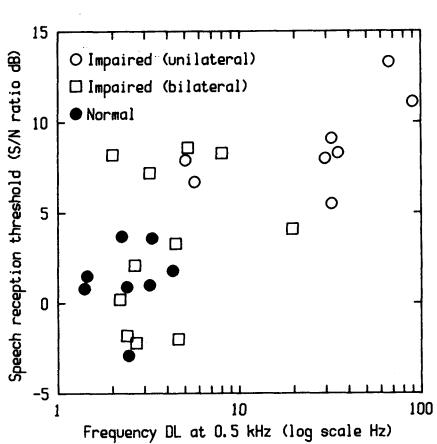


FIGURE 3. The SRTs for a noise level of 75 dB SPL are plotted as a function of the frequency DL for pulsed tones at a centre frequency of 0.5 kHz (Frequency DL averaged for conditions with and without noise).

Frequency Discrimination

Most of the measures of frequency discrimination showed positive correlations with the SRTs, but the correlations decreased with increasing frequency, and failed to reach significance at 2 kHz, either before or after partialling. Frequency DLs for pulsed tones at 0.5 kHz showed rather high correlations with the SRTs; this was true both for DLs measured in quiet and for those measured in bandpass noise, and so we decided to average the frequency DLs measured with and without noise. Table 4 and figure 3 show the results. In general the correlations with the SRTs in noise remained significant when the effect of absolute threshold was partialled out. And, as with the gap thresholds, the highest correlations were obtained with the differences between the SRTs for the normal and impaired ears.

Table 4. Correlations between the frequency DLs for pulsed tones at 0.5 kHz and the SRTs. Format is the same as for table 2.

	SRT in quiet	SRT in 60-dB noise	SRT in 75-dB noise
All ears	0.54*(0.03)	0.70*(0.50*)	0.64*(0.44*)
Impaired ears	0.61*(0.33)	0.74*(0.58*)	0.60*(0.43*)
Impaired-normal	0.93*(0.89*)	0.88*(0.89*)	0.80*(0.73*)

GENERAL DISCUSSION AND CONCLUSIONS

Our approach of using subjects with unilateral impairments appears promising, in that the differences in the SRTs between the impaired and normal ears showed rather high correlations with two of the psychoacoustic measures: gap threshold at 1.0 kHz and the frequency DL for pulsed tones at 0.5 kHz. The correlations were higher than those obtained with the raw SRTs, whether considering both impaired and normal ears or just the impaired ears. Furthermore, the correlations remained significant, or even increased, after partialling out the effects of absolute threshold. On the other hand, the difference in SRTs between the normal and impaired ears did not correlate significantly with the measures of frequency selectivity, either before or after partialling out the effects of absolute threshold.

We have argued elsewhere that the frequency discrimination of pulsed tones probably depends on the use of information contained in the fine time-structure of neural firing patterns in the auditory nerve (Moore and Glasberg, 1986a,b). Thus both of the psychoacoustic factors which we have found to be related to speech intelligibility may be considered as measures of temporal processing. Our results suggest that deficits in temporal processing associated with cochlear impairment may have more important effects on speech intelligibility than deficits in frequency selectivity.

REFERENCES

1. Bench, J. and Bamford, J. (1979). Speech Hearing Tests and the Spoken Language of Hearing Impaired Children. Academic Press, London.
2. Boothroyd, A. (1968). Developments in speech audiometry. Sound, 2, 3-10.
3. Dreschler, W.A. and Plomp, R. (1980). Relations between psychophysical data and speech perception for hearing-impaired subjects. I. Journal of the Acoustical Society of America, 68, 1608-1615.
4. Dreschler, W.A. and Plomp, R. (1985). Relations between psychophysical data and speech perception for hearing-impaired subjects. II. Journal of the Acoustical Society of America, 78, 1261-1270.

5. Festen, J.M. and Plomp, R. (1983). Relations between auditory functions in impaired hearing. Journal of the Acoustical Society of America, 76, 652-662.
6. Glasberg, B.R. and Moore, B.C.J. (1986). Auditory filter shapes in subjects with unilateral and bilateral cochlear impairments. Journal of the Acoustical Society of America, 79, 1020-1033.
7. Glasberg, B.R., Moore, B.C.J., and Bacon, S.P. (1986). Gap detection and masking in hearing-impaired subjects. Submitted to Journal of the Acoustical Society of America.
8. Kendall, M.G. and Stuart, A. (1967). The Advanced Theory of Statistics, Vol.II, Second Ed. Griffin, London.
9. Laurence, R.F., Moore, B.C.J., and Glasberg, B.R. (1983). A comparison of behind-the-ear high-fidelity linear aids and two-channel compression aids, in the laboratory and in everyday life. British Journal of Audiology, 17, 31-48.
10. Levitt, H. (1971). Transformed up-down methods in psychoacoustics. Journal of the Acoustical Society of America, 49, 467-477.
11. Moore, B.C.J., Glasberg, B.R., Hess, R.F. and Birchall, J.P. (1985). Effects of flanking noise bands on the rate of growth of loudness of tones in normal and recruiting ears. Journal of the Acoustical Society of America, 77, 1505-1513.
12. Moore, B.C.J. and Glasberg, B.R. (1986a). The relationship between frequency selectivity and frequency discrimination for subjects with unilateral and bilateral cochlear impairments. In: Auditory Frequency Selectivity, edited by B.C.J. Moore and R.D. Patterson. Plenum, New York.
13. Moore, B.C.J. and Glasberg, B.R. (1986b). The role of frequency selectivity in the perception of loudness, pitch and time. In: Frequency Selectivity In Hearing, edited by B.C.J. Moore. Academic, New York.
14. Patterson, R.D. (1976). Auditory filter shapes derived with noise stimuli. Journal of the Acoustical Society of America, 59, 640-654.
15. Patterson, R.D., Nimmo-Smith, I., Weber, D.L. and Milroy, R. (1982). The deterioration of hearing with age: Frequency selectivity, the critical ratio, the audiogram and speech threshold. Journal of the Acoustical Society of America, 72, 1788-1803.
16. Plomp, R. and Mimpen, A.M. (1979). Improving the reliability of testing the speech reception threshold for sentences. Audiology, 18, 43-52.
17. Shailler, M.J. and Moore, B.C.J. (1979). Gap detection as a function of frequency, bandwidth and level. Journal of the Acoustical Society of America, 74, 467-473.
18. Tyler, R.S., Summerfield, Q., Wood, E.J., and Fernandes, M.A. (1982). Psychoacoustic and phonetic temporal processing in normal and hearing-impaired listeners. Journal of the Acoustical Society of America, 72, 740-752.

SPEECH-RECEPTION THRESHOLD IN A FLUCTUATING BACKGROUND SOUND AND ITS POSSIBLE RELATION TO TEMPORAL AUDITORY RESOLUTION

J.M. Festen

Experimental Audiology, ENT Department, Free University Hospital, de Boelelaan 1117, 1081 HV Amsterdam, The Netherlands

INTRODUCTION

The speech-reception threshold (SRT) in noise is a frequently encountered limitation in everyday aural communication, especially for hearing-impaired listeners. The SRT for sentences can be described with a simple signal-to-noise ratio model (Plomp, 1978) containing two parameters related to hearing loss: 'A' for attenuation and 'D' for distortion. Hearing loss for speech in noise is represented by the parameter D and hearing loss for speech in quiet by the sum of A and D, both in decibels. In normal-hearing listeners the SRT for short meaningful sentences (50% correct) in noise is reached for about -5 dB signal-to-noise ratio. At threshold, hearing-impaired listeners need a better signal-to-noise ratio, up to 10 dB, depending on the kind of hearing loss ($D < 10$ dB). Although hearing loss for speech in noise may seem very moderate in terms of decibels, it should be noted that in critical conditions 1 dB increase in signal-to-noise ratio gives an 18% higher intelligibility score for sentences (Duquesnoy, 1983). As a result, in such conditions, speech that is intelligible to normal-hearing listeners may be completely unintelligible to hearing-impaired listeners.

So far we have dealt with continuous noise. However, fluctuating backgrounds constitute a more realistic kind of interference for speech and appear to give hearing-impaired listeners an even greater handicap than continuous noise. In experiments with a second speech source as the interfering sound, Duquesnoy (1983) obtained for normal-hearing listeners a 7-dB lower SRT than in continuous noise with the same level and spectrum. This profit from the relatively silent intervals in the competing speech was not found for a group of elderly listeners with marginal hearing losses. With interrupted noise (10-Hz interruption rate) de Laat and Plomp (1983) obtained for normal-hearing listeners an improvement of 23 dB in SRT over conditions with continuous noise of equal energy. For young hearing-impaired listeners they found much less improvement than for normal hearing.

Now our question is, can we explain the differences in threshold between the various fluctuating maskers, including a speech masker, in terms of properties of the auditory system. To this end we measured the SRT for sentences in modulated noise as a function of modulation frequency. Speech, however, is not only intensity-modulated but is also characterised by fluctuations in the frequency domain, like the changes caused by shifting formants. In order to include this factor and to estimate its contribution to the masking of speech, we also measured the SRT for sentences masked by a noise stimulus split

into two frequency bands (one below and one above 1000 Hz), which were modulated in anti-phase. The next question is, can we bring the speech-reception threshold in a fluctuating background in relation with temporal auditory resolution. Therefore, we measured the masking of short clicks in peaks and troughs of sinusoidally intensity-modulated (SIM) noise as a function of modulation frequency.

THE EFFECT OF MASKER FLUCTUATIONS ON THE SRT

In this experiment listeners were requested to reproduce short meaningful Dutch sentences, recorded from a female speaker, and presented against a noise background with a spectrum equal to the long-term average spectrum of the sentences. In the modulated-masker conditions the modulation frequencies were 4, 8, 16, and 32 Hz with a sinusoidal variation of the intensity and a depth of 100 %. Each of these four modulation frequencies was tested twice: once with in-phase modulation over the whole spectrum and once with the frequencies beyond 1000 Hz modulated in anti-phase. Additionally, nonmodulated noise and continuous discourse were used as maskers, making a total of ten measuring conditions. The discourse masker, read by the same female speaker who read the signal (equal long-term average spectra), was made unintelligible by time-reversed presentation. The sequence of measuring conditions was counterbalanced for every ten subjects according to a digram-balanced Latin square.

In all conditions the masker was presented at a sound-pressure level of 75 or 80 dBA to the normal-hearing and hearing-impaired listeners, respectively. For each condition the SRT was determined with a list of 13 sentences, unknown to the listener, by using a simple up-and-down procedure for the presentation level. The procedure required from the listener correct replication of the entire sentence for a correct response. The step size used was 2 dB. The average presentation level after the fourth sentence was taken as the SRT for that particular condition.

Twenty university students volunteered as normal-hearing listeners in this study and were tested monaurally at the ear of their preference (75 dBA masker level). The hearing-impaired listeners were twelve pupils of a highschool for the hearing impaired. They had sensorineural hearing losses and were tested at their better ear (80 dBA masker level). Pure-tone acuity (PTA, average hearing level for 500, 1000, and 2000 Hz) ranged from 12 to 48 dB in this group. Additionally, three listeners with a more severe loss (average PTA: 57 dB) were tested with a masker level of 90 dBA. Because the number of hearing-impaired listeners is not a multiple of ten, the measuring conditions were not completely counterbalanced for this group.

The results for the two groups of subjects are given in Fig. 1. The standard deviation between subjects, pooled over all conditions, is 2.3 dB for the hearing-impaired listeners and 1.6 dB for the normal-hearing listeners. Only with speech as a masker, much larger interindividual differences are found for the normal-hearing listeners ($\sigma=5.1$ dB). At the optimum modulation frequency of the noise, normal-hearing listeners gain nearly 5.5 dB in signal-to-noise ratio as compared to unmodulated noise. On the average nearly the same gain is obtained with competing discourse as a masker. For hearing-impaired

listeners, however, the gain with optimum modulation of the masker is only 1.2 dB, and for speech as a masker the speech-reception

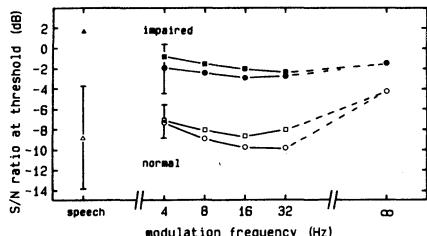


FIGURE 1, Average speech-perception threshold in modulated noise as a function of modulation frequency for 20 normal-hearing listeners (open symbols) and 12 hearing-impaired listeners (closed symbols). The circles are for in-phase modulation of the whole spectrum and the squares are for conditions in which the frequencies below and above 1000 Hz are modulated in anti-phase.

threshold is even raised by 3.2 dB. As a result the difference between the two groups of listeners increases from 2.7 dB for unmodulated noise to 10.5 dB for competing speech with the same character as the signal. For the hearing-impaired listeners a significant correlation ($r = 0.79$) is found between the SRT in unmodulated noise and the SRT with competing speech. No such correlation was found for the normal-hearing listeners. Nearly independent of the modulation frequency, for both groups of listeners a slightly higher SRT is found when a part of the masker spectrum is modulated in anti-phase. The data for the three listeners with more severe hearing losses are not shown in Fig. 1, but they are essentially similar to the data of the other hearing-impaired listeners.

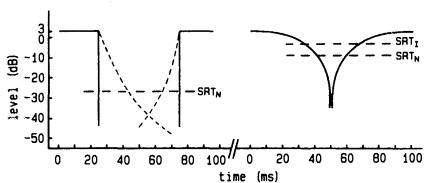


FIGURE 2, Schematic representation of the envelope of the interrupted masker used by de Laat and Plomp (1983) (left) and the SIM masker used in the present study (right). The curved dashed lines represent forward and backward masking for a masker of 80 dB as inferred from Plomp (1964) and Elliott (1971), respectively. The horizontal dashed lines represent the SRT for normal hearing (N) and impaired hearing (I).

In modulated noise, the lowest SRT is found for modulation frequencies between 16 and 32 Hz in both groups of subjects. For lower modulation frequencies complete words may be masked, with a detrimental effect on sentence intelligibility. For higher modulation frequencies the limited temporal resolution of the auditory system hampers the release from masking in the relatively silent intervals in the masker. The much smaller release from masking in hearing-impaired listeners compared to normal-hearing listeners might be

caused by a reduced temporal resolution of the auditory system. The release from masking found in this experiment is far less than the 23 dB obtained for normal-hearing listeners in interrupted noise by de Laat and Plomp (1983). A possible explanation for this large difference in SRT obtained with about the same modulation frequency is schematically presented in Fig. 2. On a decibel scale the SIM noise has a broad maximum and a very narrow minimum. The interrupted noise used by de Laat and Plomp has a very broad minimum. In this case the amount of masking in the minima is likely to be determined by the sum of forward and backward masking from the two adjacent noise bursts, as indicated with dashed lines. The speech level at threshold found for the two maskers is shown by the horizontal long-dash lines. The SRT will be determined by the signal-to-noise ratio in the masker minima. This ratio, roughly represented by the area between the speech-level dashed line and the masking curve, is about the same for the two maskers. The troughs in the masking function of the SIM noise are less deep than for the interrupted noise, and this seems to account for the threshold differences between the two experiments. In measurements with hearing-impaired listeners the SIM noise has the advantage of a smaller risk of producing results contaminated with the absolute threshold. This simple explanation for the difference in SRT between different maskers suggests a close relationship between auditory temporal resolution as inferred from the masking experiments and the SRT in modulated noise.

AUDITORY TEMPORAL RESOLUTION

Temporal resolution was deduced from the masking difference between peaks and troughs of pink SIM noise with a modulation depth of 30 dB. Listeners were presented with octave-filtered clicks (2-Hz repetition rate) in peaks or troughs of the modulated noise filtered in the same octave as the clicks. Thresholds were determined by varying the click level in a Békésy procedure, and the masking difference

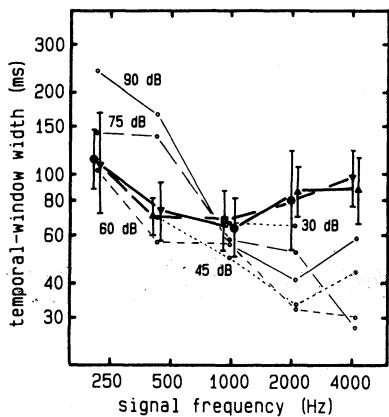


FIGURE 3. Width of the temporal window as a function of centre frequency, measured with octave-filtered masker and probe signal. Light lines are for normal-hearing listeners with masker level in dB/one-third-octave as a parameter (fully drawn, 90 dB; broken, 75 dB; dashed, 60 dB; small dash, 45 and 30 dB). Heavy lines and closed symbols show the results of the hearing-impaired listeners for 75 and 90 dB/one-third-octave with the standard deviation among listeners. The various symbols represent the number of listeners involved (squares, 9-10; circles, 7-8; upward pointing triangles, 5-6; downward pointing triangles, 3-4).

between peak and trough was measured at modulation frequencies of 2, 4, 6, 8, 12, and 16 Hz. An important precondition for these experiments is the absence of phase shift in the masking pattern of the SIM noise. This was verified beforehand by measuring the masking at eight positions in time of the envelope and studying both amplitude and phase characteristics as a function of modulation frequency. No phase shift was found at modulation frequencies at which the amplitude transfer function differed significantly from zero. From the peak-to-trough ratio as a function of modulation frequency the width of a temporal intensity-weighting function (temporal window) was calculated. When a temporal window with a fixed shape acts upon the masker envelope, the smearing of this envelope depends only on the width of the window. For each condition, determined by level and central frequency of the masker, the width is calculated for a Gaussian window which optimally fits the masking data. For normal-hearing listeners the temporal-window width as a function of masker frequency and level is shown in Fig. 3 with light lines. Eleven out of the twelve hearing-impaired listeners in the SRT experiment also participated in this experiment. They were tested for masker levels of 75 and 90 dB per one-third-octave, comparable to the masker levels in the SRT experiment, but not every listener was tested in all conditions. The average results are shown in Fig. 3, with for each data point the standard deviation among listeners and an indication of the number of listeners involved.

For normal-hearing listeners our data confirm the common finding that temporal resolution increases with frequency. However, this experiment shows a much larger time constant, and thus a much wider temporal window, than the time constants found in other studies with often a wide-band masker (cf. Viemeister, 1977). For low frequencies temporal resolution appears to be level dependent, with better resolution for lower masker levels. Temporal resolution for the hearing-impaired listeners is only slightly frequency dependent with an optimum resolution around 1000 Hz. In the two lower octave bands the hearing-impaired listeners show, at 75 and 90 dB per one-third-octave, a better resolution than in normal hearing. However, this is clearly an effect caused by the lower sensation level at which the signals are perceived by the hearing-impaired listeners. When both groups are compared at equal sensation levels, the normal-hearing listeners perform better for all centre frequencies. At higher frequencies there is no clear level effect for the normal-hearing listeners and here the hearing-impaired listeners have a wider window, irrespective of whether the comparison is made for equal sensation level or for equal stimulus level.

If it is true, that the SRT in a fluctuating noise is determined by the signal-to-noise ratio in the troughs, as is suggested in Fig. 2, then a better temporal resolution must lead to a lower SRT and vice versa. Under this plausible assumption, the very small profit that hearing-impaired listeners have from masker modulations, as shown in Fig. 1, must be due to the deterioration of their temporal resolution at the high frequencies, which was measured for comparable masker levels as used in the SRT experiment. The relation between temporal resolution at high frequencies and speech reception in noise was also demonstrated by Dreschler (1983). He measured gap detection at 500, 1000, and 2000 Hz as part of a battery of tests with hearing-impaired

listeners and found significant correlations with speech hearing loss in noise (D), which appear to be higher towards the higher frequencies.

REFERENCES

1. Dreschler, W.A. (1983). Relations between psychophysical data and speech perception for hearing-impaired subjects. Ph.D. thesis, Free University, Amsterdam, The Netherlands.
2. Duquesnoy, A.J. (1983). Effect of a single interfering noise or speech source upon the binaural sentence intelligibility of aged persons. J. Acoust. Soc. Am., 74, 739-743.
3. de Laat, J.A.P.M., and Plomp, R. (1983). The reception threshold of interrupted speech for hearing-impaired listeners. In: R. Klinke and R. Hartmann (Eds.), Hearing - Physiological Bases and Psychophysics, 359-362. Springer-Verlag, Berlin.
4. Elliott, L.L. (1971). Backward and forward masking. Audiology, 10, 65-67.
5. Plomp, R. (1964). Rate of decay of auditory sensation. J. Acoust. Soc. Am., 36, 277-282.
6. Plomp, R. (1978). Auditory handicap of hearing impairment and the limited benefit of hearing aids. J. Acoust. Soc. Am., 63, 533-549.
7. Viemeister, N.F. (1977). Temporal factors in audition: a system analysis approach. In: E.F. Evans and J.P. Wilson (Eds.), Psychophysics and Physiology of Hearing, 419-428. Academic Press London.

DIFFERENCES IN LISTENING STRATEGIES BETWEEN NORMAL AND HEARING-IMPAIRED LISTENERS*

Arjan Bosman and Guido F. Smoorenburg
Laboratory of Experimental Audiology, Department of
Otorhinolaryngology
University Hospital, Catharijnesingel 101
3511 GV Utrecht, The Netherlands

INTRODUCTION

The perception of words spoken in isolation is influenced by various factors. We studied the influence of 1) word type, with different degrees of redundancy, 2) style of articulation, and 3) presentation level. The study included a reference group of normal hearing (NH) subjects and a second group of hearing-impaired (HI) subjects, all usually wearing hearing aids. Comparison of the perception at phoneme and word level and of patterns of phoneme confusions reveals differences in listening strategies between the subjects of both groups.

METHODS

Three word lists were selected, which are commonly used in speech audiometry by different audiological centres in the Netherlands. The first list, A, consists of words of the consonant-vowel-consonant type (CVC words). The second list, B, consists of mono-syllables with a consonant cluster of 1 to 4 phonemes in initial position and a cluster of 0 to 3 consonants in final position. The third list, C, is phonetically balanced (i.e. the frequency of occurrence of the phonemes in the list corresponds to the phoneme frequencies in Dutch). List C consists of monosyllabic and bisyllabic words of 2 to 6 phonemes. All bisyllabic words contain an unstressed syllable with a schwa, which are mostly plural or infinitival forms. This results in a higher redundancy for the bisyllabic words than for the other word-types. Each list was uttered by two female speakers; speaker 1, S1, with normal articulation, and speaker 2, S2, with unnatural sounding, overprecise articulation.

The first group of subjects consisted of 24 normal-hearing students, the second group of 24, mostly older, hearing-impaired subjects. The HI subjects were selected on a maximum phoneme discrimination score in the range of 60 to 90% without using their hearing aids. In the present experiments we measured all HI subjects without hearing aids. Sixteen subjects suffered from presbyacusis with

*The measurements of the NH group were carried out by J. van Dijkhuizen, L. van Loon and M. Schelvis; the HI group was measured by A. Clemens. The authors thank prof. A. Cohen for his comments on an earlier draft of this paper. This research has been supported by the Netherlands Organisation for the Advancement of Pure Research (ZWO).

its characteristic high-frequency hearing loss; the other 8 subjects showed flat losses resulting from various auditory disorders.

For the NH group, presentation levels were 15, 20, 25, 30, and 35 dB SPL, for the HI group 50, 65, 80, 95, and 110 dB SPL. The six list-speaker combinations were presented to the subjects according to a counterbalanced design. The speech material was presented in a sound-treated room using a Madsen OB 822 audiometer and TDH 39 headphones with MX41/AR cushions.

Responses were noted down in phonemes and compiled into confusion matrices. These confusion-matrices were transformed into symmetric similarity matrices according to an algorithm suggested by Houtgast in an article by Klein et al. (1970):

$$s(i,j) = s(j,i) = 0.5 * \sum_{k=1}^N [c(i,k) + c(j,k) - |c(i,k) - c(j,k)|] \quad (1)$$

in which N is the number of stimuli, $c(i,j)$ are the elements of the confusion matrix, and $s(i,j)$ the elements of the similarity matrix.

PHONEME AND WORD SCORES

Fig. 1 shows the number of correct phonemes (phoneme score) and the number of words responded to completely correctly (word score) in each word list by, left panel, the NH group and, right panel, the HI group. The data were pooled across both speakers. The phoneme scores for list C are higher than those for list A, while list B yields the lowest scores. The high phoneme scores for list C are due to the higher redundancy in the bisyllabic words of this list; however, the word scores are not higher for this list, because word perception is governed by the non-redundant stem of the word. The type of word does not affect the word score for the HI group, whereas it does for the NH group. The latter result suggests that the normal hearing students make more use of lexical factors in word perception.

Fig. 2 shows the speaker-effect on the phoneme and word scores for both subject groups. The data were pooled across the three word lists. This figure shows that the overprecise articulation of S2 results in a somewhat higher intelligibility. The differences are greater for the NH group than for the HI group, especially at the higher levels. This will be explained in the sections below.

VOWEL CONFUSIONS

Only the vowels with a frequency of occurrence higher than 4% were used in the analysis of phoneme confusions. The schwa-phoneme was hardly ever confused with other vowels. This is probably due to the high redundancy in the bisyllabic words of list C. Therefore, the schwa was removed from the confusion matrices. The confusion matrices of the remaining 10 vowels were transformed into similarity matrices according to eq. (1). Subsequently, these matrices were subjected to the INDSCAL-algorithm of Carroll and Chang (1972), to study the effect of speaker and presentation level. With this algorithm the stimuli are represented as points in a so-called object space; the relative importance of the dimensions per stimulus condition follows from the weighting factors in the subject space.

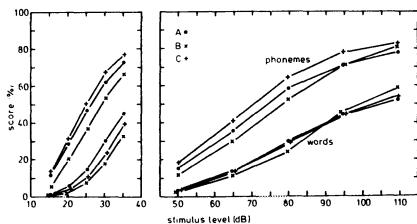


FIGURE 1, The effect of wordlist on the number of correct phonemes (phoneme score) and on the number of words responded completely correctly (word score) for each list. Scores for NH subjects are shown in the left panel, scores for HI subjects in the right panel. Data were pooled across speakers.

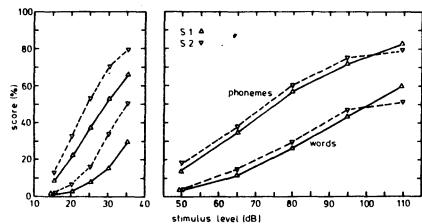


FIGURE 2, The effect of speaker on phoneme and word scores. Data were pooled across word lists. Left panel: NH subjects, right panel: HI subjects.

The results for the NH group are shown in Fig. 3. The first dimension, which explains 43% of the total variance, corresponds to the first formant F1. The second dimension, explaining 34% of the total variance, corresponds to the second formant F2. The /u/ was left out of consideration, because of its low frequency of occurrence. Therefore, one corner of the vowel triangle is missing.

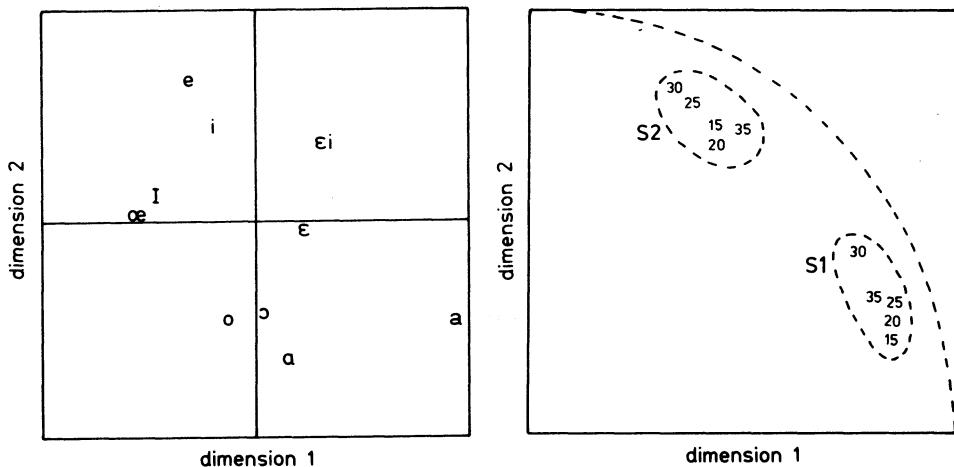


FIGURE 3, Object and subject space from a two-dimensional INDSCAL analysis of vowel confusion matrices for NH subjects. Data were pooled across wordlists. The subject space shows weighting factors for all combinations of speaker and stimulus level.

The configuration for the HI group is quite different from the configuration for the NH group. The first dimension corresponds again to the first formant F1, but in this case it accounts for as much as 65% of the variance in the data. The character of the second dimension, explaining only 14% of the variance in the data, is less outspoken. It does not represent the second formant. This can be explained by the audiograms of the HI subjects: most subjects (18) showed a predominantly high-frequency loss, causing an attenuation of the speech components especially in the range of the second formant. Further analysis showed, however, that vowel duration becomes an important acoustic factor. The number of confusions among short vowels is relatively high.

The subject space of Fig. 3 reveals a clustering of presentation levels for both speakers. In the NH group the speaker effect becomes apparent in a higher weighting of S2 on the F2-dimension. LPC-analysis revealed that the range of F2 was wider for S2, while the F1 ranges were comparable for both speakers. Thus, the NH group makes use of a wider range of F2. For the HI group we find a higher weighting of S1 on the dimension of vowel duration, while S2 articulated more emphatically. We conclude that for this speaker the emphasized articulation implies less useful information in vowel duration; the unnatural style of articulation meant 40% longer vowels and a considerably higher spread in vowel duration.

For the NH group the influence of presentation level on the relative importance of F1 and F2 is small. At lower presentation levels more of the extreme low- and high-frequency portions of the speech spectrum fall below threshold. This results in a higher number of confusions at the lower levels and thus in a shrinking of the perceptual vowel space. Since both F1 and F2 are affected, the relative importance of F1 and F2 remains almost independent of level. In the HI group, scores which are comparable with the scores of the NH group, are found across a wider range of stimulus levels. This implies that the low-frequency cue F1 exceeds the detection thresholds in most conditions. Perception of the second factor, vowel duration, does not depend on stimulus level. Thus, for the HI group we may expect little effect of stimulus level on the role of the different stimulus features in vowel perception. The experimental results were in agreement with this expectation.

CONSONANT CONFUSIONS

Two consonants, viz. /j/ and /ng/, which only appeared in list B, were discarded from the analysis because of their low frequency of occurrence (<1%). The first two dimensions of a 3-dimensional INDSCAL analysis are shown in Fig. 4 for the NH group and in Fig. 5 for the HI group.

For the NH group the three dimensions explain 50%, 18%, and 9% of the variance, respectively. The first dimension can be interpreted as voicing, as it differentiates between the voiceless plosives and fricatives /p,t,k,f,s,X/ and the voiced consonants. The second dimension can be interpreted as sonorance, as it separates the "cluster" /l,m,n/ from the other consonants. The third dimension is almost exclusively due to S2, who exaggerated the articulation of particularly the /r/ and the /s/. The subject space shows a weak level

effect for S1, whereas it is much stronger for S2. At the lowest stimulus level the weighting is about the same for both speakers, but at higher levels the weighting for S2 shifts from dimension 1 (voicing) towards the second and third dimensions. Apparently, in consonant perception the overprecise articulation of S2 has an effect only at the higher levels. At low levels voicing is all-important.

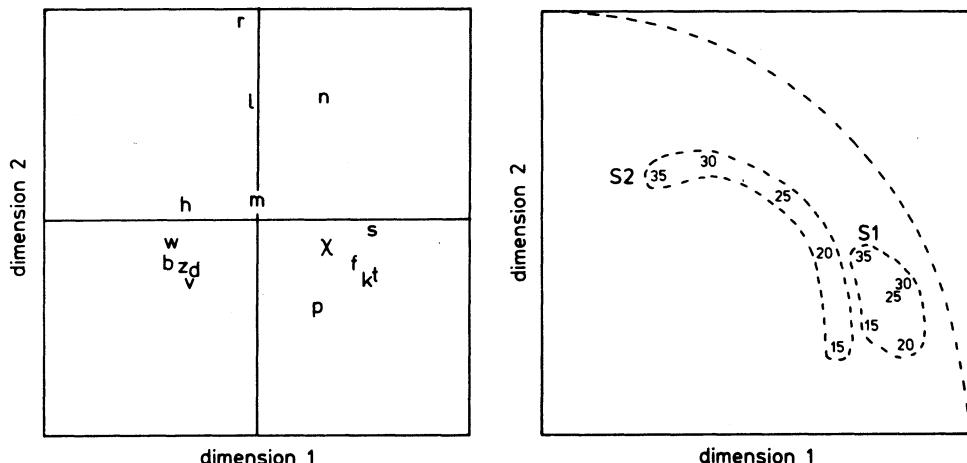


FIGURE 4, The first two dimensions of object and subject space of a three dimensional INDSCAL analysis of consonant confusions for NH subjects. Data were pooled across wordlists.

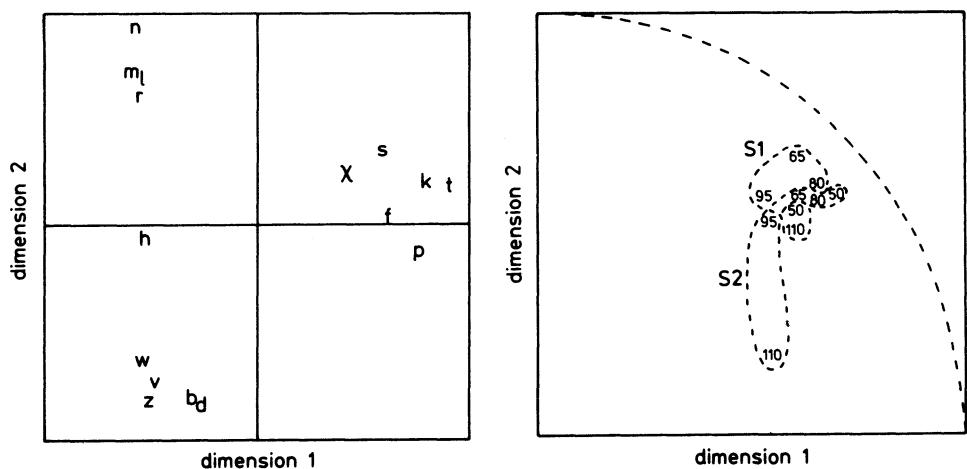


FIGURE 5, The first two dimensions of object and subject space of a three dimensional INDSCAL analysis of consonant confusions for HI subjects. Data were pooled across wordlists.

The configuration for the HI group is shown in Fig. 5. The three dimensions explain 38%, 27%, and 12% of the variance in the data, respectively. As for the NH group, the first dimension can be

interpreted as voicing and the second as sonorance. The third dimension can be interpreted as frication and is due to the lengthening of the fricatives by S2. For S1 the weighting on the dimensions 1 and 2 is about equal at all levels, whereas the weighting on dimension 3 is small. For S2 the weighting factors coincide with their positions for S1, except at the highest levels of 95 and 110 dB SPL, where they shift towards dimension 3 (friction). The size of the shift for the HI group is comparable to that for the NH group, but here there is a shift to frication.

The influence of word type was studied by carrying out separate KRUSKAL-analyses (1964a,b) on the data pooled across levels and speakers for each word list. Especially for the NH group, and to a lesser degree for the HI group, the configuration for list A shows a more outspoken clustering than the configuration for list C, with list B in between. Apparently, the higher redundancy in list C interferes with phoneme-by-phoneme perception.

DISCUSSION

The results show recurring patterns of phoneme confusions, particularly for CVC words. The patterns become less outspoken for more redundant words. Continued research shows a close resemblance of the patterns obtained with both sense and nonsense CVC words. For CVC words this indicates that lexical factors have little influence on phoneme confusions. Of course, phonological constraints as set by Dutch and coarticulation also play a role, but phoneme-by-phoneme perception is the most important factor.

Differences in listening strategies between NH and HI subjects become apparent for both vowel and consonant confusions. The vowel configurations show that NH subjects make use of F1 and F2 information, whereas HI subjects (without their hearing aids) use F1 and vowel duration. The vanished contribution of F2 for the HI subjects is due to the hearing losses in the F2-range. The first two dimensions of the consonant configuration correspond, for both subject groups, to voicing and sonorance. However, the clustering is more outspoken with the HI group. Apparently, HI subjects tend to confuse phonemes mostly with phonemes which lie in the same cluster and hardly with phonemes outside this cluster. So, phoneme categories are used more strictly by HI subjects.

REFERENCES

1. Carroll, J.D. and Chang, J.J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of the 'Eckart-Young' decomposition. *Psychometrika*, 35, 283-319.
2. Klein, W., Plomp, R., and Pols, L.C.W. (1970). Vowel spectra, vowel spaces, and vowel identification. *Journal of the Acoustical Society of America*, 48, 999-1009.
3. Kruskal, J.B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29, 1-27.
4. Kruskal, J.B. (1964b). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29: 28-42.

CRITICAL BANDS IN THE PERCEPTION OF SPEECH SIGNALS BY NORMAL AND SENSORINEURAL HEARING LOSS LISTENERS*

Robert D. Celmer
Acoustics Program and
Laboratory,
Department of Mechanical
Engineering,
University of Hartford,
West Hartford, CT, USA 06117

Gordon R. Bienvenue
Communications Department,
State University of New
York,
New Paltz, NY, USA 12561

INTRODUCTION

Existing auditory theory suggests a major role for critical bands. Scharf has defined the critical band empirically as "that bandwidth at which subjective responses rather abruptly change" (cf. Scharf, 1970, p. 159). In general, two stimuli separated in frequency by less than a critical bandwidth will interact in one of a number of ways, while two stimuli separated by more than a critical bandwidth will not. The critical band phenomenon has been observed in such perceptual phenomena as masking (Fletcher, 1940; Scharf, 1970), loudness (Zwicker, 1958), and musical consonance (Plomp, 1964).

Speech is the most pervasive and significant acoustic stimulus for the human listener, and evidence suggests that the critical band may serve in the analysis of speech (cf. Scharf, 1970). In several studies, bandwidths contributing equally to the perception of speech were approximately equal to critical bands found in pure tone psychoacoustic studies (cf. French and Steinberg, 1947; Richards and Archbald, 1956; Kryter, 1960; Castle, 1964; Chari, 1977).

The works of Fletcher (1940), Zwicker (1958), and Greenwood (1961) imply that the critical band serves to band-limit background noise. The narrower the passband of the ear as a filter, the more noise the ear can reject. Thus, a listener may be able to correctly perceive a spoken communication despite background noise simply because much of the energy associated with the noise lies outside the critical bands surrounding the formant frequencies of the speech.

Discrimination of the formant and harmonic content of both speech and non-speech signals requires that these components be separated by at least one critical band (Morton and Carpenter, 1963; Plomp, 1964; Haggard, 1974). Synthetic vowels presented to listeners by Remez (1977) showed an abrupt changeover from speech-like to non-speech-like sounds as the formant bandwidth increased to greater than a critical bandwidth.

Several researchers have reported evidence of distorted or widened critical bands in subjects with sensorineural hearing loss

*We wish to acknowledge the contributions of James Martin, Paul Michael and James Prout to this study.

(Michael, et al, 1976; Bienvenue, et al, 1977; Bennett, et al, 1978; Bonding, 1979; Bienvenue and Celmer, 1983). In addition, these data demonstrate that the width of the widened critical band is independent of the magnitude of threshold hearing loss amongst those sensorineurals with critical bandwidth distortion.

The purpose of the present study was to directly test the hypothesis that the critical band is an essential element in auditory speech discrimination.

METHODS AND MATERIALS

A. Subjects

Forty-eight normal hearing listeners, aged 19 to 31 years, and sixty-eight sensorineural hearing impaired listeners, aged 18 to 67 years, participated in the study. Each subject was classified using conventional audiometric techniques. The right ear was the test ear for all subjects.

In addition, subjects were presented with a loudness-of-complexes type of critical bandwidth test (cf. Scharf, 1970) in order to independently measure the subjects' critical bandwidth. The center frequencies of each tonal complex were located at 700 Hz, 1000 Hz, 1600 Hz, and 2150 Hz. For each trial a sub-critical tonal complex was presented to the listener and the bandwidth of the tonal complex was increased by small amounts. Ten repetitions of each center frequency were performed at a presentation level of 50 dB HL. Subjects were asked to listen to each test signal episode, and indicate the moment they perceived a change in the stimulus. The signal bandwidth required to elicit a perceptual change was recorded for each of the trials for each subject. Thus, the independent estimate of the critical bandwidth at each center frequency for each subject was based on the subjects' performance on ten trials.

B. Taped Stimulus Materials

The stimulus materials used in this study were a set of limited resolution bandwidth speech signals. Prerecorded stimuli (in analog form) were digitized by a 12-bit A/D converter at 20,000 samples per second, and stored in a computer. The signals were separated into 13 ms segments. The average spectrum of each segment was computed by first applying a Hamming window and then performing a Fast Fourier Transformation. The resolution bandwidth of this discrete spectrum is then limited to the frequency limits those recommended by Scharf (1970). The discrete frequency amplitudes that fell within each bandwidth were averaged; each of the discrete amplitudes of that band was then set equal to this r.m.s. value, limiting the resolution allowed to the preselected bandwidth. See Figure 1. Coarser and narrower filtering schemes were realized by multiplying the bandwidth limits used by a chosen factor (retaining the original center frequency), and averaging the amplitude contained within these widened limits. Each processed spectrum was then inverse transformed into the time domain, an inverse Hamming window is applied, and the output was converted to analog form via a D/A converter and recorded on audio tape.

The Nu#6 wordlist, a clinical audiometric word list, was used as the input audio material, since it includes (consonant-consonant-nucleus-consonant) sounds as opposed to only (consonant-consonant-vowel-consonant) sounds (cf. Tillman and Carhart, 1966).

The tapes generated had seven frequency resolutions: an unprocessed list (UP); a bandwidth equal to one-half the resolution of the normal critical bandwidth (HX) (cf. Scharf, 1970, Scharf and Hellman, 1980); a bandwidth equal to the resolution of the normal critical bandwidth (1x); two times normal (2x); three times normal (3x); five times normal (5x); and seven times normal (7x).

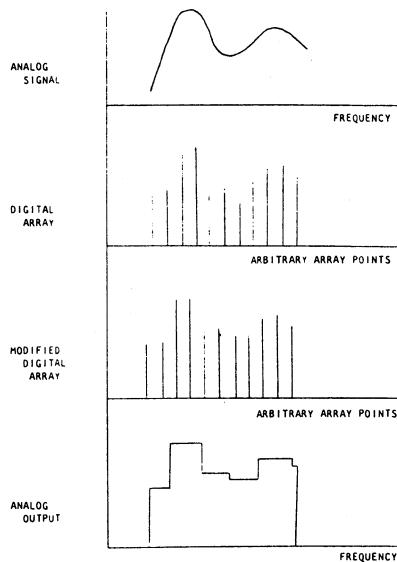


FIGURE 1, Effect of digital filtering.

C. Equipment

The processed signals described above were generated using a hybrid computer system comprised of a EAI (Electronic Associates Incorporated) Model 680 analog computer interfaced with a DEC (Digital Equipment Corporation) digital computer, Model PDP-10. The A/D and D/A conversions were both performed at a rate of 20,000 samples/second. The audio output was recorded via a Crown Model BP824 tape deck.

The discrimination tasks were performed using an Ampex AG-440 B tape recorder, a Maico Model MA-18 audiometer calibrated to ANSI 1969 standards, and TDH-39 earphones fitted with MX-41/AR cushions. The tests were performed in a Suttle Corporation Model B1 quiet room.

D. Test Procedure

After the audiometric threshold test and the independent critical bandwidth test, the subjects were presented with the seven fifty-word lists in the following resolution order: 7X, 5X, 3X, 2X, 1X, HX, UP. This sequence was chosen in order to minimize learning effects. The signal reached the earphone at a level of 50 dB HL and a signal-to-noise ratio of +10 dB. Pink noise was utilized as the masking source. Masking was used to minimize ceiling effects. Subjects were provided with answer sheets and were asked to write down the word they felt was said, guessing when necessary. Wordlists were scored on a percentage basis.

RESULTS

A. Normal Hearing Subjects

The normal hearing group's mean pure-tone thresholds ranged from 1.5 to 5.6 dB for the frequencies from 500 Hz to 4000 Hz. On the loudness-of-complexes test for independent critical bandwidth measurement, this group showed a mean critical bandwidth of 0.94 times the normal critical band as reported by Scharf; that is, 0.94X. The range of values was from 0.7X to 1.2X over the four test frequencies measured, confirming that this group did indeed have normal critical bands.

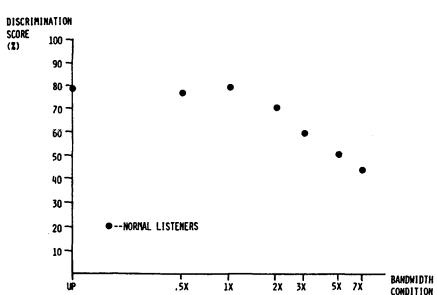


FIGURE 2, Plotted discrimination score means.

As may be seen from Figure 2, the discrimination score was a direct function of the log of bandwidth resolution, for those bandwidths wider than one critical band, but was independent of bandwidth for those bandwidth resolution conditions equal to or narrower than the critical band. The average discrimination score for the UP, HX, and 1X conditions was 77.56%. Thus, the masker did eliminate a ceiling effect.

A single factor analysis of variance indicated a significant main effect of speech processing. A Newman-Keuls follow-up test (cf. Glass and Stanley, 1970) demonstrated a significant decreasing trend for discrimination scores as the bandwidths were varied from 2X through 7X. However, no significant decreases in intelligibility scores were observed for the UP (unprocessed) through 1X conditions. In addition, the group scores at the 1X condition were significantly higher than those at the 2X condition.

B. Sensorineural Hearing Impaired

The sensorineural hearing impaired group's mean pure-tone threshold audiogram ranged from 22.4 dB at 250 Hz to 71.2 dB at 8000 Hz (See Table I). The mean speech reception threshold was 37.6 dB and the mean speech intelligibility in quiet at 50 dB HL was 74.7%. On the loudness-of-complexes test for independent critical bandwidth measurement, this group showed a range of widened critical bands, from 1.1X to 4.23X, with a mean bandwidth of 2.43X. The subjects fell into four critical bandwidth groups: a 1X group (range of bandwidths

The plotted means for processed speech intelligibility scores are presented in Figure 2. A regression line was computed for the 2X through 7X portion of the data. The bandwidth condition vs the group's speech intelligibility scores exhibited a significant correlation of $r = -.75$ with a probability of error less than 0.01.

1.1X to 1.5X), a 2X group (1.5X to 2.5X range), a 3X group (2.5X to 3.5X range), and a 4X group (3.5X to 4.23X range).

Table 1, Mean audiometric data for hearing impaired listeners

FREQUENCY (Hz)	250	500	1000	2000	4000	8000
THRESHOLD (dB)	22.4	29.3	38.3	60.2	65.6	71.2

The relationship between the critical bandwidth rating determined from the loudness-of-complexes test and the "knee" of the processed speech intelligibility data was of particular interest. This "knee" continued to represent the processing bandwidth at which the intelligibility scores dropped significantly from the unprocessed score. See Figure 3. These observations were made by performing a similar statistical analysis as used on the normal data.

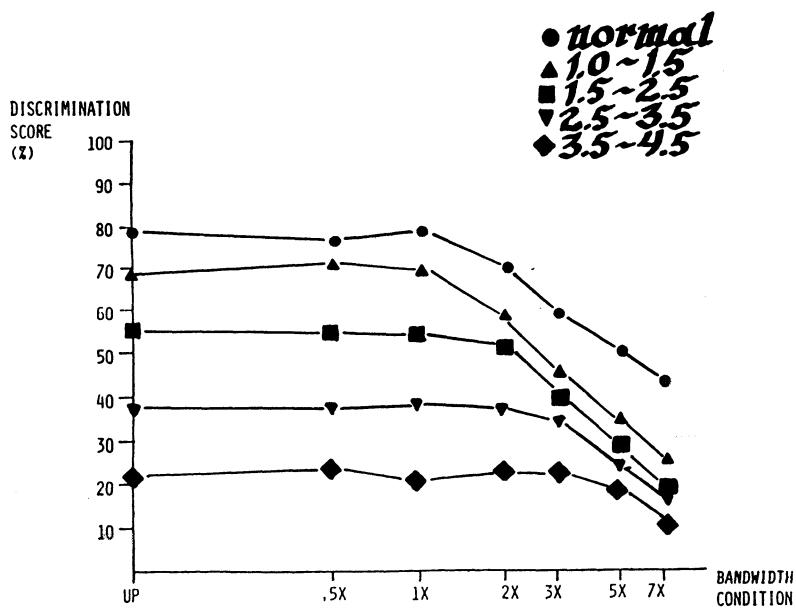


FIGURE 3, Plotted discrimination score means.

It is interesting to note that the mean speech intelligibility functions for the subjects in the present study show a "knee" that coincides with their critical bandwidth test. This is especially easy to observe when the data are grouped according to the result of

the loudness-of-complexes test. The correlation coefficient between the loudness-of-complexes bandwidth measurement and the "knee" of the speech intelligibility function was 0.8749. This is a high correlation and indicates a very strong relationship between tonal estimates of the critical band and speech intelligibility of sensorineurally hearing impaired listeners.

DISCUSSION AND CONCLUSIONS

As noted above, the peripheral auditory system has been described as performing a preliminary frequency analysis of incoming acoustic signals. The limit to which frequency information may be gated is called the critical band and has been observed in a variety of psychoacoustic contexts. In particular, the critical band mechanism performs noise-band limiting and harmonic discrimination, both of which are crucial for the correct perception of such complex acoustic stimuli as speech. Thus, it was hypothesized that the critical band is contributing factor to normal auditory speech intelligibility.

The purpose of the present study was to test the hypothesis that the critical band is an essential element in the process of speech listening. The performance of normal listeners with the processed speech test indicated three distinct trends:

- 1) A plateau effect was observed for the UP through 1X condition lists; that is, no significant differences among these scores were observed.
- 2) The 1X score was significantly higher than the 2X score.
- 3) A monotonic, decreasing trend was observed as the allowed bandwidth resolution was widened from the 2X through 7X.

The 2X through 7X scores demonstrated a close approximation to a logarithmic curve with a negative slope. The high correlation coefficient between these discrimination scores and their bandwidth conditions ($r = 0.75$) also underscored this observation.

The result of evaluating the loudness of complexes confirmed that the sampled normal listeners did indeed have normal critical bands, equal to about $0.94X$. Note that this mean bandwidth value was statistically equal to the widest processed speech list in which non-decremental performance was observed (i.e., the 1X list). In other words, this independent critical bandwidth rating corresponded to the observed point of inflection demonstrated by the speech intelligibility scores. Thus, the results for normal listeners support the notion that the critical band is a contributing factor to normal auditory speech intelligibility .

The results with the sensorineural hearing impaired group also demonstrated the same trends.

- 1) A plateau effect was noted for the processing conditions UP through the condition equal to a subject's independent critical band rating. For example, the group whose critical band was measured independently as 2X had a plateau effect occurring from UP to 2X.

2) The score of the highest processing condition in the plateau was significantly higher than the next higher processing condition score.

3) A monotonic, decreasing trend was observed as bandwidth resolution was widened beyond the "knee" in the curve.

Also significant was the correlation between the independent critical bandwidth rating and the location of the "knee" in the curve. This high correlation indicates a very strong relationship between tonal estimates of the critical band and speech intelligibility of sensorineural hearing impaired listeners. It is reasonable to conclude that the integrity of the critical band is an important factor in the understanding of speech signals.

In summary, a correlation has been demonstrated between the performance of listeners on an auditory speech intelligibility test and a test with tonal complexes as stimuli. The tonal complex test results yielded a bandwidth resolution value that was correlated to the probable point of inflection of the auditory speech intelligibility test results. Existing literature has implied, but has not clearly demonstrated the presence of such a correlation. The authors suggest that any consideration of auditory speech intelligibility among normal or sensorineural hearing impaired listeners must include an examination of the integrity of the critical band phenomenon in the subject population.

REFERENCES

1. Bennet, T., Bienvenue, G. Anthony, A., and Michael, P. (1978). Procedures for characterizing certain effects of prolonged noise exposure. Journal of the Acoustical Society of America, 63, Suppl. 1, S64.
2. Bienvenue, G. and Michael, P. (1977). The temporary effects of short term noise exposure on masking phenomenon. Unpublished EAL Research Project.
3. Bienvenue, G.R., Celmer, R.D., Prout, J.H., and Michael, P.L. (1980). Digital processing of speech materials in the study of sensorineural hearing impairment. Journal of the Acoustical Society of America, 67, Supplement 1, S60
4. Bonding, P. (1979). Critical bandwidth in presbycusis. Scandinavian Audiology, 8.
5. Boer de, E. (1961). Measurement of the critical bandwidth in cases of perception deafness. Proc. III Int. Cong. Acous., 1, 100-103, Elsevier, Amsterdam.
6. Boer de, E. and Bouwmeester, J., (1974). Clinical Psychophysics. Audiology, 14, 274-299.
7. Castle, W.E. (1964). Effects of selective narrow band filtering on the perception by normal listeners of Harvard PB-50, word lists. Journal of the Acoustical Society of America, 36, 1074.
8. Chari, N. (1977). Perception of 1/3 octave-filtered speech. Journal of the Acoustical Society of America, 36.
9. Fletcher, H. (1940). Auditory Patterns. Review Medical of Physiology, 12.

10. French, N. and Steinberg, J. (1947). Factors governing the intelligibility of speech sounds. Journal of the Acoustical Society of America, 19, 90-119.
11. Greenwood, D. (1961). Auditory masking and the critical band. Journal of the Acoustical Society of America, 33, 484-502.
12. Haggard, M. (1974). Feasibility of rapid critical bandwidth estimates. Journal of the Acoustical Society of America, 55, 304-308.
13. Kryter, K. (1960). Speech bandwidth compression through spectrum selection. Journal of the Acoustical Society of America, 32.
14. Michael, P. and Bienvenue, G. (1976). A procedure for the early detection of noise susceptible individuals. American Industrial Hygiene Association Journal.
15. Morton, J. and Carpenter A. (1963). Experiments relating to the perception of formants. Journal of the Acoustical Society of America, 35, 475-480.
16. Plomp, R. (1964). The ear as a frequency analyzer. Journal of the Acoustical Society of America, 36, 1628-1636.
17. Plomp, R. and Levelt, W. (1965). Tonal consonance and critical bandwidth. Journal of the Acoustical Society of America, 38, 548-560.
18. Rabiner, L. and Schafer, R. (1978). Digital processing of speech signals, ed. A. Oppenheim, (Englewood Cliffs: Prentice-Hall, 1978).
19. Remez, R. (1977). Adaptation of the category boundary between speech and nonspeech: A case against feature detectors. Haskins Laboratories: Status Report on Speech Research, SR-50, 151-167.
20. Richards, D. and Archbald, R. (1956). A development of the Collard principle of articulation calculation. Proceedings of the IEE, 103B.
21. Scharf, B. (1961). Complex sounds and critical bands. Psychological Bulletin, 58.
22. Scharf, B. (1970). Critical bands. Foundations of Modern Auditory Theory, ed. J. Tobias, (New York: Academic Press, 1970).
23. Tillman, T. and Carhart, R. (1966). An expanded test for speech discrimination utilizing CNC monosyllabic words (Northwestern University auditory test no. 6). Technical Report, SAM-TR-66-55, United States Air Force School of Aerospace Medicine, Aerospace Medical Division (AFSC), Brooks Air Force Base, Texas.
24. Zwicker, E. (1958). Über psychologische und methodische Grundlagen der Lautheit. Acustica, 8, Akust. Beihefte 1, 237-258.

PHASE AND THE HEARING-IMPAIRED*

Stuart Rosen

Department of Phonetics & Linguistics, University College London,
4 Stephenson Way, London NW1 2HE, England

INTRODUCTION

Although Helmholtz, on the basis of experiments with 8-component harmonic complexes of fundamental frequencies near 119 and 238 Hz, claimed to "have never experienced the slightest difference in the quality of tone" with changes in relative phase among the components (Helmholtz, 1954), more recent studies have modified his conclusions (e.g., Mathes and Miller, 1947; Goldstein, 1967). It is now apparent that the primary determinant of the perceptibility of a given phase change is the frequency spacing between the sound's constituent sinusoidal components. When relative phase changes are made in components that are "close enough" together, they are perceptible; when they are made to widely spaced components, they are not. Phase sensitivity is thus understood to reflect the failure of frequency resolution - only when a sound's constituent sinusoids interact (i.e., lie sufficiently within a single critical band, or auditory filter) will a phase change be detectable. (For a discussion of other factors, see Rosen, 1986).

Especially relevant for estimating the importance of phase on the perception of speech (in particular, for vowel-like sounds) are studies like those of Licklider (1957) and Schroeder (1959), who restricted their attention to harmonic complexes, noting that changes in timbre and pitch were readily produced by phase manipulations. What seemed to have been the final word along these lines was an impressive multi-dimensional scaling study by Plomp and Steeneken (1969), who concluded that the effect of phase on timbre (in the limited sense of the perceptual attribute which distinguishes periodic sounds of identical pitch and loudness), although real, was small compared to the effect of the relative amplitude of the components.

All these studies, though, used normal listeners. There has been almost no investigation of the role of phase in determining the percepts of the hearing-impaired (the notable exceptions being Hoekstra and Ritsma [1977] and Hoekstra [1979]). Given that phase sensitivity is supposed to be constrained by auditory frequency selectivity, and that many impaired listeners have impaired selectivity,

* This work has been supported by the Medical Research Council of the U.K. Many thanks to V. Ball, C. Bootle, C.M.Green, V. Hazan, and H. Wall for their extensive participation as listeners, and P. Howell for proofreading.

it seems likely that phase will play a larger role for them than for normal-hearing listeners (Rosen, 1984; Rosen and Fourcin, 1986).

METHODS

Test stimuli were synthesized digitally by a DEC PDP-12 computer running a 10-bit DAC at a sampling frequency of 10 kHz. The phase and amplitude of each stimulus component was corrected (except where noted) for the phase and amplitude distortion produced by the headphones (a Connevans CE8, chosen for its relatively low nonlinear distortion at low frequencies and high levels). This was determined with a small electret microphone mounted on the grid protecting the headphone diaphragm, thus allowing monitoring of the sound pressure while the listener wears the headphones (Dominitz, 1975; Rosen and Nevard, *in press*). Preliminary measurements on a KEMAR manikin indicate that the sound pressure measured by the headphone-mounted microphone will be within 6 dB and 10° of that at the listener's tympanic membrane at 1.8 kHz (the maximum frequency in the following studies), with improving accuracy as frequency is lowered. The same phase and amplitude corrections (a mean of 8 ears) were applied for all listeners. The headphone output was intermittently monitored, using a real-time spectrum analyzer, to set levels and check the waveform. Typically, amplitudes were within ± 1.5 dB, and phase within -5 to +10° of those specified.

All sounds had a steady-state duration of 400 ms, with 50-55 ms raised-cosine rises and decays added. They were presented, after low-pass filtering and amplification, to a single earphone in a sound-treated room. Spurious spectral components in the sounds, measured at the headphones, were at least 40 dB down from the smallest component of the complex. Masking noise, when present, was band-pass (20 Hz to 2-3 kHz) at about 30-35 dB SPL/Hz.

A 3-interval 3-alternative forced-choice (3I-3AFC) task was used for testing the discrimination of phase shifts, while a 2I-2AFC task was used for assessing abilities to discriminate changes in fundamental frequency. Inter-stimulus intervals were about 580 ms. Feedback as to the correctness of response was given. During a particular session, the two sounds whose discriminability was being tested remained constant.

At the start of each 30-trial session, listeners were given the opportunity of unlimited practice with the pair of sounds to be tested. At this time, they effectively controlled the presentation of the stimuli. The experimenter, too, often trained the listeners using this facility. This initial practice was crucial, especially when phase discrimination was being tested. It was frequently reported in such tests that all three sounds were identical at the start of practice, but a difference could often be found after listening for a little while.

HYPER-SENSITIVITY TO PHASE CHANGES IN THE HEARING-IMPAIRED

Rosen (1984) argued that if phase sensitivity reflects the failure of frequency resolution, then at least some impaired listeners (with widened auditory filters and relatively intact temporal analyzing capabilities) should be, under appropriate circumstances, more sensitive

to phase changes than normal listeners. This possibility was tested using stimuli that have seen extensive use in phase perception studies, so-called 100%-SAM (for sinusoidally amplitude modulated) and QFM (for quasi-frequency modulated) sounds (Mathes and Miller, 1947). Both sounds have an identical amplitude spectrum: a central sinusoidal component, and two sinusoidal side-bands, 6 dB lower in amplitude than, and equally-spaced in linear frequency from (by an amount given by the modulation rate) the central component. They differ only in their phase spectrum, a 90° change in the central component.

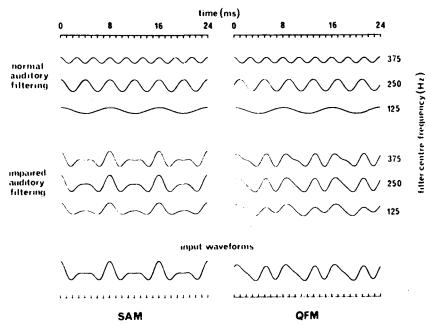


FIGURE 1, The outputs of hypothetical normal and impaired auditory filter banks to two three-component harmonic complexes which differ only in the relative phase of their central component (SAM and QFM sounds with a carrier frequency of 250 Hz and a modulation rate of 125 Hz). The auditory filters are centered at the frequencies of the harmonic components. Normal auditory filtering uses the rounded exponential model and bandwidths given by Moore and Glasberg (1983), while impaired auditory filtering assumes the same trend of bandwidth with frequency, but with absolute values ten times larger than in the normal case. From Rosen and Fourcin (1986).

Figure 1 shows graphically why impaired auditory filtering might lead to better discrimination of SAM from QFM sounds. With normal auditory filtering (when the spectral components for this particular sound are essentially resolved), in order to distinguish SAM from QFM sounds there needs to be some way for comparing the time of events across auditory filters (which available evidence suggests is not possible). In an impaired auditory filter bank, the same phase change is expressed as a within channel change.

Figure 2 shows an instance in which an impaired listener (XG) did, indeed, evidence increased sensitivity to phase changes in a SAM/QFM discrimination task centered at 400 Hz. XG was a young (late twenties), successful hearing-aid user with a relatively flat loss of 30-50 dB across the frequency range 0.125-8 kHz. Her degree of frequency selectivity was assessed at 500 Hz using the "notched-noise" technique of Patterson et al. (1982). The difference between the threshold of a 500 Hz tone obtained in a broadband noise, and one with a "notch" in its spectrum (400 Hz wide) centered linearly on the tone frequency, was determined. The bigger this difference in thresholds is, the narrower are the auditory filters. For the noise level of 60 dB SPL/Hz used, normal listeners obtain about a 20 dB

change between the two conditions (Rosen and Stock, in preparation). XG showed only a 10 dB change in threshold, with an estimated equivalent rectangular bandwidth (ERB) approximately twice that of normal listeners. She also showed better than normal performance for SAM/QFM sounds centered at 500 Hz.

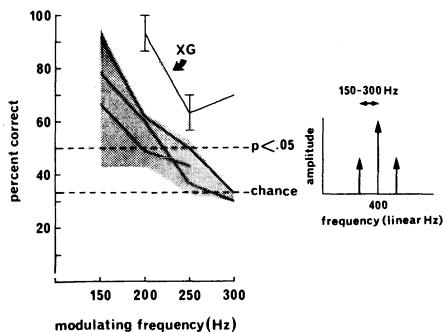


FIGURE 2, The performance of three normal-hearing listeners compared to that of a single hearing-impaired listener in a SAM/QFM discrimination task, all with approximately equal exposure to the task. The carrier frequency of the stimuli was always 400 Hz, with modulation frequency varying from 150 to 300 Hz. The level of the carrier was equal for all listeners, but varied somewhat (from about 93-98 dB SPL) with modulation rate. The inset at right shows the amplitude spectrum for the stimuli.

Although the amplitude of the three stimulus components was adjusted to give the correct amplitude relationships at the output of the headphones, all stimuli were in sine phase at their input. Thus the true phase relationships among the sounds varied with modulation rate, although the QFM stimulus always differed from the SAM stimulus by a + 90° phase shift at 400 Hz. Inspection of the output of the headphones to the SAM stimulus complex of 200, 400 and 600 Hz showed that the three components were, coincidentally, very nearly in -cosine phase. The shaded area shows the range of performances obtained across all the normal listeners, while the solid lines within the shaded area show their individual results, averaged across sessions. Note how overall performance decreases with increasing modulation rate. The solid line at the top of the figure shows the mean results obtained from a hearing-impaired listener, while the bars show the range of performance that was exhibited. XG was shown to have a loss of frequency selectivity at 500 Hz.

That the degradation in frequency selectivity is the important factor in accounting for this "hyper-sensitivity" is supported by results from another hearing-impaired listener. His audiogram shows a loss sloping from 15 dB at 125 Hz to 40-50 dB at 1-8 kHz. Even with a loss of 35 dB at 500 Hz, he shows normal frequency selectivity there, and is also within the normal range for detecting phase changes in stimulus complexes centered at that frequency. At 1 kHz, however, where his selectivity is degraded (the ERB about 50% bigger than normal), he is better than any normal listener tested at distinguishing SAM from QFM at modulation rates from 600 to 800 Hz. At the same frequency, he is able, under certain circumstances, to distinguish complexes which are added together in cosine phase from those added together in sine phase (Rosen, 1986). Normal listeners were unable to perform this particular discrimination.

Of course, not every hearing-impaired listener will show better discrimination performance than normal listeners. For one thing, the detection of phase changes clearly relies on sufficiently good temporal analyzing ability. Although an impaired listener may theoretically gain an edge by broadened auditory filters, s/he may just as well lose it through impaired temporal processing. What is striking, though, is that only very rarely are impaired listeners, even when they are impaired to a profound degree, much inferior to normal listeners in this task. As the impaired often become less sensitive to changes in the amplitude spectrum, the relative role of phase is still likely to be greater than that found, for example, by Plomp and Steeneken (1969) for normal listeners, even if they are not more acute in absolute terms.

THE PERCEPTUAL IMPORTANCE OF INCREASED SENSITIVITY TO PHASE

Given that the frequently broadened auditory filters of the hearing-impaired will allow a greater interaction between spectral components, and hence a greater role for phase, there are likely to be two main ways in which this will influence the perception of speech.

Firstly, in so far as temporal information is important in the perception of spectral shape, as has been proposed, for example, by Sachs and Young (1979) and Young and Sachs (1979), impaired listeners will hear changes in phase as changes in vowel quality. In fact, many listeners, both normal and impaired, report phase changes in harmonic complexes as changes in vowel quality. Darwin and Gardner (1986) have shown changes in vowel labelling performance with changes in phase in normal listeners, and those effects are likely to be stronger in impaired listeners.

Secondly, in so far as pitch perception relies on a temporal analysis of waveforms after a preliminary frequency analysis (as in the models of Moore and Glasberg [1986] and van Noorden [1982]), the perception of voice pitch is likely to depend on the relative phases of the constituent components of the sounds in a much stronger way than is found in normal listeners. Such effects have been shown by Hoekstra and Ritsma (1977) and Hoekstra (1979), albeit for sounds that are only remotely related to speech. They used SAM and QFM complexes with a centre frequency of 2 kHz and modulation rates near 200 Hz. Instead of requiring listeners to discriminate between SAM and QFM sounds at the same modulating frequency (as in the experiments reported in the previous section), they were asked to discriminate changes in modulation rate with sounds that were both SAM or QFM. This is roughly equivalent to perceiving changes in the fundamental frequency of a speech sound from three upper harmonics (the so-called "residue"). Hoekstra (1979) reported that three of five hearing-impaired listeners were significantly better at discriminating changes in modulating frequency for the SAM complex ("in phase" components), than for the QFM complex. Normal listeners (and the other two impaired listeners) showed no difference between the two conditions. It seems likely that in the impaired listeners who showed this difference, widened auditory filters allowed spectral components to interact, thus affecting the waveform presented to the temporal analyzers by the auditory filters. We might well suppose that "in phase" spectral components would reflect the modulating frequency of

the signal in a clearer way than "out of phase" components (figure 1).

A similar result has been obtained with rather more speechlike sounds. Rosen and Fourcin (1983) extensively investigated the auditory capabilities of one profoundly impaired listener who lost his hearing in his mid-forties as the result of a skull fracture from a fall. One ear was made totally deaf, while the other had a so-called "left-hand corner" audiogram (70 dB HL at 125 Hz, falling off to 115 dB HL at 1 and 2 kHz, with thresholds greater than 120 dB HL for 4 and 8 kHz). They found that his discrimination of changes in fundamental frequency in the voice frequency range was better when the stimuli were sinusoids, than when they were pulse trains or speech. This is the opposite pattern to that found in normal listeners, where discrimination of fundamental frequency in sounds with multiple harmonic components is generally better than that found for sinusoids at the fundamental (e.g., Henning and Grosberg, 1968). Rosen and Fourcin argued that the almost certainly impaired frequency selectivity of the listener (difficult to measure in such a profound loss), was allowing harmonic components to interact to a much greater extent than in the normal case. Thus the temporal analyzers were presented with more complex waveforms in the case of pulse trains than in the case of sinusoids, and this was the cause of the worsened discriminability.

Confirmation of this hypothesis is found in further studies of the same patient, who was tested on his ability to discriminate changes in fundamental frequency for three-harmonic complexes with fundamental frequencies near 240 Hz. The fundamentals of the two stimuli to be compared were fixed at 228.6 and 252 Hz (a change of 10.25%), and could be either SAM or QFM (equivalent to what would be obtained by a modulation frequency of exactly half the carrier frequency). The phase relationships of all the stimuli within a session were always the same (i.e., appropriate either for SAM or QFM). Here, a 2I-2AFC task was used, in which the listener was required to label the direction of the pitch change. To prevent the use of any possible loudness differences between the stimuli (which seems unlikely, anyway), each sound was jittered by a different random amount each presentation, over a range of ± 2 dB. Table 1 shows that the change in fundamental frequency between the two stimuli is more salient for the complexes in SAM phase. Note though that phase corrections were not applied to these sounds. At the input to the headphones, all components were in sine phase, but this led to SAM complexes that were in approximate -cosine phase.

Table 1. Percent correct in discriminating a fixed change in fundamental frequency in a three-component harmonic complex when the relative phases of the components in the pair of stimuli to be compared are varied.

phase relationship		statistical significance of the difference
SAM	QFM	
65.6% (of 90)	47.8% (of 90)	p<0.05

"In phase" components do not always lead to more salient

pitches. Table 2 shows the results in a similar task by the same listener. The sounds are now four-component harmonic complexes with all components at the same amplitude. The fundamental frequencies of the two stimuli to be compared again differed by 10.25% (238.1 and 262.5 Hz), and the components could be either in sine phase (at the headphone output, as phase corrections were applied), or with alternating sine and cosine terms. No amplitude jitter was present. These results show that the detailed temporal structure of the sound needs to be taken into account, along with the phase distortions imposed by the listener's auditory system.

Table 2, Percent correct in discriminating a fixed change in fundamental frequency in a four-component harmonic complex when the relative phases of the components in the pair of stimuli to be compared are varied.

phase relationship		statistical significance of the difference
sine	alternating	
60.0% (of 110)	100.0% (of 110)	p<0.001

FINAL REMARKS

In considering the perceptual consequences of hearing impairment, we tend to assume that abilities to make perceptual distinctions, if not uniformly degraded, certainly do not become any better. Phase may be one instance in which a feature that is of relatively little consequence in determining the percepts of normal listeners (although possibly of great use in investigating temporally-based models of spectral feature extraction), becomes much more important in the hearing impaired. For example, increased sensitivity to phase may well be part of the reason why impaired listeners are often disturbed by reverberation (or, as one profoundly impaired listener put it to me: "The echoes, to people like us, are a bit disconcerting"), even though the randomization of phase relationships caused by reverberation has little perceptual effect in normal listeners (Plomp and Steeneken, 1973).

Not all phase effects may be negative. As impaired listeners can hear changes in phase as changes in vowel quality, even when they are poor at distinguishing vowels on the basis of their amplitude spectra, phase manipulations could provide a way of signalling useful information. The problem, though, is that if phase changes also affect the saliency (and indeed even the perceived value) of the voice pitch, how is one to independently manipulate the two perceptual features?

REFERENCES

1. Darwin, C.J. and Gardner, R.B. (1986). Mistuning a harmonic of a vowel: grouping and phase effects on vowel quality. *J. Acoust. Soc. Am.*, **79**, 838-845.
2. Dominitz, R.H. (1975). Headphone monitoring system for binaural experiments below 1 kHz, *J. Acoust. Soc. Am.*, **58**, 510-511.
3. Goldstein, J.L. (1967). Auditory spectral filtering and monaural phase perception, *J. Acoust. Soc. Am.*, **41**, 458-479.

4. Helmholtz, H. (1954). On the Sensations of Tone. New York.
5. Henning, G.B. and Grosberg, S.L. (1967). Effect of harmonic components on frequency discrimination. J. Acoust. Soc. Am., 44, 1386-1389.
6. Hoekstra, A. (1979). Frequency discrimination and frequency analysis in hearing. Ph.D. Thesis, University of Groningen.
7. Hoekstra, A. and Ritsma, R.J. (1977). Perceptive hearing loss and frequency selectivity. In: E.F. Evans and J.P. Wilson (Eds.), Psychophysics and Physiology of Hearing. Academic, London.
8. Licklider, J.C.R. (1957). Effects of changes in the phase pattern upon the sound of a 16-harmonic tone. J. Acoust. Soc. Am., 29, 780 (abstract).
9. Mathes, R.C. and Miller, R.L. (1947). Phase effects in monaural perception. J. Acoust. Soc. Am., 19, 780-797.
10. Moore, B.C.J. and Glasberg, B.R. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns, J. Acoust. Soc. Am., 74, 750-753.
11. Moore, B.C.J. and Glasberg, B.R. (1986). The role of frequency selectivity in the perception of loudness, pitch and time. In: B.C.J. Moore (Ed.), Frequency Selectivity in Hearing. Academic, London.
12. Noorden, L. van (1982). Two channel pitch perception. In: M. Clynes (Ed.), Music, Mind and Brain. Plenum, New York.
13. Patterson, R.D., Nimmo-Smith, I., Weber, D.L., and Milroy, R. (1982). The deterioration of hearing with age: Frequency selectivity, the critical ratio, the audiogram, and speech threshold. J. Acoust. Soc. Am., 72, 1788-1803.
14. Plomp, R. and Steeneken, H.J.M. (1969). Effect of phase on timbre of complex tones. J. Acoust. Soc. Am., 46, 409-421.
15. Plomp, R. and Steeneken, H.J.M. (1973). Place dependence of timbre in reverberant sound fields. Acustica, 28, 50-58.
16. Rosen, S. (1986). Monaural phase sensitivity: Frequency selectivity and temporal processes. In: B.C.J. Moore and R.D. Patterson, (Eds.) Auditory Frequency Selectivity. Plenum, New York.
17. Rosen, S. (1984). Hyperacute monaural phase sensitivity in the hearing-impaired, Brit. J. Audiol., 18, 257-258 (abstract).
18. Rosen, S. and Fourcin, A.J. (1986). Frequency selectivity and the perception of speech. In: B.C.J. Moore (Ed.), Frequency Selectivity in Hearing. Academic, London.
19. Rosen, S. and Fourcin, A.J. (1983). When less is more further work. Speech, Hearing and Language: Work in Progress, 1, 3-27 (Phonetics and Linguistics, University College London).
20. Rosen, S. and Nevard, S. (in press). A headphone monitoring system for low-frequency psychoacoustics. Brit. J. Audiol., (abstract).
21. Rosen, S. and Stock, D. (in preparation). Auditory filter bandwidths as a function of level at low (125 Hz-1 kHz) frequencies.
22. Sachs, M.B. and Young, E.D. (1979). Encoding of steady-state vowels in the auditory nerve: representation in terms of discharge rate. J. Acoust. Soc. Am., 66, 470-479.
23. Schroeder, M.R. (1959). New results concerning monaural phase sensitivity. J. Acoust. Soc. Am., 34, 1579 (abstract).
24. Young, E.D. and Sachs, M.B. (1979). Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers. J. Acoust. Soc. Am., 66, 1381-1403.