Explaining Phonetic Variation: A Sketch of the H&H Theory

Book · January 1990

DOI: 10.1007/978-94-009-2037-8_16

CITATIONS
359

READS
3,912

3 authors, including:

Björn Lindblom
Stockholm University
167 PUBLICATIONS **9,015** CITATIONS

SEE PROFILE

Alain Marchal
Centre Hospitalier Territorial de Nouvelle-Calédonie
19 PUBLICATIONS **497** CITATIONS

SEE PROFILE

EXPLAINING PHONETIC VARIATION: A SKETCH OF THE H&H THEORY

B LINDBLOM
Departments of Linguistics
University of Texas at Austin
Austin 78712-1196 Texas and
University of Stockholm
Stockholm S-10691 Sweden

ABSTRACT. The H&H theory is developed from evidence showing that speaking and listening are shaped by biologically general processes. Speech production is adaptive. Speakers can, and typically do, tune their performance according to communicative and situational demands, controlling the interplay between production-oriented factors on the one hand, and output-oriented constraints on the other. For the ideal speaker, H&H claims that such adaptations reflect his tacit awareness of the listener's access to sources of information independent of the signal and his judgement of the short-term demands for explicit signal information. Hence speakers are expected to vary their output along a continuum of **hyper-** and **hypospeech**. The theory suggests that the lack of invariance that speech signals commonly exhibit (Perkell and Klatt 1986) is a direct consequence of this adaptive organization (cf MacNeilage 1970). Accordingly, in the H&H program the quest for phonetic invariance is replaced by another research task: Explicating the notion of **sufficient discriminability** and defining the class of speech signals that meet that criterion.

1. Introduction

1.1 EXPLAINING PHONETIC VARIATION

One of the classical topics in phonetic theory is speech sounds in context. The modifications that phonetic segments undergo in the speech of a single individual are known to be extensive and to give rise to what is called the invariance problem: The difficulty of giving a physical phonetic definition of a given linguistic category that is constant and always free of context.

W. J. Hardcastle and A. Marchal (eds.), Speech Production and Speech Modelling, 403–439.

The main theme of the present chapter is: How do we make sense of this pervasive intra-speaker phonetic variation? My account is not new. It uses many well-known facts and traditional ideas. It may nevertheless differ in significant ways from answers currently obtainable from competing frameworks such as the Motor Theory, the Quantal Theory, the Direct Realism and Action Theory approaches, and 'Icebergs' (Fujimura 1989) just to mention a few (see Section 5.1).1)

The ultimate goal of the H&H program is to work out a quantitative and testable theory. At present I shall offer no more than a verbal sketch with a few hints about quantification. For short, I shall refer to it as the H&H theory, H&H referring to 'hyper'- and 'hypo'-articulation.

1.2 THE H&H ARGUMENT

To anticipate where the present review will lead us, I shall state the main argument of the paper right away.

We start out from two basic observations about perception and two about production:

a. Speech perception involves **discrimination** among items stored in the listener's lexicon. Lexical access is thus a function of the distinctiveness (rather than invariance) of the acoustic stimulus.

b. The process of discrimination is facilitated by processes not in the signal and whose contributions show short-term variations. Accordingly lexical access is assumed to be driven also by 'knowledge', that is by **signal-complementary processes.**

c. Speech motor control is future-oriented. It is purpose-driven and prospectively organized. Key words: plasticity and **output-oriented** control. Cf 'hyperspeech'.

d. As the output constraints on a movement become less severe, it tends to default to some low-cost form of behavior. Key words: economy and **system-oriented** control. Cf 'hypospeech'.

Combining those four observations we deduce that:

---

1) The H&H theory has grown out of my collaboration with Peter MacNeilage and Michael Studdert-Kennedy (Lindblom and MacNeilage 1986, Lindblom, MacNeilage and Studdert-Kennedy in preparation). Since language and speech are products of evolution, a minimum requirement on any theorizing in our area is that it be evolutionarily plausible. That means that theories of speech processes and sound patterns must be consistent with current Neo-Darwinian models of biological and cultural variation and selection. In line with such goals, H&H theory makes an attempt to address the systematic nature of intra-speaker phonetic variation.

e. The amount of explicit signal information minimally required for successful lexical access will vary between and within utterances. Cf the predictability of the word "nine" in the following examples (Lieberman 1963):
(i) The next word is _____ .
(ii) A stitch in time saves _____ .
f. In the **ideal case**, the speaker estimates the running contribution that signal-complementary processes will make during the course of an utterance, and dynamically tunes the production of its elements to the short-term demands for either output-oriented control (hyperspeech) or system-oriented control (hypospeech). What he/she needs to control is - not that linguistic units are actualized in terms of **physical invariants** (higher-order or whatever) - but that their signal attributes possess **sufficient contrast**, that is discriminative power that is sufficient for lexical access.
These considerations suggest the H&H working hypothesis:
g. During speech development and adult language use, real speakers adapt to the above-mentioned conditions. Speakers develop a 'feel' for the 'survival value' of phonetic forms through a process not unlike **natural selection**. Hence their speech will tend to vary systematically along an H&H dimension.

2. What Do Listeners Do?

what info do we get from the signal, and at which step/level does the disambiguation (which leads to correct lexical access) happen?

2.1 SIGNAL-DERIVED INFORMATION

In this section a short, selective review will be presented of research on the initial stages of speech signal processing.

A recent theme issue of the Journal of Phonetics (Greenberg 1988) provides a useful overview of what is currently known about the auditory representation of speech signals. It exemplifies the portrayal of signals up to the level of the ventral cochlear nucleus, a projection site of the auditory nerve. It appears that "the cochlea acts as a faithful translator of incoming waveforms" (Geisler 1988:34). However, certain stimulus features are enhanced. And there is evidence for onset detectors, fundamental frequency tracking mechanisms and automatic-gain-control properties of the transduction from the inner hair cells to nerve fibers. Moreover, frequency analysis is non-linear in a way that makes spectral information resistant to noise. With particular interest we note that at these early stages of auditory processing there does not yet seem to be any sorting of stimuli into phonetic categories (Geisler, p 34), no processing in any way special to speech.

Great importance is generally attributed to the **dynamic aspects** of speech signals (Bladon 1985, Remez, Rubin, Pisoni and Carrell 1981, Strange 1989). Note though the experimental work indicating that, under certain conditions, dynamic events may cause blurring rather than enhancement of stimulus cues (Lacerda 1987). Despite widespread agreement on the role of dynamic attributes, the fact is that we do not yet have a great deal of information on precisely what the relevant dynamic aspects might be. Few speech perception models and speech recognition algorithms make explicit proposals about specific dynamic parameters. Klatt (1987) mentions a study by Rabiner (1987) who found that a smoothed estimate of spectral change had a dramatic effet on the performance of many speech recognition systems. Error scores were approximately halved.

A mechanism that derives invariants from the signal was recently presented by Miller (1989) who proposes an F0-dependent version of the so-called **formant-ratio-theory** to account for cross-speaker vowel quality constancy effects. Invariant quality is a function of F0 and fixed logarithmic distances between formants.

Fowler (1986:4) states that "...perception must be direct and, in particular, unmediated by cognitive processes of inference or hypothesis testing, which introduce the possibility of error." This view reflects her attempt to apply the program of **direct perception** (Gibson 1972, 1979) to speech. The phenomenon of cross-speaker vowel color constancy that Miller addresses appears compatible with a direct realist account. Leaving complications such as "non-uniform scaling" (Fant 1973) aside for the moment, we could see Miller's model as an explicit specification of a mechanism that extracts what Gibsonians would call higher-order invariants directly from the signal.

Mechanisms of a similar sort form part of the **feature-based lexical access** scheme proposed by Stevens (1986). Here relationally defined stimulus attributes, offering more stability and invariance than the raw spectral data, are assumed to be extracted from the signal.

The phenomenon of durational contrast - invoked by Diehl and co-workers in developing a **theory of auditory enhancement** (Diehl and Kluender 1989; Diehl, Kluender, Walsh and Parker in press) - highlights an interesting aspect of auditory processing. In one of their experiments they address the widely observed interplay between closure duration and preceding vowel duration as a cue for signaling voicing contrasts in medial stops (Kluender, Diehl and Wright 1988). Two series of VCV-stimuli were used each varying only in the duration of the first vowel: The first series was derived by editing a natural /apa/. It ranged perceptually from /aba/ to /apa/. The second series

replaced the vowels by square-wave segments. Although their temporal and amplitude-envelope characteristics were those of the speech stimuli, these stimuli were not at all speech-like. For both series the task was to classify a given stimulus as one of the end-point stimuli. When plotted as a function of the duration of the medial silent interval, the results showed that a long initial vowel duration produced more /b/ responses than the short initial vowel condition. Significantly, a long initial square-wave segment analogously favored more short-gap responses. In other words, the longer initial segment durations made the silent gaps sound shorter in both cases. The authors see their parallel speech and non-speech results as arising from durational contrast, a general signal-processing phenomenon, and suggest that this effect explains the voicing-dependent vowel duration effect traditionally attributed to production factors.

In a recent formulation of the **Motor Theory** (MT), Liberman and Mattingly (1985:7) state: " ... adaptations of the motor system for controlling the organs of the vocal tract took precedence in the evolution of speech. These adaptations made it possible, not only to produce phonetic gestures, but also to coarticulate them so that they could be produced rapidly. A perceiving system specialized to take account of the complex acoustic consequences, developed concomitantly." How this specialized mechanism works remains to be specified. Nevertheless, the current version of MT assumes that phonetic invariance is gestural and counts on its being derivable from the signal: "...the invariant source of the phonetic percept is somewhere in the processes by which the sounds of speech are produced" (p 21).

2.2 SIGNAL-COMPLEMENTARY PROCESSES

Studying speech signal processing physiologically, psycho-acoustically or by numerical modeling, will be highly relevant to obtain more sophisticated formulations of the invariance problem. How far a better understanding of signal-derived processes will take us towards a solution of that problem is not known. Assuming, as the direct realists do, that "the information is in the signal" is a methodologically important move since it forces the investigator to squeeze all he/she can out of the stimulus (Ohala 1986).

Nevertheless, there is a great deal of evidence indicating that speech percepts are also influenced by the current dynamic state of the processing system, that is by

signal-complementary processes.2) That conclusion is sometimes vividly illustrated in introductory classes on auditory analysis and phonetic transcription. When students know a language and are thoroughly familiar with the transcription conventions they show a high degree of interperson agreement. The segments and their phonetic values are by and large "heard correctly" (=according to the conventions taught). However, as soon as the speech sample comes from an unknown language subjects tend to differ more widely both with respect to segmentation and quality judgements. Interestingly, the transcription errors generally make good sense from a physical-acoustic viewpoint and in terms of phonetic similarity.

A similar situation arises when the trained phonetician compares his own transcriptions with spectrograms. Although he knows the language and has considerable experience from listening analytically, he is nevertheless continually surprised by the spectrographic patterns that typically show omissions and contextual modifications in excess of his expectations.

What is going on here? It seems that, if we know a certain language, we cannot help imposing that knowledge on the signal. Physically ambiguous information is disambiguated and incomplete stimulus information is restored. It appears as if the signal-complementary processes modulate the input and shape the percept in a most tangible way. And the process is highly automatic.

adjustments in perception-automatic, not cognitive

Introspectively, there is something utterly compelling about it. As listeners we do not in any way have the impression that disambiguation and perceptual filling in are products of "cognitive processes of inference or hypothesis testing". The compelling nature of the process is illustrated by examples of the following sort: 1. What is your homework assignment? - 2. Less**on** five. - 3. How many came? - 4. Less **than** five. Although the pronunciation of 2 and 4 may both contain physically identical **less'n**, there is no ambiguity. In either case listeners perceive the ambiguous syllabic nasal as intended and are unaware of any signal overlap.

We here enter a huge literature including topics such as intelligibility of filtered speech in noise, effect of increasing access to signal context, perceiving continuity in pulsating signals in noise, phonemic restoration and fluent restoration in shadowing. There is also the phenomenon of **amodal completion** as when subjects are asked to identify an utterance on the basis of visual and reduced

---

2) Cf Lashley (1951:112): "... the input is never into a quiescent or static system, but always into a system which is already actively excited and organized." Also Miller, Galanter and Pribram (1960).

auditory cues. Consider experiments using acoustic stimuli generated from a natural utterance by intensity- and F0-modulation of a steady-state complex buzz spectrum. When these stimuli are presented in synchrony with the video recording of the speaker some of the segments reported are demonstrably physically absent (Risberg 1979).

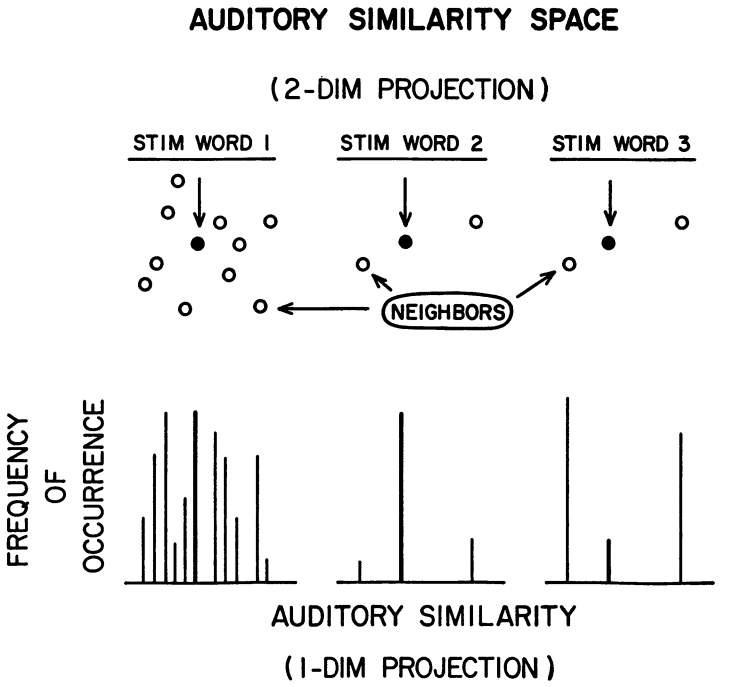

Figure 1. Auditory similarity and frequency of occurrence as determinants of word recognition. Adapted from Luce (1986).

It is impossible to do justice to all the experimental paradigms that elucidate the nature of signal-complementary processes. Let us focus on a particular topic: the **word frequency effect** in auditory word recognition. A recent thesis by Luce (1986) takes a new look at this phenomenon. A model, the **Neighborhood Activation Model** (NAM) is developed and tested. The literature indicates that high-frequency words exhibit a number of processing advantages. The higher the frequency of a word the higher its probability of being correctly recognized. That is true in general but NAM predicts departures from this basic rule. It views word identification as the task of discriminating

among lexical items in memory and makes this discrimination a function of the number and nature of the elements that the input signal activates. The recognition process is assumed to be influenced not only by the frequency of the stimulus word itself but also by how auditorily similar it is to other words in the "neighborhood", by how many such neighbors it has and by what the frequencies of these neighbors are. Figure 1 is our attempt to capture those ideas.

At the top the locations of three stimulus words in a hypothetical "auditory similarity space" are indicated. For simplicity only two dimensions are drawn. Unfilled circles refer to neighbors, solid dots to stimulus words. Stimulus word 1 is in a crowded neighborhood. In contrast, stimulus words 2 and 3 have only two neighbors each. The auditory similarity structures of 2 and 3 are assumed to be the same as indicated by the identical relative positions of the points. At the bottom we see a one-dimensional projection of the stimulus and neighbor points onto the abscissas. The ordinate is word frequency. Comparing stimulus words 2 and 3 we notice that they have the same frequency. However the neighbors of 2 have lower frequencies than those of 3. Although they are similar with respect to acoustic-auditory confusability and frequency of occurrence, NAM predicts that they may nevertheless behave differently on certain processing tasks. To test the model Luce ran an identification test.

About 900 English words of CVC structure were drawn from Webster's dictionary. The selected items were listed in the Brown corpus of frequency counts and had been rated with respect to familiarity in a parallel study. The stimuli were spoken by a male talker and presented to listeners at three S/N-ratios. A set of confusion-matrix experiments were also carried out. One test contained the CV sequences formed by by combining 23 initial consonants (including a null consonant) with 15 vowels. Another set was produced from 15 vowels and 22 final consonants (again including a null element). These tests were run at the same three S/N-ratios.

NAM defines the probability of correctly recognizing a stimulus word in the context of its neighbors as:

$$p=S/(S+\Sigma N) \qquad (1)$$

where S is the product of the discriminability and frequency of the stimulus word. ΣN is the sum of terms associated with individual neighbors, each computed in a manner analogous with the calculation of S, that is an individual N is the product of the confusability and frequency of a given neighbor. The discriminability of a given word is derived from the confusion matrices. The

probability of correctly recognizing the word "beat", from its acoustic waveform only, is computed as the product of three terms all available from the confusion-experiment data: The probability of responding by /b/ when presented with /b/, the probability of responding by /i/ when presented with /i/ and the probability of responding by /t/ when presented with /t/. Similarly, the confusability of of a neighbor, say "meat", is the product of the three terms: probability of responding by /m/ when presented with /b/, the probability of responding by /i/ when presented with /i/ and the probability of responding by /t/ when presented with /t/.

Let us examine Eq (1), as Luce did, by considering four cases obtained by systematically combining a high and a low value of S with a high and a low value of ΣN. A word with a highly distinctive acoustic pattern and with a high frequency would make S large. A high value of ΣN would be observed in a case where there are many neighbors that could easily be confused with the stimulus word and whose frequencies are high. On the other hand, a low frequency and a high confusability with neighbors makes S low. A low value of ΣN would result when the stimulus word has few neighbors whose frequencies and similarity scores are low.

TABLE I Testing the NAM model. A comparison of identification performance (%) and qualitative predictions (in parentheses) derived from Eq (1).

		value of S	
		low	high
value of ΣN	low	51 (intermed)	64 (high)
	high	38 (low)	55 (intermed)

Table I presents predictions (in parentheses) derived by plugging high and low values for S and ΣN into Eq (1). Also shown are the percent correct identifications observed for the four groups of words meeting the criteria of the matrix

cells. The percentages can be seen to vary as expected from the predictions.

Our summary of NAM serves the purpose of demonstrating that factors independent of the signal - in this case word frequency and lexical sound structure - do indeed play a role in perceptual processing. Secondly, how these factors come into play can be studied in a rigorous manner by building quantitative models.

2.3 DIRECT PERCEPTION

It might at first seem that the conclusions reached in the preceding paragraphs are incompatible with the Direct Realism approach. This is not necessarily so. The source of signal-complementary processes is knowledge ('internal representations') of various kinds that interacts with the incoming signal in percept formation. If we assume that that knowledge would become accessible more or less instantly upon stimulation (cf parallel processing), then postulating a role for "information not the signal" need not include "mediation by cognitive processes or hypothesis testing". The signal-complementary information embodied in the lexicon network would start modulating the signal immediately and in a direct manner.

In his Gibson Memorial Lecture, Shepard (1984) presents a theory that attempts to reconcile the direct nature of perception with a role for 'internal representations', notions normally avoided and usually explicitly rejected by Gibsonians. Basic to the ecological approach is the belief in "smart" perceptual processes capable of extracting higher-order invariants in an automatic way. Shepard agrees with Gibson that "the brain has evolved to extract invariants under favorable conditions ..".

But he adds that we must also assume that it has "evolved to serve the organism under less favorable conditions of nighttime, obstructed, and spatially or temporally limited viewing and, even, of structural damage to the brain itself" (p 419). He goes on to develop the idea that, to deal with non-optimal stimulus conditions, brains were driven to internalize significant external conditions, and to develop an ability to build mental models of the world. This process which underlies the ability to remember, to anticipate, and to imagine objects and to plan events in their absence is not separate from perceptual functions. This is where Shepard differs from Gibson.

If, as Shepard suggests, signals are assumed to "resonate with" internalized knowledge, direct perception and signal-complementary processes ('internal representations') are perfectly compatible notions.

3. What Do Speakers Do?

3.1 CLUES FROM BIOLOGY I: ECONOMY

The themes of the next two sections are summarized in Figure 2. It shows a window cleaner and the compensatory articulation of a bite-block vowel (Lindblom 1983). Two important movement characteristics are illustrated here: economy and plasticity. The bottom diagrams illustrate the expected normal behaviors in which the feet and the jaw synergistically participate to facilitate the two tasks, cleaning the pane on the right and producing the vowel /i/. The top diagrams show compensatory behaviors. The feet and jaw assume positions that necessitate extreme displacements of the arm-body and the tongue respectively. The point is this: **Unconstrained, a motor system tends to default to a low-cost form of behavior.**

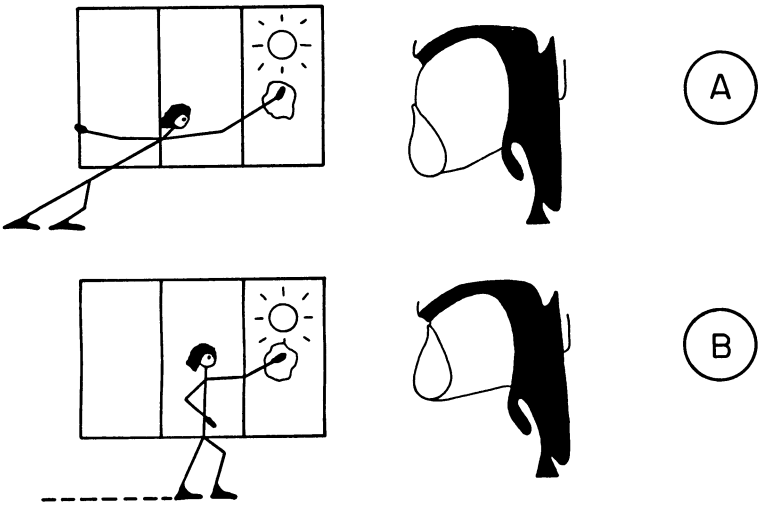

Figure 2. Movement economy and plasticity. A: Compensatory behaviors. B: Normal behaviors.

This trend is pervasive in all kinds of motor behavior including speech. First let us review some non-speech cases. Hoyt and Taylor (1981) observed rates of energy consumption (measured as amount of oxygen consumed to move a certain distance) for a walk, a trot and a gallop in three small horses. The animals were trained to vary their speed within each gait while moving on a motorized

treadmill. 'Cost functions' were derived by plotting oxygen consumption against running speed. They form parabolic contours that are concave upward and have clear minima at speeds that depend on the type of gait.

Data on oxygen consumption and speed are presented also for free movement. They indicate that, within each gait, speeds tended to be selected near the minima of the cost functions established in the first experiment. The authors conclude that "the natural gait at any speed indeed entails the smallest possible energy expenditure." Using biomechanical models Milsum (1966) worked out cost functions for breathing and bipedal walking that are similar in form to those reported by Hoyt and Taylor. Rates predicted on the basis of least power expenditure compare favorably with actual rates observed under conditions of rest and exercise.

Nelson, Perkell and Westbury (1984) compared jaw movements during increasingly rapid opening and closing alternations ('wags') and during increasingly rapid repetitions of the syllable /sa/. Both for the speech and the non-speech task the results indicate that mandible excursions tend to decrease as movement times become shorter. They show that for jaw displacements to remain large and independent of gesture duration, it would be necessary to increase peak velocities from 40 mm/s (for movements of long durations) to over 250 mm/s (for movements of short durations). In fact the observed maximum velocities cluster around 100-120 mm/s. They argue that, since peak velocity represents a theoretically plausible measure of "biomechanical effort" (Nelson 1983), their findings support the idea that speech as well as non-speech movements are constrained by a principle of physical "economy".

The results of Nelson et al are compatible with the so-called **duration-dependent undershoot** model proposed to explain vowel reduction (Lindblom 1963): As vowel duration in a CVC syllable becomes shorter and shorter, the extent of the movement towards the vowel target is reduced: Hence both articulatory and acoustic undershoot. However, several studies (Kuehn and Moll 1976, Nord 1986, Engstrand 1988) provide data that refute the idea that presence/absence of undershoot depends only on duration as implied by the original undershoot model. For instance, it is evident that sometimes targets can be reached also at very short durations with no trace of undershoot (e.g. Gay 1978).

A theoretical analysis along the lines of Nelson (1983) and Nelson, Perkell and Westbury (1984) provides the clue to resolving this issue. For simplicity consider a single articulator, say the jaw, which we model biomechanically as a strongly damped spring-mass system. Further let us analyze movements in a CVC syllable as the response of such

a system to a sequence of stepwise alternating force values: "close!"-"open!"-"close!" This metaphor helps us understand on the one hand why the system will exhibit undershoot when the commands arrive too fast for the system to follow. On the other hand, it also shows how undershoot behavior could in principle be avoided: Compensation for rapid timing of the force commands can be achieved by increasing movement velocities which in turn is brought about by increasing force amplitudes (cf velocity data in Kuehn and Moll (1976)). However, as pointed out by Nelson et al increasing peak velocities entails greater biomechanical power expenditure. If the speech system operates so as to minimize 'articulatory effort' (peak velocities), we should expect it to undershoot phonetic targets quite often, but not necessarily in every single instance. The key point is: **Speakers have a choice.**

3.2 CLUES FROM BIOLOGY II: PLASTICITY

The second point made in Figure 2 is the output-oriented

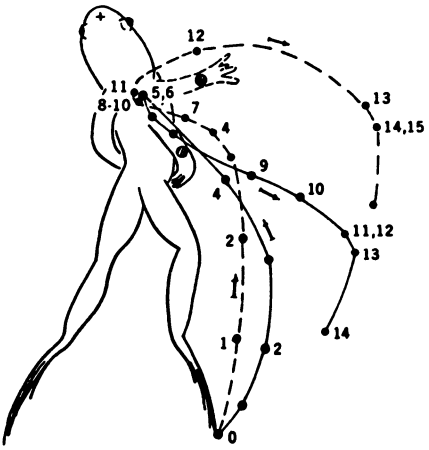

Figure 3. Wiping reflex of spinal frog as a function of foreleg position. (Adapted from Grillner 1982).

nature of motor control, the fact that movements appear to be 'purpose-driven' (Sherrington 1941, Granit 1979). The end products of the compensatory and unconstrained /i/-articulations (or window cleaning) are the same although the movements achieving them are very different. This prospective means-end organization is not special to speech. A lot of evidence shows that it is a perfectly general feature of motor control.

Grillner (1982) draws attention to a study of the wiping reflex in a frog whose spinal cord had been transected at a high cervical level (Fukson, Berkinblit and Feldman 1980). The trajectory of the frog's hindlimb was traced from film frames. When an irritant substance was applied to the forelimb, the frog responded by trying to wipe it away. Significantly, the path of the trajectory varied depending on the position of the foreleg at the moment of stimulation (Figure 3). Although the details of the computational mechanism in the spinal cord remain unknown, the result clearly demonstrates the presence of both adaptive plasticity and goal-orientation in non-speech movements.

Recall also the followng familiar observation. With ease we often recognize a friend's handwriting, whether the letters are small and occur on a piece of paper, or they are large and appear on a blackboard.

OUTPUT-ORIENTED MOTOR CONTROL

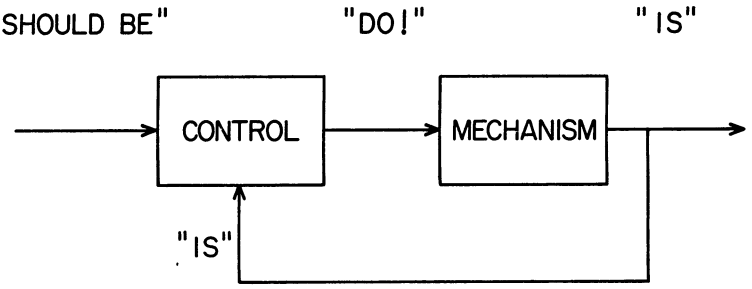

Figure 4. Some components of a feedback system.

Personal characteristics are sometimes preserved under unusual circumstances, as when right-handers are asked to try upside-down mirror writing with the left hand and with eyes closed (Lashley 1951:124).

Experimental observations of this sort suggest that, paradoxically, the execution of a given set of movements is preceded by the construction of a representation of the end-product of those very movements - a representation that also shapes their implementation (Sherrington 1941, Lashley 1951, MacNeilage 1970, Granit 1979).

Figure 4 is an attempt to summarize our discussion so far. On the one hand, we have stated that unconstrained a movement tends to default to a low-cost form of behavior. Accordingly, when an /i/ is produced without a bite-block, a tongue gesture is invoked that deviates little from a neutral configuration ('economy'). However, when a bite-block is in fact present, the system is perfectly capable of compensating ('plasticity').

Common to both situations is the goal, the "SHOULD-BE" value, and an "IS"-signal, the current state of the system as reported to the CONTROL component by various sensory feedback channels. Given "IS" and "SHOULD-BE" the controller computes a course of action and tells the mechanism what to do (the "DO"-signal).

What can we learn from this simplification of the complex phenomena that go on in speech? Recall our discussion of reduction and undershoot. We concluded that within limits speakers appear to have a choice whether to undershoot or not to undershoot. We also noted that avoiding undershoot at short segment durations entails a higher biomechanical cost. The model in Figure 4 offers a way of thinking about how such choices could be implemented.

The control component operates on the basis of error information. It looks at the difference between "SHOULD-BE" and "IS". The current error value can be amplified, or attenuated, by varying the "gain of the feedback loop". When the error is allowed to have a strong influence, the system response will exhibit faster and more accurate target attainment. This corresponds to a compensatory and output-oriented mode. When the role of the error information is attenuated, the "DO"-signal drives the system less forcefully. In this non-compensatory and system-oriented mode target attainment is less efficient and the response is shaped more by system constraints. There is little compensation for the intrinsic dynamic characteristics of individual articulators, should they interfere with the implementation of the "SHOULD-BE"-command.

By conceptualizing speech in this way we obtain a way of describing the mechanism for tuning target attainment

according to the demands of the listener and the situation, the notion of "controlling the gain of the feedback loop". This, we suggest, is how speakers move around along the dimension of **hyper-** and **hypospeech**. In global terms, this is how they are able to over- or underarticulate.

4. The H&H Argument Illustrated

We have looked at two sources of information for speech perception: signal-driven and signal-independent processes. We have singled out two significant characteristics of speech motor control: economy and plasticity. We are ready to move on by putting two and two together (cf 1.2, **e-g**).

Let us examine H&H against the background of speech as produced not only in the laboratory but also in its natural, ecological settings. Without aspiring to be exhaustive, in Figure 5 we survey factors that shape the phonetic variation in the speech of a single individual. The simplification that H&H imposes is that phonetic variation is basically one-dimensional: When output constraints dominate, we expect to see hyperforms, whereas with system constraints dominating, hypospeech will be observed.

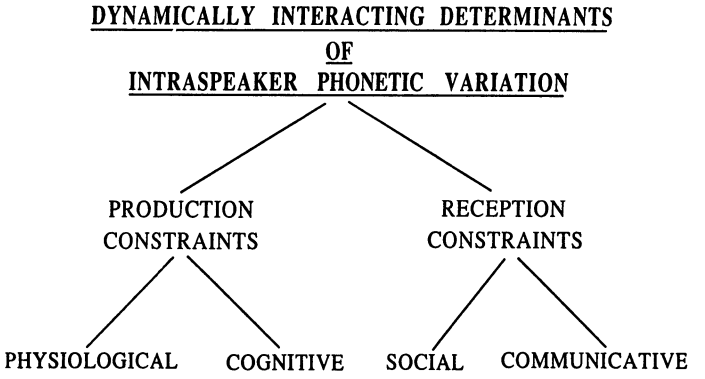

Figure 5: Factors causing the pronunciation of spoken forms to vary.

In keeping with this view Figure 5 distributes the factors influencing intra-speaker variation between production constraints and perception constraints. Under these two

nodes we place physiological factors (mostly involuntary modulations such as emotion, disease, etc...), cognitive factors (speaking to oneself, propositional vs automatic styles (Bates 1979:359), etc...), social and communicative factors (channel, listener, situation, degree of formality ...).

These complex phenomena are mentioned to make clear that the assumption about H&H variation being one-dimensional is a deliberate simplification which is likely to be revised in the course of further work. However, for the moment we regard this assumption as heuristic and methodologically justified.

4.1 OUTPUT-ORIENTED CONTROL

It might be suggested that the fact that subjects do well on experimental tasks involving compensatory articulation (e g bite-block speech) is interesting but irrelevant to phonetic theory since bite-blocks create unnatural, non-speech conditions. We favor an alternative view: Good performance on compensatory tasks is a consequence of the fact that motor behavior in general is organized to be inherently compensatory. Sundberg's work on singing (Sundberg 1987) provides some instructive non-speech examples of output-oriented control.

Female opera singers often produce pitches whose frequencies radically exceed the normal values of the first formant. This is particularly true for close vowels. Acoustic theory predicts that when there are no strong harmonics to support the first formant, the intensity of the sound will go down. Sundberg has shown that singers "track" the increases in fundamental frequency by adjusting the frequency of F1. This is done by lowering the mandible so as to raise F1 in step with F0, thus maintaining an output of sufficient intensity.

Another example is the account of the origin and function of the "singer's formant" (Sundberg 1987:118). This is a spectral prominence primarily due to a clustering of the third, fourth and fifth formants. According to Sundberg's measurements and studies of vocal tract models, this cluster seems to be brought about by larynx lowering which has the effect of making the vocal tract longer as well as making certain structures (laryngeal ventricle and sinus piriformis) wider. He measured long-term average (LTA) spectra for three types of signal: orchestra, speech, orchestra + singing. The LTA-spectra of the orchestra and the speech were found to coincide closely, whereas that of the singer exhibited a salient peak in the region of the "singer's formant" rising above the orchestra envelope. Recordings of singing - produced with and without the "singer's formant" and superimposed on noise with the

long-term spectral properties of orchestral sound - demonstrate that the "singer's formant" makes the singing more audible in the presence of an orchestra. Sundberg notes that the "singer's formant" is typical of all voiced sounds produced in Western concert-style singing except soprano singing. Why not sopranos? According to Sundberg the higher pitch range of sopranos causes their LTA-spectra to overlap less extensively with the LTA-spectra of orchestral sound. Their audibility may therefore be sufficiently high without a "singer's formant" - a circumstance fully in line with his explanation and a further instance of the notion of output-oriented motor control.

4.2 OUTPUT- VS SYSTEM-ORIENTED CONTROL

So far our story conforms with accounts of speech production that view it as a continual **tug-of-war** between demands on the output on the one hand and system-based constraints on the other (cf the papers by Keller and by Kohler, this volume). Here is another finding from the Sundberg research team that highlights this balance of forces.

In an X-ray investigation Johansson, Sundberg and Willbrand (1983) studied the vowels /i: a: u:/ spoken and sung. The fundamental frequencies of the sung versions were 230, 465 and 940 Hz. Mid-sagittal contours of the tongue body were traced (Sundberg 1987:Fig 5.25). For the lowest pitch the tongue contours of sung and spoken vowels coincide rather closely. At 930 Hz, however, essentially the same tongue configuration was used for all three vowels. It is interesting that this shape comes closest to the most neutral of the vowels /a:/. This result is a good illustration of the rule proposed in Section 3.1: "Unconstrained, a motor system tends to default to a low-cost form of behavior".

at high pitch while singing- why unconstrained?
oh, as in, when all the vowels do have to converge, they converge to the position that is most low-cost

4.3 PARTIAL COMPENSATION

Compensations can be seen as cases where output constraints overrule system constraints. It is an interesting fact about most articulatory compensations that, when examined closely, they turn out to be partial, not perfect. Such observations support the correctness of the present tug-of-war scenario. Let us look at four examples of compensatory speech production behavior.

In a study of the duration of open and close vowels (Lindblom 1967) I measured lip and jaw movement in /I'bVb(:)I/-sequences and found evidence supporting the Extent of Movement Hypothesis. This hypothesis was proposed by Eli Fischer-Jorgensen (1964) to explain the durational

differences between open and close vowels: I found that the jaw began its opening gesture for the V during the first /b/-segment and returned to a high position around the middle of the second /b/. In test words with open and close vowels, say /I'ba:bI/ and /I'bi:bI/, the jaw contributed to an earlier separation of the lips in the first /b/ and delayed their coming together again for the final /b/.

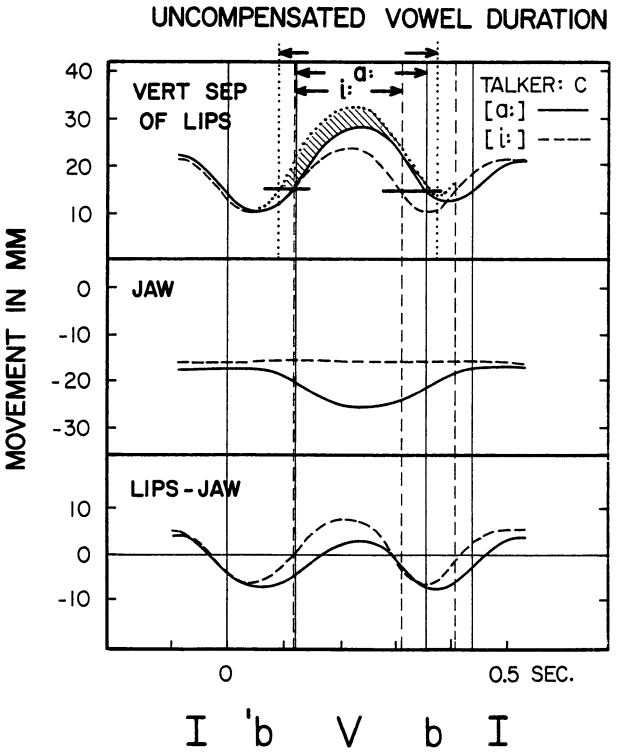

Figure 6: Compensated and uncompensated vowel duration. Vertical lines indicate acoustic segment boundaries. Solid and dashed lines show measurements from which the dotted lines and the dotted curve were derived.

Since open vowels had larger excursions of the jaw than close vowels, the effect on the timing of the vowel segment

boundaries was greater for the open vowels. It was thus, everything else being equal, responsible for the longer duration of those segments.

The reason for returning to those findings here is that the condition of "everything else being equal" was not always met. It was violated in an interetsting way. In words with open vowels the lip gestures for the /b/:s (corrected for superimposed jaw movement) were found to actively resist a premature opening of the first /b/ and to be closer to each other for the second /b/. Had this vowel-dependent reorganization of the lip gestures not taken place, the durational differences would have exceeded the observed maximum of 45 ms by a large margin.

This conclusion is reached by "synthesizing" vowel durations according to the method described by Schulman (1989). In Figure 6 above we present data on /a:/ and /i:/ taken from Fig I-A-14 of Lindblom (1967). We derive vowel durations by systematically cross-adding the /i:/ and /a:/ jaw movements (middle panel) to the /i:/ and /a:/ lip gestures (bottom panel) and noting where in time the resulting synthesized lip separation curves reach the criterion value for separation/closure. For instance, in Figure 6 above the top dotted curve was produced by adding the jaw curve for /I'ba:bI/ to the lip component for /I'bi:bI/. This exercise results in a vowel duration of 310 ms instead of the observed 235 ms. The point is this: Clearly there is compensation in the lips for the superimposed jaw movement but that compensation is no more than partial since, in the observed cases, the open vowels nevertheless remain longer than the close ones.

Loud speech generally exhibits much wider jaw openings for vowels. Producing loud speech is thus like speaking with a "natural bite-block". Schulman (1989) collected lip and jaw movement data from four speakers who produced tokens of twelve Swedish vowels in an /i'b_b/ frame. His study showed up to threefold increases in stressed vowel jaw opening, a shortening of the /b/:s and a lengthening of the stressed vowel for the loud condition. To interpret these changes, he constructed a simple model of loud speech on the assumption that loud speech movements are related to normal movements simply in terms of linear amplification. Normal lip and jaw movements were multiplied by scale factors. Adding the derived jaw and lip traces he obtained a curve representing the vertical separation of lips (Schulman 1989, Figure 12). This procedure demonstrates that loud speech involves a more complex transform than mere linear amplification of movements. For instance, for the loud vowels the jaw makes excursions that are 2 to 3 three times larger than those of the normal vowels. However, for the surrounding consonants in loud tokens, the jaw is not thus scaled. It remains in its normal high

position. Schulman's simulation shows us why: To keep the effort-dependent increases in vowel duration within bounds and so as not to totally jeopardize the attainment of stop closure. We conclude that the loud-speech transform is goal-directed and involves articulatory manoeuvres that are compensatory. Note that this compensation is partial. The system tolerates some increase of loud vowel durations but shows restraint.

The interaction of jaw movement and segment durations is also the topic of Lindblom, Lubker, Lyberg, Branderud and Holmgren (1987). In this study reiterant speech samples were used modeled on real Swedish mono-, bi- and trisyllabic words: ['bab: 'bab:ab ba'bab: 'bab:abab ba'bab:ab baba'bab:]. The subjects were all unaware of the goal of the experiment. They spoke the test words both normally and with sizeable bite-blocks. The difference in jaw opening between normal and bite-block conditions was of the order of 15-20 mm. The question addressed was this: Would speakers be able to achieve /b/-closures in spite of the fixed and abnormally low jaw position? And if so, what segment durations would they produce under such conditions? It was found that bite-block segment durations reproduced normal vowel and consonant durations within 15 ms and that there was a tendency for the bite-block condition to induce complementary consonant shortenings and vowel lengthenings (cf Schulman's findings). It appears that these "results do not preclude that the reorganization of articulatory gestures takes place with respect to segment duration only in so far as it is necessary to maintain perceptual constancy of stress patterns. Hence we observe **selective** or partial segment duration **compensation**" (p 178).

Gay, Lindblom and Lubker (1981) obtained articulatory data on normal and bite-block steady-state vowels from still lateral X-ray pictures. Normal and bite-block contours were compared in terms of a maxilla-based coordinate system. Compensations were close to perfect near maximum constrictions, but incomplete at points with large cross-sectional areas. Area functions derived from the X-ray tracings were subjected to a theoretical perturbation analysis which showed that the strategy actually followed by the subjects was the acoustically optimal one. Once again the same pattern emerges: Although compensations are invoked that are selective, acoustic equivalence is nevertheless achieved.

4.4 SUFFICIENT CONTRAST: AN ANALYSIS OF COARTICULATION

One influential view of coarticulation is the one that has guided the development of the Motor Theory of Speech Perception. Recall the earlier quotation from Liberman and Mattingly (1985:7): The MT "assumes that adaptations of the

motor system for controlling the organs of the vocal tract took precedence in the evolution of speech. These adaptations made it possible, not only to produce phonetic gestures, but also to coarticulate them so that they could be produced rapidly. A perceiving system, specialized to take account of the complex acoustic consequences, developed concomitantly." In other words, coarticulation is seen as an adaptation to a demand for a preferred speaking tempo. "A function of coarticulation is to evade" ... "that each unit would become a syllable, in which case talkers could speak only as fast as they could spell" (p 13). Also, reviewing the multiplicity and variety of cues, Liberman and Mattingly write: "... we should conclude that there is simply no way to define a phonetic category in purely acoustic terms" (p 12).

Another well-known, but entirely different view is that explored by Stevens and co-workers (Stevens and Blumstein 1978, 1981; Blumstein and Stevens 1979, 1981). This research program involves a theoretically motivated search for acoustic properties that remain invariant across talkers and phonetic contexts and whose perceptual role is demonstrated experimentally.

Below I shall discuss two sets of data that bear on theories of coarticulation. In particular, we shall examine the possibility of interpreting coarticulation phenomena in terms of "sufficient contrast" as suggested by the H&H theory.

The first point is a restatement of Öhman's (1966) findings. Öhman's speech samples were symmetrical and asymmetrical VCV sequences. They were produced by generating all possible combinations of /b d g/ and /y ö a o u/ and were spoken by a Swedish subject. His Tables II and IV contain formant frequency measurements for the initial and final vowels and for the points corresponding to the VC and CV boundaries (here called "locus" patterns). Vowel-consonant coarticulation is extensive. The locus patterns and the formant transitions are not unique. "... place information for a given consonant is carried by a rising transition in one vowel context and a falling transition in another" (Liberman, Delattre, Cooper and Gerstman 1954). At the CV boundary formant patterns depend not only on the identity of V_2 but also on V_1. Conversely, at the VC boundary they depend on both V_1 and V_2. At first glance one might feel tempted to agree with Liberman and Mattingly conceding "that there is simply no way to define a phonetic category in purely acoustic terms".

Is such a conclusion really justified? In Figure 7 we plot F2 measured at the CV-boundary along the x-axis, F3 also measured at the CV-boundary along the y-axis and the F2 vowel target on the z-axis. This 3-D view displays three "clouds" that enclose all the V_1bV_2, V_1dV_2 and V_1gV_2

measurements. Although the consonant formant patterns exhibit massive context-dependence, the 3-D space shows three configurations that do not overlap. The point is this: If other dimensions were also considered (spectral dynamics, aspects of the burst spectra etc), we would expect even better separability. Owing to the large coarticulation effects there are no absolute invariants. However, the phonetic correlates of the three stop categories are nevertheless distinct. They meet the condition of "sufficient contrast". For similar results for English stops see Sussman (1989).

"Unconstrained, a motor system tends to default to a low-cost form of behavior". Is the preceding analysis of coarticulation compatible with that rule? If it is true that the constraints on perceptual distinctiveness tolerate some variation in locus patterns - manifested as clouds rather than points in a multi-dimensional representation - what is the source of that variation? Why is it there? H&H theory assumes that locus coarticulation is there - not to serve perception in the first instance (although it sometimes provides valuable cues) - but primarily to facilitate production. If so, in what sense is vowel-consonant coarticulation "a low-cost form of behavior"? An extension of work previously done on vowels (Lindblom and Sundberg 1971) sheds some light on those questions.

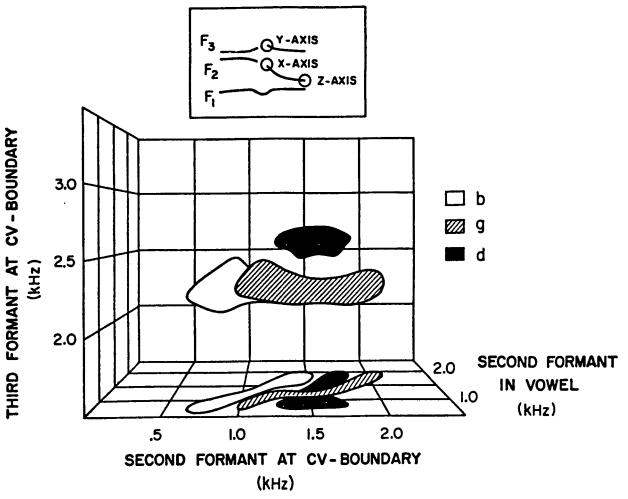

Figure 7: Acoustic separability of stop consonants. Adapted from Öhman (1966).

Two sets of experiments were done in which the following articulatory parameters were controlled: (1) Lip configuration: The subject (JS) held a cylindrical brass tube between his lips so that it made contact with his upper incisors. This allowed us to control the length and the cross-sectional area of the lip section; (2) Jaw opening: A bite-block creating a 7 mm jaw opening was introduced between the subject's molar teeth; (3) Place of apical closure: The place of tongue tip was marked with a small pellet on a false palate.

In the first series of experiments we collected data on coarticulation effects in the formant patterns of Swedish dental and retroflex stops in VCV utterances (cf Öhman 1966). We included naturally spoken VCV utterances as well as tokens produced with the set-up described above. The second series was aimed at determining the acoustically relevant properties of the cavity in front of the apical closure. Using the lip tube, the bite-block and the false palate with the marker, the subject produced VCV utterances in which the place of articulation of the consonant was systematically varied. The subject was instructed to freeze his articulatory movement as he reached the closure. While he thus held his articulation steady, a sine-wave sweep-tone was applied to the front cavity. The frequency at which its amplitude had a maximum was taken to be the resonance frequency of the front cavity system. The introduction of the lip tube made that system similar to a Helmholtz resonator with a neck corresponding to the lip tube and a volume corresponding to the space in front of the occlusion. This circumstance enabled us to use the resonance frequencies to estimate front cavity volumes as a function of point of articulation (Sundberg and Lindblom 1989).

In parallel with these measurements we used our articulatory model to derive a set of nomograms relating variations in articulatory parameters to their acoustic consequences. This model has a jaw-based coordinate system for the lips and the tongue (Lindblom and Sundberg 1971). It uses two dimensions for generating tongue body contours: constriction position (where?) and deviation from neutral (how much?). The tongue tip parameters are analogous: retraction-protrusion (where?) and elevation (how much?) (Lindblom, Pauli and Sundberg 1975). This framework first generates a profile from specifications of jaw, lip parameters, tongue tip, tongue body and larynx position. The profile is coded as a table of vocal tract cross-distances. The distances are converted into cross-sectional areas following two methods. Posterior to the closure we used our previously established rules (Lindblom and

Sundberg 1971). In front of the closure we derived areas from the volume data gathered in the sweep-tone experiment.

To investigate the patterns of coarticulation observed in the spectrographic VCV measurements, we adopted an analysis-by-synthesis procedure addressing the following question: What model configurations show formant patterns identical with those measured in the VCV words? Recall that in one of the VCV experiments we controlled articulation with the aid of the lip tube, a 7 mm jaw opening and indications of place of apical closure. With the model we had obtained parallel information simulating those conditions. For any given place of articulation the parameters free to vary were the retraction-protrusion, constriction position and deviation from neutral. For each place of apical articulation we prepared a set of diagrams showing how F2 and F3 of the model varied as a function of these three parameters. Then we adopted the following graphical procedure for the comparing the measurements with the model: (i) Input the F2 and F3 observed for a given place of apical closure; (ii) Determine the best acoustic match that the model produces for that place; (iii) Read off the articulatory parameter values associated with that match.

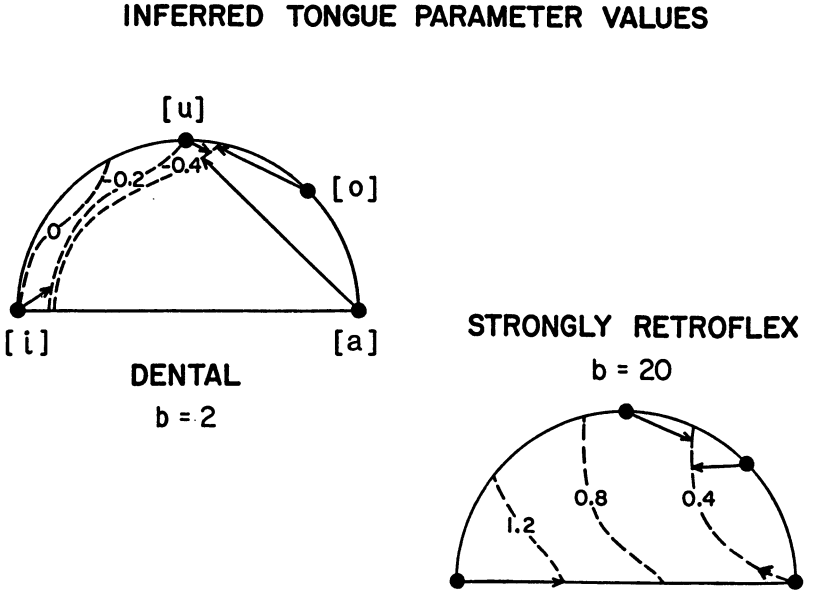

Figure 8: Coarticulation in dental and retroflex stops. Values of tongue body parameters were inferred from articulatory model simulations and matching the derived F-patterns to the observed formant data.

Figure 8 shows the analysis-by-synthesis results for apical stops produced at 2 and 20 mm behind the upper incisors in five contexts: /i-i/, /ɛ-ɛ/, /a-a/, /o-o/ and /u-u/. Formant measurements were made at VC boundaries. The semicircles stylize the tongue body space. A point within the semicircle is specified by two numbers: angle of radius (constriction location) and displacement from origin along a given radius (deviation from neutral). Vowel contexts are shown as solid dots. The arrows indicate the values of the best matches. The dashed wavy lines are **compatibility contours**. They represent families of tongue body shapes, that is those tongue body contours that are compatible with a given place of articulation, **b**, and a given degree of retraction-protrusion. (The more positive the number next to dashed line the greater the retraction).

The point that Figure 8 makes is this: As indicated by the compatibility contours in both semicircles, the model makes available a large range of tongue shapes. However, the arrows do not show arbitrary positions on those contours. They occur in the local neighborhood of the vowel dots.

The significance of that finding is simply that the tongue shapes of apical stops tend to resemble those of the adjacent vowels. That, of course, is not a new observation. However, when we view it in terms of the matched model parameters we arrive at an answer to the questions raised earlier: What is the source of the locus pattern variations? Why are they there? In what sense is vowel-consonant coarticulation "a low-cost form of behavior"? The answer is that they are there because the tongue tip and body do their job following a pattern of synergy we also saw in the window cleaner and the bite-block /i/. The displacement of the tongue body invoked to facilitate apical closures is a fraction of the maximum possible. It is in this sense that vowel-consonant coarticulation constitutes an example of low-cost motor behavior.

4.5 VOWEL REDUCTION IN CLEAR SPEECH

Recent work (Moon and Lindblom 1989) indicates that "clear speech" is not merely normal speech produced louder. It also involves reorganization of articulatory gestures and acoustic patterns. Five speakers of American English read lists of words with syllables exhibiting large F2 transitions: wheel, will, well, wail. To obtain vowel duration variations the syllables occurred as monosyllbles and were embedded in bi- and trisyllabic frames generated by appending -ing and -ingby or -ingham (place name endings): e g well, welling, Wellingby. These lists were

read in two ways: First at a rate and loudness spontaneously chosen by the subject (citation forms); Then in response to an explicit instruction to "overarticulate", and "to speak as clearly as possible". Citation-form style repetitions of each vowel in an /h_d/ environment were also recorded for each subject. Vowel durations and F2 and F3 at vowel midpoints were measured for all tokens. Formant frequencies were plotted as a function of vowel duration to see whether there were undershoot effects. All speakers showed undershoot. Its degree varied somewhat depending on vowel, talker and style.

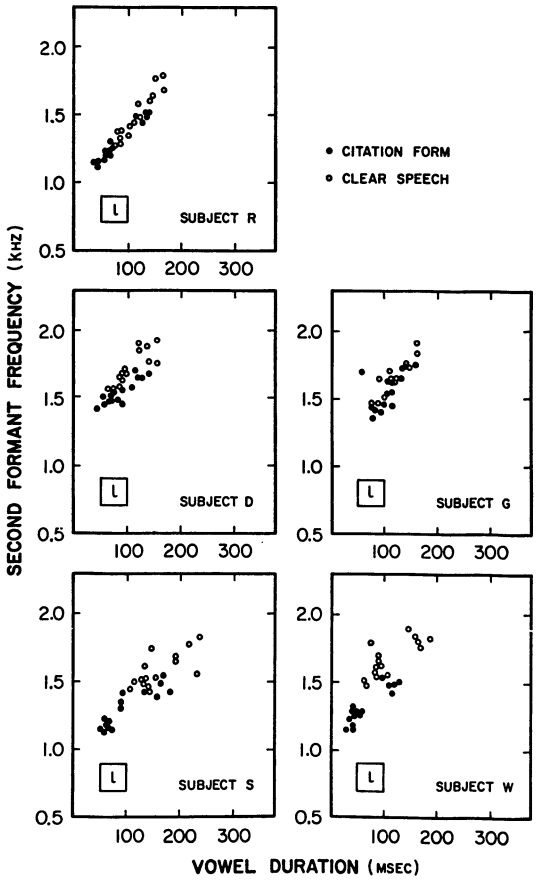

Figure 9: Duration-dependent undershoot in clear and citation-form speech. Measurements from syllable will-.

Both clear and citation-form speech showed undershoot. But clear speech formants were generally closer to the /h_d/ reference values than the corresponding citation-form measurements. (This effect was not simply due to clear speech tokens being longer (which they often were). When exponentials were fitted to the formant-vs-duration plots (cf Lindblom 1963), separate curves were needed to describe the two styles (Moon (in prep), cf also Engstrand and Krull (1989))

Why does clear speech have these properties? Both the existence and the nature of a clear-speech mode make eminently good sense from the H&H viewpoint. So does the preliminary evidence that formant patterns approximate targets more closely in clear than in citation-form speech. Should further analyses establish that clear-speech transforms are indeed communicatively functional - that is, they increase both the audibility and the intelligibility of speech (Moon in prep) - a natural interpretation would immediately become available in terms of the H&H theory.

5. Implications for the Invariance Issue

5.1 IN-PRINCIPLE SOLUTIONS

The Motor Theory (Liberman and Mattingly 1985), the Quantal Theory (Stevens 1989) and the Direct Realism approach (Fowler 1986) can all be said to offer in-principle solutions to the invariance problem. Although different in other respects, MT and DR explicitly assume that phonetic invariance is gestural.

Liberman and Mattingly (1985:22) "Though we have a great deal to learn before we can account for the variation in instances of the same gesture, it is nonetheless clear that, despite such variation, the gestures have a virtue that the acoustic cues lack: instances of a particular gesture always have certain topological properties not shared by any other gesture."

And Fowler (1986:11): "By hypothesis, the organization of the vocal tract to produce a phonetic segment is invariant over variation in segmental and suprasegmental context." Cf also (p 13): "How do listeners recover phonetic structure from such a signal? One thing is clear; the functional parsing of the acoustic signal for the perceiver is not one into acoustic segments. Does it follow that perceivers impose their own parsing on the signal? There must be a "no" answer to this question for any event theory devised from a direct-realist perspective to be viable. The perceived parsing must be in the signal; the special role of the perceptual system is not to create it, but only to select it."

The QTS is developed from the assumption that acoustic/auditory stability plays a major role in the selection of the phonetic segments and features of the languages of the world. There seems to be a clear link between the functional value attributed to stability in QTS and the experimental search for acoustic invariants that Stevens has been engaged in. In several papers the approach taken builds on the fact that the "work of Fant (1960), Jakobson et al (1963) and Halle et al (1957) in particular, has suggested that spectral analysis of the speech signal results in distinctively different patterns corresponding to individual phonetic features." ... "The notion that spectral analyses reveal distinct patterns for individual phonetic features suggests that there may indeed be an invariance of acoustic patterns which can characterize place of articulation independent of the following vowel" (Blumstein and Stevens 1980:648).

These research programs represent distinct working hypotheses about the resolution of the invariance problem. However, they do have one theme in common: They all share the assumption that the ultimate solution will be found in the signal.

In contrast the H&H theory assumes that, in all instances, speech perception is the product of both signal-driven and signal-independent information, that the contribution made by the signal-independent processes show short-term fluctuations and that speakers adapt to those fluctuations. It says that - whether communicatively successful or not - that adaptive behavior is the reason for the alleged lack of invariance in the speech signal. Hence it predicts that the quest for signal-based definitions of invariance will continue to remain unsuccessful as a matter of principle. In the H&H model the need to solve the invariance issue disappears. But the problem is replaced by another - one that appears equally forbidding but should, in principle, have better prospects of success: That of describing the class of speech signals that satisfy the condition of "sufficient discriminative power".

Let us pause to try to envision a hypothetical situation in the future. Imagine that the invariance problem has been solved and that one of the signal-based theories is the winner. Assume that this theory is successful because it defines the physical correlates of linguistic units in a rigorous and quantitative way that makes them independent of context. No matter what instance of phonetic gesture or acoustic pattern we come up with, this theory will tell how to make the measurement that establishes the one-to-one correspondence between the linguistic category and the signal or gesture attributes.

Given this state of affairs, what will the status of "intra-speaker phonetic variation" be? It would be true, would it not, that once a successful new theory emerges, the notorious variability of speech signals would by definition vanish. That variability would turn into epiphenomena arising from our incorrect 1989 way of looking at speech signals. Nonetheless, as soon as we looked at speech as we do today, all the extensive modifications (coarticulation, reduction etc) would still be there and would appear as systematic as they do now. What will successful signal-based theories of invariance have to say about that variability?

Do we really want our theories to succeed in solving the invariance problem without having anything to say about the structure of intra-speaker phonetic variation? No, the systematicity of phonetic variation is very real and demands an explanation.

5.2 CLUES FROM THE TYPOLOGY OF SOUND SYSTEMS

The reality of phonetic variation comes to the fore also when we begin to think about the relationship between on-line phonetic processes and the phonetic structure of lexicons and phonologies. The following findings make that evident.

The UPSID database (Maddieson 1984) lists the vowel and consonant inventories of 317 languages. 58 phonetic dimensions are used to specify several hundred consonant and vowel segment types. For descriptive purposes these elements can be naturally divided into three categories: Basic, Elaborated and Complex segments. To illustrate this classification consider [d d̤ d̻ d̤̻]. The middle two symbols are Elaborated segments in that they show a superposition of retroflexion and breathy voice on the more elementary articulation of the Basic [d]. Any segment with a combination of at least two elaborating processes, e.g. retroflexion and breathy voice, is treated as Complex. Classifying all the segments of UPSID in this way we found a very robust pattern that we have called the **Size Principle** (Lindblom and Maddieson 1988): Small systems use only Basic segments, medium-sized systems use both Basic and Elaborated articulations. Large systems show all three kinds. This principle emerges clearly when the number of B, E and C segments that a language uses is plotted as a function of its inventory size. In systems with six to ten consonants the recruitment of B segments grows in step with system size. Then B segments seem to reach a saturation point at which E segments are brought in growing linearly in number. For inventories with more than 25 consonants, C segments are invoked in addition to the other types. They too exhibit a linear growth with system size. The picture

is similar for the vowels and highly lawful making quantitative size-dependent predictions of the number of B, E and C segments possible.

It appears likely that a **sufficient contrast** constraint underlies the formation of these regularities. In small systems demands for perceptual distinctiveness are less than in larger systems. Articulatorily complex articulations seem to be brought into play only in so far as intrasystemic competition for contrast calls for them.

Clearly there is a parallel between the use of B, E and C segments in these inventories and the use of H&H forms in on-line speech production: A similar tug-of-war and balancing of production-oriented and listener-oriented forces. What mechanism underlies the selection and shaping of phonetic inventories? We cannot seriously address that question here. Suffice it to say that, if natural speech varies along an H&H continuum, then the raw materials are available for building inventories structured according the Size Principle. Elsewhere (Lindblom, MacNeilage and Studdert-Kennedy in preparation) we propose that a process not unlike the variation-selection mechanism of evolutionary biology provides the appropriate model for understanding also the phonological fossilization of on-line phonetic variation.

What do clues from the typology of sound systems tell us about phonetic variation? As we have just seen, they suggest that it is very real. It builds phonologies.

6. References

Bates E (1979): The Emergence of Symbols, Academic Press:New York.

Bladon A (1985): "Diphthongs: A Case. Study of Dynamic Auditory Processing", Speech Communication 4:145-154.

Blumstein S and Stevens K N (1979): "Acoustic Invariance in Speech Production: Evidence from Measurement of the Spectral Characteristics of Stop Consonants", J Acoust Soc Am 72, 43-50.

Blumstein S and Stevens K N (1981): "Phonetic Features and Acoustic Invariance in Speech", Cognition 10, 25-32.

Diehl R L and Kluender K R (1989): "On the Objects of Speech Perception", Ecological Psychology 1(2), 121-144.

Diehl R L, Kluender K R, Walsh M A and Parker E M (in press): "Auditory Enhancement in Speech Perception and Phonology", to appear in Hoffman, R and Palermo, D (eds): Cognition: The State of the Art, LEA:Hillsdale, NJ.

Engstrand O (1988): "Articulatory Correlates of Stress and Speaking Rate in Swedish VCV Utterances", J Acoust Soc Am 83:1863-1875.

Engstrand O and Krull D (1989): "Determinants of Spectral Variation in Spontaneous Speech", pp 88-91 in Proceedings of Speech Research '89, Budapest.

Fant G (1973): Speech Sounds and Features, MIT Press:Cambridge, MA.

Fischer-Jørgensen E (1964): "Sound Duration and Place of Articulation", Zeitschrift für Sprachwissenschaft und Kommunikationsforschung 17:175-207.

Fowler C A (1986): "An Event Approach to the Study of Speech Perception from a Direct-Realist Perspective", J of Phon 14:1, 3-28.

Fujimura O (1989): "Articulatory Perspectives of Speech Organization" lecture presented at the Nato Advanced Institute on Speech Production and Speech Modeling, see this volume.

Fukson O I, Berkinblit A G and Feldman A G (1980): "The Spinal Frog Takes into Account the Scheme of its Body during the Wiping Reflex", Science 209, 1261-1263.

Gay T (1978): "Effect of Speaking Rate on Vowel Formant Movements", J Acoust Soc Am 63(1):223-230.

Gay T, Lindblom B and Lubker J (1981): "Production of Bite-Block Vowels: Acoustic Equivalence by Selective Compensation", J Acoust Soc Am 69(3), 802-810.

Geisler C D (1988): "Representation of Speech Sounds in the Auditory Nerve", J of Phon 16:1, 19-35.

Gibson J J (1972): "Outline of a Theory of Direct Visual Perception", in Royce, J R and Rozeboom, WW (eds): The Psychology of Knowing, Gordon&Breach:New York.

Gibson J J (1979): The Ecological Approach to Visual Perception, Houghton Mifflin:Boston, MA.

Granit R (1979): The Purposive Brain, MIT Press:Cambridge MA.

Greenberg S (1988): Representation of Speech in the Auditory Periphery, J of Phon 16:1-149 (theme issue).

Grillner S (1982): "Possible Analogies in the Control of Innate Motor Acts and the Production of Speech", 217-229 in Grillner S, Lindblom B, Lubker, J and Person, A (eds): Speech Motor Control, Pergamon Press:Oxford.

Hoyt D F and Taylor C R (1981): "Gait and the Energetics of Locomotion in Horses", Nature 292, 239-240.

Johnson A, Sundberg J and Willbrand H (1983): "'Kölning': A Study of Phonation and Articulation in a Type of Swedish Herding Song", 187-202 in Askenfelt A, Felicetti S, Jansson E and Sundberg J (eds): Proc of SMAC 83 (vol 1), Royal Swedish Academy of Music:Stockholm.

Keller E (1989): "Speech Motor Timing", lecture presented at the Nato Advanced Institute on Speech Production and Speech Modeling, see this volume.

Kelso J A S, Saltzman, E L and Tuller, B (1986): "The Dynamical Perspective on Speech Production: Data and Theory", J of Phon 14:1, 29-59.

Klatt D H (1987): "Review of Selected Models of Speech Perception", to be published in Marslen-Wilson W D (ed): Lexical Representation and Process, MIT Press:Cambridge, MA.

Kluender K R, Diehl R L and Wright B A (1988): "Vowel-Length Difference before Voiced and Voiceless Consonants: An Auditory Explanation", J of Phon 16, 153-169.

Kohler K J (1989): "Segmental Reduction in Connected Speech in German: Phonological Facts and Phonetic Explanations", lecture presented at Nato Advanced Institute on Speech Production and Speech Modeling, see this volume.

Kuehn D P and Moll K L (1976): "A Cineradiographic Study of VC and CV Articulatory Velocities", J of Phon 4:303-320.

Lacerda F (1987b): "Effects of Peripheral Auditory Adaptation on the Discrimination of Speech Sounds", dissertation monograph, Perilus VI, Department of Linguistics, Stockholm University.

Lashley K S (1951): "The Problem of Serial Order in Behavior", 112-146 in Jeffress, L A (ed): Cerebral Mechanisms in Behavior, Wiley:New York.

Liberman A M and Mattingly I G (1985): "The Motor Theory of Speech Perception Revised", Cognition 21:1-36.

Liberman A M, Harris K S, Hoffman H S and Griffith B C (1957): "The Discrimination of Speech Sounds within and across Phoneme Boundaries", J of Experimental Psychology 54:358-368.

Lieberman P (1963): "Some Effects of Semantic and Grammatical Context on the Production and Perception of Speech", Language and Speech 6:172-187.

Lindblom B (1963): "Spectrographic Study of Vowel Reduction", J Acoust Soc Am 35:1773-1781.

Lindblom B (1967): "Vowel Duration and a Model of Lip Mandible Coordination", STL-QPSR 4/1967, 1-29, (Department of Speech Communication, RIT, Stockholm).

Lindblom B (1983): "Economy of Speech Gestures", 217-245 in MacNeilage, P.F. (ed): Speech Production, Springer Verlag:New York.

Lindblom B and Sundberg J (1971): "Acoustical Consequences of Lip, Tongue, Jaw and Larynx Movement", J Acoust Soc Am 50(4):1166-1179.

Lindblom B, Pauli S and Sundberg J (1975): "Modeling Coarticulation in Apical Stops", 87-94 in Fant G (ed): Proceedings of the Speech Communication Seminar, Vol. 2, Almqvist&Wiksell:Stockholm.

Lindblom B, and Maddieson I (1988): "Phonetic Universals in Consonant Systems", 62-78 in Hyman L M and Li C N (eds): Language, Speech and Mind, Routledge:London and New York.

Lindblom, B and MacNeilage, P (1986): "Action Theory: Problems and Alternative Approaches", J of Phon 14:1, 117-132.

Lindblom B, Lubker J, Lyberg B, Branderud P and Holmgren K (1987): "The Concept of Target and Speech Timing", 161-182 in Channon R and Shockey L (eds): In Honor of Ilse Lehiste, Foris:Dordrecht, Holland.

Lindblom B and Sundberg J (in prep): Acoustical Consequences of Articulatory Movement.

Lindblom B, MacNeilage P and Studdert-Kennedy M (in prep): Evolution of Spoken Language, Orlando, FL:Academic Press.

Luce P A (1986): Neighborhoods of Words in the Mental Lexicon, Doctoral dissertation, Department of Psychology, Indiana University.

MacNeilage P (1970): "Motor Control of Serial Ordering of Speech", Psychological Review 77:182-196.

Maddieson I (1984): Patterns of Sound, Cambridge University Press:Cambridge.

Miller G A, Galanter, E and Pribram, K (1960): Plans and the Structure of Behavior, Holt, Rinehart & Winston:New York.

Miller J D (1989): "Auditory-Perceptual Interpretation of the Vowel", J Acoust Soc Am 85(5):2114-2133.

Milsum J H (1966): Biological Control Systems Analysis, McGraw-Hill:New York.

Moon S-J and Lindblom, B (1989): "Formant Undershoot in Clear and Citation-Form Speech: A Second Progress Report", 121-123 in STL-QPSR 1/1989, (Dept of Speech Communication, RIT, Stockholm).

Nelson W L (1983): "Physical Principles for Economies of Skilled Movements", Biol Cybernetics 46, 135-147.

Nelson W L, Perkell, J S and Westbury, J R (1984): "Mandible Movements during Increasingly Rapid Articulations of Single Syllables: Preliminary Observations", J Acoust Soc Am 75(3):945-951.

Nord L (1986): "Acoustic Studies of Vowel Reduction in Swedish", 19-36 in STL-QPSR 4/1986, (Dept of Speech Communication, RIT, Stockholm).

Ohala J J (1986): "Against the Direct Realist View of Speech Perception", J of Phonetics 14:1, 75-82.

Öhman S (1966): "Coarticulation in VCV Utterances: Spectrographic Measurements", J Acoust Soc Am 39(1):151-168.

Perkell J and Klatt D (1986): Invariance and Variability in Speech Processes, LEA:Hillsdale, N J.

Rabiner L R (1987): "Use of Spectral Change Information Can Significantly Reduce the Error Rate in Speech Recognition", oral presentation at DARPA meeting at Bolt Beranek and Newman Inc, Cambridge, MA, Nov 1987.

Remez R E, Rubin P E, Pisoni D B and Carrell T D (1981): "Speech Perception without Traditional Speech Cues", Science 212, 947-950.

Risberg A (1979): Bestämning av hörkapacitet och talperceptionsförmåga vid svåra hörselskador, Doctoral dissertation, Royal Institute of Technology, Stockholm.

Schulman R (1989): "Articulatory Dynamics of Loud and Normal Speech", J Acoust Soc Am 85(1):295-312.

Shepard R N (1984): "Ecological Constraints on Internal Representation: Resonant Kinematics of Perceiving, Imagining, Thinking and Dreaming", Psychological Review 91(4), 417-447.

Sherrington C S (1941): Man on His Nature, MacMillan:London.

Stevens K N (1986): "Models of Phonetic Recognition II: A Feature-Based Model of Speech Recognition", 66-67 in Mermelstein P (ed): Proceedings Montreal Satellite Symposium on Speech Recognition, Twelfth International Congress on Acoustics.

Stevens K N (1989): "On the Quantal Nature of Speech", J of Phonetics 17:1/2, 3-45.

Stevens K N and Blumstein S (1978): "Invariant Cues for Place of Articulation in Stop Consonants", J Acoust Soc Am 64, 1358-1368.

Stevens K N and Blumstein S (1981): "The Search for Invariant Acoustic Correlates of Phonetic Features", 1-38 in Eimas P and Miller J (eds): Perspectives on the Studies of Speech, LEA:Hillsdale, N J.

Strange W (1989): "Evolving Theories of Vowel Perception", J Acoust Soc Am 85(5), 2081-2087.

Sundberg J (1987): The Science of the Singing Voice, Northern Illinois University Press:Delkalb, Illinois.

Sundberg J and Lindblom B (1989): "Area Functions for Apical Stops and Some Acoustic Problems", submitted to J Acoust Soc Am.

Sussman H M (1989): "The Representation of Stop Place in Multi-Dimensional Space: A Graphic and Statistical Investigation Consonantal Separability as a Function of Vowel Place", submitted to J Acoust Soc Am.