Speech Processing in the Auditory System: An Overview

Chapter · September 2006

DOI: 10.1007/0-387-21575-1_1

4 authors, including:

Steven Greenberg
University of California, Berkeley
148 PUBLICATIONS **4,930** CITATIONS
SEE PROFILE

Arthur N Popper
University of Maryland, College Park
422 PUBLICATIONS **16,205** CITATIONS
SEE PROFILE

Richard Fay
Loyola University Chicago
214 PUBLICATIONS **7,976** CITATIONS
SEE PROFILE

Chapter 1

Speech Processing in the Auditory System: An Overview

Steven Greenberg and William Ainsworth

Speech Processing in the Auditory System
Steven Greenberg, William Ainsworth,
Arthur Popper and Richard Fay, editors
Springer Handbook of Auditory Research

Contact Information

Steven Greenberg, The Speech Institute, 9 Sereno Circle, Oakland, CA 94619, USA. Tel: (510) 531-1215, Fax: (510) 531-1215, Email: steveng@cogsci.berkeley.edu

William Ainsworth, MacKay Institute for Communication and Neuroscience, Keele University, Keele, Staffordshire, ST5 5BG, United Kingdom (deceased)

1. Introduction

Although our sense of hearing is exploited for many ends, its communicative function stands paramount in our daily lives. Humans are, by nature, a vocal species and it is perhaps not too much of an exaggeration to state that what makes us unique in the animal kingdom is our ability to communicate via the spoken word (Hauser et al. 2002). Virtually all of our social nature is predicated on verbal interaction, and it is likely that this capability has been largely responsible for *Homo sapiens'* rapid evolution over the millennia (Lieberman 1990; Wang 1998). So intricately bound to our nature is language that those who lack it are often treated as less than human (Shattuck 1980).

Our verbal capability is often taken for granted, so seamlessly does it function under virtually all conditions encountered. The intensity of the acoustic background hardly matters – from the hubbub of a cocktail party to the roar of waterfall's descent, humans maintain their ability to verbally interact in a remarkably diverse range of acoustic environments. Only when our sense of hearing falters does the auditory system's masterful role become truly apparent (cf. Edwards, Chapters 7 and Clark, Chapter 8). For under such circumstances the ability to communicate becomes manifestly difficult, if not impossible. Words "blur," merging with other sounds in the background, and it becomes increasingly difficult to keep a specific speaker's voice in focus, particularly in noise or reverberation (cf. Assmann and Summerfield, Chapter 5). Like a machine that suddenly grinds to a halt by dint of a faulty gear, the auditory system's capability of processing speech depends on the integrity of most (if not all) of its working elements.

Clearly, the auditory system performs a remarkable job in converting physical pressure variation into a sequence of meaningful elements composing language. And yet, the process by which this transformation occurs is poorly understood despite decades of intensive investigation.

The role of the auditory system has traditionally been viewed as a frequency analyzer (Ohm 1843; Helmholtz 1863), albeit of limited precision (Plomp 1964), providing a faithful representation of the spectro-temporal properties of the acoustic waveform for higher-level processing. According to Fourier theory any waveform can be decomposed into a series of sinusoidal constituents, which mathematically describe the acoustic waveform (cf. Proakis and Manolakis 1996; Lynn and Fuerst 1998). By this analytical technique it is possible to describe all speech sounds in terms of an energy distribution across frequency and time. Thus, the Fourier spectrum of a typical vowel is composed of a series of sinusoidal components whose frequencies are integral multiples of a common (fundamental) frequency (f_0), and whose amplitudes vary in accordance with the resonance pattern of the associated vocal-tract configuration (cf. Fant 1960; Pickett 1980). The vocal-tract transfer function modifies the glottal spectrum by selectively amplifying energy in certain regions of the spectrum (Fant 1960). These regions of energy maxima are commonly referred to as “formants” (cf. Fant 1960; Stevens 1998). The spectra of non-vocalic sounds, such as stop consonants, affricates and fricatives differ from vowels in a number of ways potentially significant for the manner in which they are encoded in the auditory periphery. These segments typically exhibit formant patterns in which the energy peaks are considerably reduced in magnitude relative to those of vowels. In certain articulatory components, such as the stop release

and frication, the energy distribution is rather diffuse, with only a crude delineation of the underlying formant pattern. In addition, many of these segments are voiceless, their waveforms lacking a clear periodic quality which would otherwise reflect the vibration of the vocal folds of the larynx. The amplitude of such consonantal segments is typically 30 – 50 dB SPL, up to 40 dB less intense than adjacent vocalic segments (Stevens 1998). In addition, the rate of spectral change is generally greater for consonants, and they are usually of brief duration compared to vocalic segments (Avendaño et al., Chapter 2; Diehl and Lindblom, Chapter 3). These differences have significant consequences for the manner in which consonants and vowels are encoded in the auditory system.

Within this traditional framework each word spoken is decomposed into constituent sounds, known as phones (or phonetic segments), each with its own distinctive spectral signature. The auditory system need only encode the spectrum, time-frame by time-frame, in order to provide a complete representation of the speech signal for conversion into meaning by higher cognitive centers. Within this formulation (known as Articulation Theory), speech processing is a matter of frequency analysis and little else (e.g., French and Steinberg 1947; Fletcher and Gault 1950; Pavlovic et al. 1986; Allen 1994). Disruption of the spectral representation, by whatever means, results in phonetic degradation and therefore interferes with the extraction of meaning. This “spectrum-über-alles” framework has been particularly influential in the design of automatic speech recognition systems (cf. Morgan et al., Chapter 6), as well as in the development of

algorithms for the prosthetic amelioration of sensori-neural hearing loss (cf. Edwards, Chapter 7; Clark, Chapter 8).

However, this view of the ear as a mere frequency analyzer is inadequate for describing the auditory system's ability to process speech. Under many conditions its frequency-selective properties bear only a tangential relationship to its ability to convey important information concerning the speech signal, relying rather on the operation of integrative mechanisms to isolate information-laden elements of the speech stream and provide a continuous event stream from which to extract the underlying message. Hence, cocktail party devotees can attest to the fact that far more is involved in decoding the speech signal than merely computing a running spectrum (Bronkhorst 2000). In noisy environments a truly faithful representation of the spectrum could actually serve to hinder the ability to understand due to the presence of background noise or competing speech. It is likely that the auditory system uses very specific strategies to focus on those elements of speech most likely to extract the meaningful components of the acoustic signal (cf. Brown and Cooke 1994; Cooke and Ellis 2001). Computing a running spectrum of the speech signal is a singularly inefficient means to accomplish this objective, as much of the acoustics is extraneous to the message. Instead, the ear has developed the means to extract the information-rich components of the speech signal (and other sounds of biological significance) that may resemble the Fourier spectral representation only in passing.

As the chapters in this volume attest, far more is involved in speech processing than mere frequency analysis. For example, the spectra of speech sounds change over time, sometimes

slowly, but often quickly (Liberman et al. 1956; Pols and van Son 1993; Kewley-Port 1983; van Wieringen and Pols 1994, 1998; Kewley-Port and Neel 2003). These dynamic properties provide information essential for distinguishing among phones. Segments with a rapidly changing spectrum sound very different than those whose spectra modulate much more slowly (e.g., van Wieringen and Pols 1998, 2003).

Thus, the concept of "time" is also important for understanding how speech is processed in the auditory system (cf. Figure 1.1). It is not only the spectrum that changes with time, but also energy. Certain sounds (typically vowels) are far more intense than others (usually consonants). Moreover, it is unusual for a segment's amplitude to remain constant, even over a short interval of time. Such modulation of energy is probably as important as spectral variation (cf. Van Tassell 1987; Drullman et al. 1994a, 1994b; Kollmeier and Koch 1994; Drullman 2003; Shannon et al. 1995), for it provides information crucial for segmentation of the speech signal, particularly at the syllabic level (Greenberg 1996b; Shastri et al. 1999).

Segmentation is a topic rarely discussed in audition, yet is of profound importance for speech processing. The transition from one syllable to the next is marked by appreciable variation in energy across the acoustic spectrum. Such changes in amplitude serve to delimit one linguistic unit from the next, irrespective of spectral properties. Smearing segmentation cues has a profound impact on the ability to understand speech (Drullman et al. 1994a, 1994b; Arai and Greenberg 1998; Greenberg and Arai 1998), far more so than most forms of spectral distortion (Licklider 1951; Miller 1951; Blesser 1972). Thus, the auditory processes involved in coding syllable-length

fluctuations in energy are likely to play a key role in speech processing (Plomp 1983; Drullman et al. 1994a; Grant and Walden 1996a; Greenberg 1996b).

Accompanying modulation of amplitude and spectrum is a variation in fundamental frequency that often spans hundreds, or even thousands of milliseconds (e.g., Ainsworth 1986; Ainsworth and Lindsay 1986; Lehiste 1996). Such f_0 cues are usually associated with prosodic properties such as intonation and stress (Lehiste 1996), but are also relevant to emotion and semantic nuance embedded in an utterance (Williams and Stevens 1972; Lehiste 1996). In addition, such fluctuations in fundamental frequency (and its perceptual correlate, pitch) may be important for distinguishing one speaker from another (e.g., Weber et al. 2002), as well as locking onto to a specific speaker in a crowded environment (e.g., Brokx and Nooteboom 1982; Cooke and Ellis 2001). Moreover, in many languages (e.g., Chinese and Thai), pitch (referred to as "tone") is also used to distinguish among words (Wang 1972), providing yet another context in which the auditory system plays a key role in the processing of speech.

Perhaps the most remarkable quality of speech is its multiplicity. Not only is its spectrum, pitch and amplitude constantly changing, but the variation in these properties occur, to a certain degree, independently of each other, and are decoded by the auditory system in such seamless fashion that we are rarely conscious of the "machinery" underneath the "hood." This multi-tasking capability is perhaps the auditory system's most important capability, the one enabling a rich stream of information to be securely transmitted to the higher cognitive centers of the brain.

Despite the obvious importance of audition for speech communication, the neurophysiological mechanisms responsible for decoding the acoustic signal are not well understood, either in the periphery or in the more central stations of the auditory pathway (cf. Palmer and Shamma, Chapter 4). The enormous diversity of neuronal response properties in the auditory brainstem, thalamus and cortex (cf. Irvine 1986; Popper and Fay 1992; Oertel et al. 2002) are of obvious relevance to the encoding of speech and other communicative signals, but the relation between any specific neuronal response pattern and information contained in the speech signal has not been precisely delineated.

Several factors limit our ability to generalize from brain physiology to speech perception. First, it is not yet possible to record from single neuronal elements in the auditory pathway of humans due to the invasive nature of the recording technology. For this reason, current knowledge concerning the physiology of hearing is largely limited to studies on non-human species lacking linguistic capability. Moreover, most of these physiological studies have been performed on anesthesized, non-behaving animals, rendering the neuronal responses recorded of uncertain relevance to the awake preparation, particularly with respect to the dorsal cochlear nucleus (Rhode and Kettner 1987) and higher auditory stations.

Second, it is inherently difficult to associate the neuronal activity recorded in any single part of the auditory pathway with a specific behavior given the complex nature of decoding spoken language. It is likely that many different regions of the auditory system participate in the analysis

and interpretation of the sound patterns associated with speech, and therefore the conclusions that can be made via recordings from any single neuronal site are limited.

Ultimately, sophisticated brain-imaging technology using such methods as functional magnetic resonance imaging (e.g., Buchsbaum et al. 2001) and magnetoencephalography (e.g., Poeppel et al. 1996) is likely to provide the sort of neurological data capable of answering specific questions concerning the relation between speech decoding and brain mechanisms. Until the maturation of such technology much of our knowledge will necessarily rely on more indirect methods such as perceptual experiments and modeling studies.

One reason why the relationship between speech and auditory function has not been delineated with precision is that, historically, hearing has been largely neglected as an explanatory framework for understanding the structure and function of the speech signal itself. Traditionally, the acoustic properties of speech have been ascribed largely to biomechanical constraints imposed by the vocal apparatus (e.g., Ohala 1983; Lieberman 1984). According to this logic, the tongue, lips and jaw can move only so fast and so far in a given period of time, while the size and shape of the oral cavity set inherent limits on the range of achievable vocal-tract configurations (e.g., Ladefoged 1971; Lindblom 1983; Lieberman 1984).

Although articulatory properties doubtless impose important constraints, it is unlikely that such factors, in and of themselves, can account for the full constellation of spectro-temporal properties of speech. For there are sounds which the vocal apparatus can produce, such as coughing and spitting, which do not occur in any language's phonetic inventory. And while the

vocal tract is capable of chaining long sequences composed exclusively of vowels or consonants together in succession, no language relies on either segmental form alone, nor does speech contain long sequences of acoustically similar elements. And although speech can be readily whispered, it is only occasionally done.

Clearly, factors other than those pertaining to the vocal tract per se, are primarily responsible for the specific properties of the speech signal. One important clue as to the nature of these factors comes from studies of the evolution of the human vocal tract, which anatomically has changed dramatically over the course of the past several hundred thousand years (Lieberman 1984, 1990, 1998). No ape is capable of spoken language, and the vocal repertoire of our closest phylogenetic cousins, the chimpanzees and gorillas, is impoverished relative to that of humans[1] (Lieberman 1984). The implication is that changes in vocal anatomy and physiology observed over the course of human evolution are linked to the dramatic expansion of the brain (cf. Wang 1998), which in turn suggests that a primary selection factor shaping vocal-tract function (Carré and Mrayati 1995) is the capability of transmitting *large* amounts of information *quickly* and *reliably.*

However, this dramatic increase in information transmission has been accompanied by relatively small changes in the anatomy and physiology of the human auditory system. Whereas a quantal leap occurred in vocal capability from ape to human, auditory function has not changed all that much over the same evolutionary period. Given the conservative design of the auditory system across mammalian species (cf. Fay and Popper 1994) it seems likely that the evolutionary innovations responsible for the phylogenetic development of speech were shaped to a significant

degree by anatomical, physiological and functional constraints imposed by the auditory nervous system in its role as transmission route for acoustic information to the higher cortical centers of the brain (cf. Ainsworth 1976; Greenberg 1995, 1996b, 1997a; Greenberg and Ainsworth 2003).

2. How Does the Brain Proceed from Sound to Meaning?

Speech communication involves the transmission of ideas (as well as desires and emotions) from the mind of the speaker to that of the listener via an acoustic (often supplemented by a visual) signal produced by the vocal apparatus of the speaker. The message is generally formulated as a sequence of words chosen from a large, but finite set known to both the speaker and the listener. Each word contains one or more syllables, which are themselves composed of sequences of phonetic elements reflecting the manner in which the constituent sounds are produced. Each phone has a number of distinctive attributes, or features, which encode the manner of production and place of articulation. These features form the acoustic pattern which the listener decodes in order to understand the message.

The process by which the brain proceeds from sound to meaning is not well understood. Traditionally, models of speech perception have assumed that the speech signal is decoded phone by phone, analogous to the manner in which words are represented on the printed page as a sequence of discrete orthographic characters (Klatt 1979; Pisoni and Luce 1987; Goldinger et al. 1996). The sequence of phones thus decoded enables the listener to match the acoustic input to an abstract phone-sequence representation stored in the brain’s mental lexicon. According to this perspective the process of decoding is a straightforward one in which the auditory system performs

a spectral analysis over time that is ultimately associated with an abstract phonetic unit known as the phoneme.

Such sequential models assume that each phone is acoustically realized in comparable fashion from one instance of a word to the next, and that the surrounding context does not affect the manner in which a specific phone is produced. A cursory inspection of a speech signal (e.g., Figure 2.5 in Avendaño et al., Chapter 2) belies this simplistic notion. Thus, the position of a phone within the syllable has a noticeable influence on its acoustic properties. For example, a consonant at the end (coda) of a syllable tends to be shorter than its counterpart in the onset. Moreover, the specific articulatory attributes associated with a phone also vary as a function of its position within the syllable and the word. A consonant at syllable onset is often articulated differently than its segmental counterpart in the coda. For example, voiceless, stop consonants, such as [p], [t] and [k] are usually produced with a complete articulatory constriction ("closure") followed by an abrupt release of oral pressure, whose acoustic signature is a brief (ca. 5-10 ms) transient of broadband energy spanning several octaves (the "release"). However, stop consonants in coda position rarely exhibit such a release. Thus, a [p] at syllable onset often differs substantially from one in the coda (although they share certain features in common, and their differences are largely predictable from context).

The acoustic properties of vocalic segments also vary greatly as a function of segmental context. The vowel [ ] (as in the word "hot") varies dramatically, depending on the identity of the preceding and/or following consonant, particularly with reference to the so-called formant

transitions leading into and out of the vocalic nucleus (cf. Avendaño et al., Chapter 2; Diehl and Lindblom, Chapter 3). Warren (2003) likens the syllable to a “temporal compound” in which the identity of the individual constituent segments are not easily resolvable into independent elements, but rather garner their functional specificity through combination within a larger, holistic entity.

Such context-dependent variability in the acoustics raises a key issue – precisely “where” in the signal does the information associated with a specific phone reside? And is the phone the most appropriate unit with which to decode the speech signal? Or does the “invariant” cues reside at some other level (or levels) of representation?

The perceptual invariance associated with a highly variable acoustic signal has intrigued scientists for many years and remains a topic of intense controversy to this day. The issue of invariance is complicated by other sources of variability in the acoustics, either of environmental origin (e.g., reverberation and background noise), or those associated with differences in speaking style and dialect (i.e., pronunciation variation). There are literally dozens of different ways in which many common words are pronounced (Greenberg 1999), and yet listeners rarely have difficulty understanding the spoken message. And in many environments acoustic reflections can significantly alter the speech signal in such a manner that the canonical cues for many phonetic properties are changed beyond recognition (cf. Figure 5.1 in Assmann and Summerfield, Chapter 5). Given such variability in the acoustic signal how do listeners actually proceed from sound to meaning?

The auditory system may well hold the key for understanding many of the fundamental properties of speech and answer such age-old questions as:

(a) *What* is the information conveyed in the acoustic signal?

(b) *Where* is it located in time and frequency?

(c) *How* is this information encoded in the auditory pathway and other parts of the brain?

(d) What are the mechanisms for *protecting* this information from the potentially deleterious effects of the acoustic background to ensure reliable and accurate transmission?

(e) What are the *consequences* of such mechanisms and the structure of the speech signal for higher-level properties of spoken language?

Based on this information-centric perspective we can generalize from such queries to formulate several additional questions:

(f) To what extent can general auditory processes account for the major properties of speech perception? Can a comprehensive account of spoken language be derived from a purely auditory-centric perspective, or must speech-specific mechanisms (presumably localized in higher cortical centers) be invoked in order to fully account for what is known about human speech processing (e.g., Liberman and Mattingly 1989)?

(g) How does the structure and function of the auditory system shape the spectro-temporal properties of the speech signal?

(h) How can we use knowledge concerning the auditory foundations of spoken language to

benefit humankind?

We shall address these questions during the course of this chapter as a means of providing the background for the remainder of volume.

3. Static versus Dynamic Approaches to Decoding the Speech Signal

As described earlier in this chapter, the traditional approach to spoken language assumes a relatively *static* relationship between segmental identity and the acoustic spectrum. Hence, the spectral cues for the vowel [i^y] ("heat") differ in specific ways from the vowel [ae] ("hat") (cf. Avendaño et al., Chapter 2), the anti-resonance (i.e., spectral zero) associated with an [m] is lower in frequency than that of an [n], and so on. This approach is most successfully applied to a subset of segments such as fricatives, nasals and certain vowels which can be adequately characterized in terms of relatively steady-state spectral properties. However, many segmental classes (such as the stops and diphthongs) are not so easily characterizable in terms of a static spectral profile. Moreover, the situation is complicated by the fact that certain spectral properties associated with a variety of different segments are often vitally dependent on the nature of speech sounds preceding and/or following (referred to as "co-articulation").

3.1 The Motor Theory of Speech Perception

An alternative approach is a dynamic one in which the core information associated with phonetic identity is bound to the movement of the spectrum over time. Such spectral dynamics reflect the movement of the tongue, lips and jaw over time (cf. Aveñdano et al., Chapter 2). Perhaps the

invariant cues in speech are contained in the underlying articulatory gestures associated with the spectrum? If so, then all that would be required is for the brain to back-compute from the acoustics to the original articulatory gestures. This is the essential idea underlying the "motor" theory of speech perception (Liberman et al. 1967; Liberman and Mattingly 1985), which tries to account for the brain's ability to reliably decode the speech signal despite the enormous variability in the acoustics. Although the theory elegantly accounts for a wide range of articulatory and acoustic phenomena (Liberman et al. 1967), it is not entirely clear precisely how the brain proceeds from sound to (articulatory) gesture (but cf. Ivry and Justus 2001; Studdert-Kennedy 2002) on this basis alone. The theory implies (among other things) that those with a speaking disorder should experience difficulty understanding spoken language, which is rarely the case (Lenneberg 1962; Fourcin 1975). Moreover, the theory assumes that articulatory gestures are relatively stable and easily characterizable. However, there is almost as much variability in the production as there is in the acoustics, for there are many different ways of pronouncing words, and even gestures associated with a specific phonetic segment can vary from instance to instance and context to context. Ohala (1994), among others, has criticized production-based perception theories on several grounds: (a) the phonological systems of languages (i.e., their segment inventories and phonotactic patterns) appear to optimize sounds, rather than articulations (cf. Liljencrants and Lindblom 1971; Lindblom 1990); (b) infants, and certain non-human species can discriminate among certain sound contrasts in human speech even though there is no reason to believe they know how to produce these sounds; and (c) humans can differentiate many complex non-speech

sounds such as those associated with music and machines, as well as bird and monkey vocalizations, even though humans are unable to recover the mechanisms producing the sounds.

Ultimately, the motor theory deals with the issue of invariance by displacing the issues concerned with linguistic representation from the acoustics to production without any true resolution of the problem (Kleunder and Greenberg 1989).

3.2 The Locus Equation Model

An approach related to motor theory, but more firmly grounded in acoustics is known as the "locus equation" model (Sussman et al. 1989). Its basic premise is as follows: although the trajectories of formant patterns vary widely as a function of context, they generally "point" to a locus of energy in the spectrum ranging between 500 and 3000 Hz (at least for stop consonants). According to this perspective, it is not the trajectory itself that encodes information, but rather the frequency region thus implied. The locus model assumes some form of auditory extrapolation mechanism capable of discerning endpoints of trajectories in the absence of complete acoustic information (cf. Kleunder and Jenison 1992). While such an assumption falls within the realm of biological plausibility, detailed support for such a mechanism is currently lacking in mammals.

3.3 Quantal Theory

Stevens (1972, 1989) has observed that there is a nonlinear relation between vocal tract configuration and the acoustic output in speech. The oral cavity can undergo considerable change over certain parts of its range without significant alteration in the acoustic signal, while over other parts of the range even small vocal tract changes result in large differences. Stevens suggests that

speech perception takes advantage of this quantal character by categorizing the vocal tract shapes into a number of discrete states for each of several articulatory dimensions (such as voicing, manner and place of articulation), thereby achieving a degree of representational invariance.

4. Amplitude Modulation Patterns

Complementary to the spectral approach is one based on modulation of energy over time. Such modulation occurs in the speech signal at rates ranging between 2 and 6,000. Those of most relevance to speech perception and coding lie between 2 and 2,500 Hz.

4.1 Low-frequency Modulation

At the coarsest level, slow variation in energy reflects articulatory gestures associated with the syllable (Greenberg 1997b, 1999) and possibly the phrase. These low-frequency (2-20 Hz) modulations encode not only information pertaining to syllables but also phonetic segments and articulatory features (Jakobson et al. 1952), by virtue of variation in the modulation pattern across the acoustic spectrum. In this sense the modulation approach is complementary to the spectral perspective. The latter emphasizes energy variation as a function of frequency, while the former focuses on such fluctuations over time.

In the 1930's Dudley applied this basic insight to develop a reasonably successful method for simulating speech using a Vocoder (Dudley 1939). The basic idea is to partition the acoustic spectrum into a relatively small number (20 or fewer) of channels and to capture the amplitude fluctuation patterns in an efficient manner via low-pass filtering of the signal waveform (cf.

Avendaño et al., Chapter 2). Dudley was able to demonstrate that the essential information in speech is encapsulated in modulation patterns lower than 25 Hz distributed over as few as 10 discrete spectral channels. The Vocoder thus demonstrates that much of the detail contained in the speech signal is largely "window dressing" with respect to information required to decode the message contained in the acoustic signal.

Houtgast and Steeneken took Dudley's insight one step further by demonstrating that modulation patterns over a restricted range, between 2 and 10 Hz, can be used as an objective measure of intelligibility (the Speech Transmission Index or STI) for quantitative assessment of speech transmission quality over a wide range of acoustic environments (Houtgast and Steeneken 1973, 1985). Plomp and associates extended application of the STI to clinical assessment of the hearing impaired (e.g., Plomp 1983; Humes et al. 1986; cf. Edwards, Chapter 7).

More recently, Drullman and colleagues have demonstrated a *direct* relationship between the pattern of amplitude variation and the ability to understand spoken language through systematic low-pass filtering of the modulation spectrum in spoken material (Drullman et al. 1994a, 1994b).

The modulation approach is an interesting one from an auditory perspective, as certain types of neurons in the auditory cortex have been shown to respond most effectively to amplitude-modulation rates comparable to those observed in speech (Schreiner and Urbas 1988). Such studies suggest a direct relation between syllable-length units in speech and neural response patterns in the auditory cortex (Greenberg 1996b; Wong and Schreiner 2003). Moreover, human listeners appear to be most sensitive to modulation within this range (Viemeister 1979, 1988).

Thus, the rate at which speech is spoken may reflect not merely biomechanical constraints (cf. Boubana and Maeda 1994) but also an inherent limitation in the capacity of the auditory system to encode information at the cortical level (Greenberg 1996b).

4.2 Fundamental-frequency Modulation

The vocal folds in the larynx vibrate during speech at rates between 75 and 500 Hz, and this phonation pattern is referred to as "voicing." The lower portion of the voicing range (75-175 Hz) is characteristic of adult male speakers, while the upper part of the range (300-500 Hz) is typical of infants and young children. The mid-range (175-300 Hz) is associated with the voice pitch of adult female speakers.

As a function of time, approximately 80% of the speech signal is voiced, with a quasi-periodic, harmonic structure. Among the segments, vowels, liquids ([l] and [r]), glides ([y], [w]) and nasals (m], [n], [ng]) ("sonorants") are almost always voiced (certain languages manifest voiceless liquids, nasals or vowels in certain restricted phonological contexts), while most of the consonantal forms (i.e., stops, fricatives, affricates) can be manifest as either voiced or not (i.e., unvoiced). In such consonantal segments voicing often serves as a phonologically contrastive feature distinguishing among otherwise similarly produced segments (e.g., [p] vs. [b], [s] vs. [z], cf. Diehl and Lindblom, Chapter 3).

In addition to serving as a form of phonological contrast, voice pitch also provides important information about the speaker's gender, age and emotional state. Moreover, much of the prosody in the signal is conveyed by pitch, particularly in terms of fundamental frequency variation over the

phrase and utterance (Halliday 1967). Emotional content is also transmitted in this manner (Mozziconacci 1995), as is grammatical and syntactic information (Bolinger 1986, 1989).

Voice pitch also serves to "bind" the signal into a coherent entity by virtue of common periodicity across the spectrum (Bregman 1990; Langner 1992; Cooke and Ellis 2001). Without this temporal coherence various parts of the spectrum could perceptually fission into separate streams, a situation potentially detrimental to speech communication in noisy environments (cf. Assmann and Summerfield, Chapter 5; Cooke and Ellis 2001).

Voicing also serves to shield much of the spectral information contained in the speech signal from the potentially harmful effects of background noise (see Assmann and Summerfield, Chapter 5). This protective function is afforded by intricate neural mechanisms in the auditory periphery and brainstem synchronized to the fundamental frequency (cf. Section 9). This "phase-locked" response increases the effective signal-to-noise ratio of the neural response by 10-15 dB (Rose et al. 1967; Greenberg 1988) and thereby serves to diminish potential masking effects exerted by background noise.

4.3 Periodicity Associated with Phonetic Timbre and Segmental Identity

The primary vocal-tract resonances of speech range between 225 and 3200 Hz (cf. Avendaño et al., Chapter 2). Although there are additional resonances in the higher frequencies, it is common practice to ignore those above the third formant, as they are generally unimportant from a perceptual perspective, particularly for vowels (Pols et al. 1969; Carlson and Granström 1982; Klatt 1982; Chistovich 1985; Lyon and Shamma 1996). The first formant varies between 225 Hz

(the vowel [i^{y}]) and 800 Hz ([ ]). The second formant ranges between 600 Hz ([‹]) and 2500 ([i^{y}]), while the third formant usually lies in the range of 2500 to 3200 Hz for most vowels (and many consonantal segments).

Strictly speaking, formants are associated exclusively with the vocal-tract resonance pattern and are of equal magnitude. It is difficult to measure formant patterns directly (but cf. Fujimura and Lundquist 1971); therefore speech scientists rely on computational methods and heuristics to estimate the formant pattern from the acoustic signal (cf. Avendaño et al., Chapter 2; Flanagan 1972). The procedure is complicated by the fact that spectral maxima reflect resonances only indirectly (but are referred to as “formants” in the speech literature). This is because the phonation produced by glottal vibration has its own spectral roll-off characteristic (ca. –12 dB/octave) that has to be convolved with that of the vocal tract. Moreover, the radiation property of speech, upon exiting the oral cavity has a +6 dB/octave characteristic that also has to be taken into account. To simplify what is otherwise a very complicated situation, speech scientists generally combine the glottal spectral roll-off with the radiation characteristic, producing a –6 dB/octave roll-off term that is itself convolved with the transfer function of the vocal tract. This means that the amplitude of a spectral peak associated with a formant is essentially determined by its frequency (Fant 1960). Lower-frequency formants are therefore of considerably higher amplitude in the acoustic spectrum than their higher-frequency counterparts. The specific disparity in amplitude can be computed using the –6 dB/octave rolloff approximation described above. There can be as much as a 20-dB difference in sound pressure level between the first and second formants (as in the vowel [i^{y}]).

5. Auditory Scene Analysis and Speech

The auditory system possesses a remarkable ability to distinguish and segregate sounds emanating from a variety of different sources, such as talkers or musical instruments. This capability to filter out extraneous sounds underlies the so-called “cocktail-party” phenomenon in which a listener “filters out” background conversation and non-linguistic sounds to focus on a single speaker's message (cf. von Marlsburg and Schneider 1986). This feat is of particular importance in understanding the auditory foundations of speech processing. Auditory scene analysis refers to the process by which the brain reconstructs the external world through intelligent analysis of acoustic cues and information (cf. Bregman 1990; Cooke and Ellis 2001).

It is difficult to imagine how the ensemble of frequencies associated with a complex acoustic event, such as a speech utterance could be encoded in the auditory pathway purely on the basis of (tonotopically organized) spectral place cues – there are just too many frequency components to track through time. In a manner yet poorly understood, the auditory system utilizes efficient parsing strategies to encode not only information pertaining to a sound's spectrum, but also tracks that signal's acoustic trajectory through time and space, grouping neural activity into singular acoustic events attached to specific sound sources (e.g., Darwin 1981; Cooke 1993).

There is an increasing body of evidence suggesting that neural temporal mechanisms play an important role. Neural discharge synchronized to specific properties of the acoustic signal, such as the glottal periodicity of the waveform (which is typically correlated with the signal's fundamental frequency) as well as onsets (Bregman 1990; Cooke and Ellis 2001), can function to mark activity

as coming from the same source. The operational assumption is that the auditory system, like other sensory systems, has evolved to focus on acoustic events rather than merely performing a frequency analysis of the incoming sound stream. Such relevant signatures of biologically relevant events include common onsets and offsets, coherent modulation and spectral trajectories (Bregman 1990). In other words, the auditory system performs intelligent processing on the incoming sound stream to recreate as best it can the physical scenario from which the sound emanates.

This ecological acoustical approach to auditory function stems from the pioneering work of Gibson (1966, 1979), who considered the senses as intelligent computational resources designed to recreate as much of the external physical world as possible. The Gibsonian perspective emphasizes the deductive capabilities of the senses to infer the conditions behind the sound, utilizing whatever cues are at hand. The limits of hearing capability are ascribed to functional properties interacting with the environment. Sensory systems need not be any more sensitive or discriminating than they need to be in the natural world. Evolutionary processes have assured that the auditory system works sufficiently well under most conditions. The direct realism approach espoused by Fowler (1986, 1996) represents a contemporary version of the ecological approach to speech. We shall return to this issue of intelligent processing in Section 11.

6. Auditory Representations

6.1 Rate-Place Coding of Spectral Peaks

In the auditory periphery the coding of speech and other complex sounds is based on the activity of thousands of auditory-nerve fibers (ANFs) whose tuning characteristics span a broad range in

terms of sensitivity, frequency selectivity and threshold. The excitation pattern associated with speech signals is inferred through recording the discharge activity from hundreds of individual fibers to the same stimulus. In such a “population” study the characteristic (i.e., most sensitive) frequency (CF) and spontaneous activity of the fibers recorded are broadly distributed in a tonotopic manner thought to be representative of the overall tuning properties of the auditory nerve. Through such studies it is possible to infer how much information is contained in the distribution of neural activity across the auditory nerve pertinent to the speech spectrum (cf. Young and Sachs 1979; Palmer and Shamma, Chapter 4).

At low sound pressure levels (< 40 dB) the peaks in the vocalic spectrum are well resolved in the population response with discharge rate roughly proportional to the cochlear-filtered energy level. Increasing the sound pressure level by 20 dB alters the distribution of discharge activity such that the spectral peaks are no longer so prominently resolved in the tonotopic place-rate profile. This is a consequence of the fact that the discharge of fibers with CFs near the formant peaks has saturated relative to those with CFs corresponding to the spectral troughs. As the stimulus intensity is raised still further, to a level typical of conversational speech, the ability to resolve the spectral peaks on the basis of place-rate information is compromised even further.

On the basis of such population profiles, it is difficult to envision how the spectral profile of vowels and other speech sounds could be accurately and reliably encoded on the basis of place-rate information at any but the lowest stimulus intensities. However, a small proportion of AN fibers (15%), with spontaneous (background) rates (SR) less than 0.5 spikes/s, may be capable of

encoding the spectral envelope on the basis of rate-place information, even at the highest stimulus levels (Sachs et al. 1988; Blackburn and Sachs 1990). Such low-SR fibers exhibit extended dynamic response ranges and are more sensitive to the mechanical suppression behavior of the basilar membrane than their higher SR counterparts (Schalk and Sachs 1980; Sokolowski et al. 1989). Thus, the discharge rate of low-SR fibers, with CFs close to the formant peaks, will continue to grow at high sound pressure levels, and the activity of low-SR fibers responsive to the spectral troughs should, in principle, be suppressed by energy associated with the formants. However, such rate suppression also reduces the response to the second and third formants (Sachs and Young 1980), thereby decreasing the resolution of the spectral peaks in the rate-place profile at higher sound pressure levels. For this reason it is not entirely clear that lateral suppression, by itself, actually functions to provide an adequate rate-place representation of speech and other spectrally complex signals in the auditory nerve.

The case for a rate-place code for vocalic stimuli is therefore equivocal at the level of the auditory nerve. The discharge activity of a large majority of fibers is saturated at these levels in response to vocalic stimuli. Only a small proportion of ANFs resolve the spectral peaks across the entire dynamic range of speech. And the representation provided by these low-SR units is less than ideal, particularly at conversational intensity levels (i.e., 75 dB SPL).

The rate-place representation of the spectrum may be enhanced in the cochlear nucleus and higher auditory stations relative to that observed in the auditory nerve. Such enhancement could be a consequence of preferential projection of fibers or through the operation of lateral inhibitory

networks that sharpen still further the contrast between excitatory and background neural activity (Shamma 1985b; Palmer and Shamma, Chapter 4).

Many chopper units in the anteroventral cochlear nucleus respond to steady-state vocalic stimuli in a manner similar to that of low-SR auditory-nerve fibers (Blackburn and Sachs 1990). The rate-place profile of these choppers exhibit clearly delineated peaks at CFs corresponding to the lower formant frequencies, even at 75 dB SPL (Blackburn and Sachs 1990). In principle, a spectral peak would act to suppress the activity of choppers with CFs corresponding to less intense energy, thereby enhancing the neural contrast between spectral maxima and minima. Blackburn and Sachs have proposed that such lateral inhibitory mechanisms may underlie the ability of AVCN choppers to encode the spectral envelope of vocalic stimuli at sound pressure levels well above those at which the average rate of the majority of ANFs saturate. Palmer and Shamma discuss such issues in greater detail in Chapter 4.

The evidence is stronger for a rate-place representation of certain consonantal segments. The amplitude of most voiceless consonants is sufficiently low (< 50 dB SPL) as to evade the rate saturation attendant in the coding of vocalic signals. The spectra of plosive bursts, for example, is generally broadband, with several local maxima. Such spectral information is not likely to be temporally encoded due to its brief duration and the lack of sharply defined peaks. Physiological studies have shown that such segments are adequately represented in the rate-place profile of all spontaneous rate groups of ANFs across the tonotopic axis (e.g., Miller and Sachs 1983; Delgutte and Kiang 1984).

Certain phonetic parameters, such as voice-onset time, are signaled through absolute and relative timing of specific acoustic cues. Such cues are observable in the tonotopic distribution of ANF responses to the initial portion of these segments (Miller and Sachs 1983; Delgutte and Kiang 1984). For example, the articulatory release associated with stop consonants has a broadband spectrum and a rather abrupt onset, which evokes a marked flurry of activity across a wide CF range of fibers. Another burst of activity occurs at the onset of voicing. Because the dynamic range of ANF discharge is much larger during the initial rapid adaptation phase (0-10 ms) of the response, there is relatively little or no saturation of discharge rate during this interval at high sound pressure levels (Sachs et al. 1983; Sinex and Geisler 1983). In consequence, the onset spectra serving to distinguish the stop consonants (Stevens and Blumstein 1978, 1981) are adequately represented in the distribution of rate-place activity across the auditory nerve (Delgutte and Kiang 1984) over the narrow time window associated with articulatory release.

This form of rate information differs from the more traditional “average” rate metric. The underlying parameter governing neural magnitude at onset is actually the probability of discharge over a very short time interval. This probability is usually converted into effective discharge rate normalized to units of spikes per second. If the analysis window (i.e. bin width) is sufficiently short (e.g., 100 μs), the apparent rate can be exceedingly high (up to 10,000 spikes/s). Such high onset rates reflect two properties of the neural discharge – the high probability of firing correlated with stimulus onset and the small degree of variance associated with this first-spike latency. This measure of onset response magnitude is one form of instantaneous discharge rate. Instantaneous,

in this context, refers to the spike rate measured over an interval corresponding to the analysis bin width, which generally ranges between 10 - 1000 μs. This is in contrast to average rate which reflects the magnitude of activity occurring over the entire stimulus duration. Average rate is essentially an integrative measure of activity which counts spikes over relatively long periods of time and weights each point in time equally. Instantaneous rate emphasizes the clustering of spikes over small time windows and is effectively a correlational measure of neural response. Activity which is highly correlated in time, upon repeated presentations will, over certain time intervals, have very high instantaneous rates of discharge. Conversely, poorly correlated response patterns will show much lower peak instantaneous rates whose magnitudes are close to that of the average rate. The distinction between integrative and correlational measures of neural activity is of critical importance for understanding how information in the auditory nerve is ultimately processed by neurons in the higher stations of the auditory pathway.

Place-rate models of spectral coding do not function well in intense background noise. Because the frequency parameter is coded though the spatial position of active neural elements the representation of complex spectra is particularly vulnerable to extraneous interference (Greenberg 1988). Intense noise or background sounds with significant energy in spectral regions containing primary information about the speech signal possess the capability of compromising the auditory representation of the speech spectrum. This vulnerability of place representations is particularly acute when the neural information is represented in the form of average rate. This vulnerability is a consequence of there being no neural marker other than tonotopic affiliation with which to

convey information pertaining to the frequency of the driving signal. In instances where both fore- and background signals are sufficiently intense it will be exceedingly difficult to distinguish that portion of the place representation driven by the target signal from that driven by interfering sounds. Hence, there is no systematic way of separating the neural activity associated with each source purely on the basis of rate-place-encoded information. We shall return to the issue of information coding robustness in Section 9.

The perceptual implications of a strictly rate-place model are counter-intuitive, for it is implied that the intelligibility of speech should decline with increasing sound pressure level above 40 dB. Above this level the rate-place representation of the vocalic spectrum for most AN fibers becomes much less well defined, and only the low-SR fiber population continues to encode the spectral envelope with any degree of precision. In actuality speech intelligibility *improves* above this intensity level at a point where the rate-place representation is not nearly so well delineated.

6.2 Latency-Phase Representations

In a linear system the phase characteristics of a filter are highly correlated with its amplitude response. On the skirts of the filter, where the amplitude response diminishes quickly, the phase of the output signal also changes rapidly. The phase response, by itself, can thus be used in such a system to infer the properties of the filter (cf. Huggins 1952). For a nonlinear system, such as pertains to signal transduction in the cochlea, phase and latency (group delay) information may provide a more accurate estimate of the underlying filter characteristics than average discharge rate because they are not as sensitive to such cochlear nonlinearities as discharge-rate compression and

saturation which typically occur above 40 dB SPL.

Several studies suggest that such phase and latency cues are exhibited in the auditory nerve across a very broad range of intensities. A large phase transition is observed in the neural response distributed across ANFs whose CFs span the lower tonotopic boundary of a dominant frequency component (Anderson et al. 1971), indicating that the high-frequency skirt of the cochlear filters is sharply tuned across intensity. A latency shift of the neural response is observed over a small range of fiber CFs. The magnitude of the shift can be appreciable, as much as half a cycle of the driving frequency (Anderson et al. 1971; Kitzes et al. 1978). For a 500-Hz signal this latency change would be on the order of 1 ms. Because this phase transition may not be subject to the same nonlinearities that result in discharge-rate saturation, fibers with CFs just apical to the place of maximal response can potentially encode a spectral peak in terms of the onset phase across a wide range of intensities.

Interesting variants of this response-latency model have been proposed by Shamma (1985a, 1985b, 1988) and Deng et al. (1988). The phase transition for low-frequency signals should, in principle, occur throughout the entire response, not just at the beginning, as a result of ANFs' phase-locking properties. Such ongoing phase disparities could be registered by some form of neural circuitry presumably located in the cochlear nucleus. The output of such networks would magnify activity in those tonotopic regions over which the phase and/or latency changes rapidly through some form of cross-frequency-channel correlation. In the Shamma model, the correlation is performed through the operation of a lateral inhibitory network, which subtracts the AN output

of adjacent channels. The effect of this cross-channel subtraction is to null out activity for channels with similar phase and latency characteristics, leaving only that portion of the activity pattern where rapid phase transitions occur. The Deng model uses cross-channel correlation (i.e., multiplication) instead of subtraction to locate the response boundaries. Correlation magnifies the activity of channels with similar response patterns and reduces the output of dissimilar adjacent channels. Whether the cross-channel comparison is performed through subtraction, multiplication or some other operation, the consequence of such neural computation is to provide “pointers” to those tonotopic regions where a boundary occurs that might otherwise be hidden if analyzed solely on the basis of average rate. These pointers could, in principle, act in a manner analogous to peaks in the excitation pattern but with the advantage of being preserved across a broad range of sound pressure levels.

6.3 Synchrony-place Information

Place and temporal models of frequency coding are generally discussed as if they are diametrically opposed perspectives. Traditionally, temporal models have de-emphasized tonotopic organization in favor of the fine-temporal structure of the neural response. However, place and temporal coding need not be mutually exclusive. The concept of the central spectrum (Goldstein and Srulovicz 1977; Srulovicz and Goldstein 1983) attempts to reconcile the two approaches through their combination within a single framework for frequency coding. In this model, both place and temporal information are used to construct the peripheral representation of the spectrum. Timing information, as reflected in the interval histogram of ANFs, is used to estimate the driving

frequency. The model assumes that temporal activity is keyed to the tonotopic frequency representation. In some unspecified way, the system "knows" what sort of temporal activity corresponds to each tonotopic location, analogous to a matched filter.

The central spectrum model is the intellectual antecedent of the peripheral representational model of speech proposed by Young and Sachs (1979), whose model is based on the auditory-nerve population response study discussed in Section 6.1. As with place schemes in general, spectral frequency is mapped onto tonotopic place (i.e., ANF characteristic frequency), while the amplitude of each frequency is associated with the magnitude of the neural response synchronized to that component by nerve fibers whose CFs lay within close proximity (1/4 octave). The resulting average localized synchronized rate (ALSR) representation of the stimulus spectrum represents a parsimonious representation of the signal spectrum (cf. Figures 4.5 and 4.7 in Palmer and Shamma, Chapter 4). The ALSR is a computational procedure for estimating the magnitude of neural response in a given frequency channel based on the product of firing rate and temporal correlation with a pre-defined frequency band. The spectral peaks associated with the three lower formants (F_1, F_2 F_3) are clearly delineated in the ALSR representation, in marked contrast to the rate-place representation.

The mechanism underlying the ALSR representation is referred to as "synchrony suppression" or "synchrony capture." At low sound pressure levels, temporal activity synchronized to a single low-frequency (< 4 kHz) spectral component is generally restricted to a circumscribed tonotopic region close to that frequency. Increasing the sound pressure level results in a spread of the

synchronized activity, particularly towards the region of high-CF fibers. In this instance, the spread of temporal activity occurs in roughly tandem relation with the activation of fibers in terms of average discharge rate. At high sound pressure levels (ca. 70-80 dB) a large majority of ANFs with CFs below 10 kHz are phase-locked to low-frequency components of the spectrum. This upward spread of excitation into the high-frequency portion of the auditory nerve is a consequence of the unique filter characteristics of high-CF mammalian nerve fibers. Although the filter function for such units is sharply bandpass within 20-30 dB of rate threshold, it becomes broadly tuned and low-pass at high sound pressure levels. This tail component of the high-CF fiber frequency-threshold curve (FTC) renders such fibers extremely responsive to low-frequency signals at sound pressure levels typical of conversational speech. The consequence of this low-frequency sensitivity, in concert with the diminished selectivity of low-CF fibers, is the orderly basal recruitment (toward the high-frequency end of the auditory nerve) of ANFs as a function of increasing sound pressure level.

Synchrony suppression is intricately related to the frequency selectivity of auditory-nerve fibers. At low sound pressure levels, most low-CF nerve fibers are phase-locked to components in the vicinity of their CF. At this sound pressure level the magnitude of a fiber's response, measured in terms of either synchronized or average rate, is approximately proportional to the signal energy at the unit CF, resulting in rate-place and synchrony-place profiles relatively isomorphic to the input stimulus spectrum. At higher sound pressure levels, the average-rate response saturates across the tonotopic array of nerve fibers, resulting in significant degradation of the rate-place

representation of the formant pattern, as described above. The distribution of temporal activity also changes, but in a somewhat different manner. The activity of fibers with CFs near the spectral peaks remain phase-locked to the formant frequencies. Fibers whose CFs lie in the spectral valleys, particularly between F_1 and F_2, become synchronized to a different frequency, most typically F_1.

The basis for this suppression of synchrony may be as follows: the amplitude of components in the formant region (particularly F_1) are typically 20 to 40 dB greater than that of harmonics in the valleys. When the amplitude of the formant becomes sufficiently intense, its energy “spills” over into neighboring frequency channels as a consequence of the broad tuning of low-frequency fibers referred to above. Because of the large amplitude disparity between spectral peak and valley, there is now more formant-related energy passing through the fiber's filter than energy derived from components in the CF region of the spectrum. Suppression of the original timing pattern actually begins when the amount of formant-related energy equals that of the original signal. Virtually complete suppression of the less intense signal results when the amplitude disparity is greater than 15 dB (Greenberg et al. 1986). In this sense, encoding frequency in terms of neural phase-locking acts to enhance the peaks of the spectrum at the expense of less intense components.

The result of this synchrony suppression is to reduce the amount of activity phase-locked to frequencies other than the formants. At higher sound pressure levels, the activity of fibers with CFs in the spectral valleys are indeed phase-locked, but to frequencies distant from their CFs. In the ALSR model the response of these units contributes to the auditory representation of the signal spectrum only in an indirect fashion, since the magnitude of temporal activity is measured only for

frequencies near the fiber CF. In this model, only a small subset of ANFs, with CFs near the formant peaks, directly contribute to the auditory representation of the speech spectrum in the model.

6.4 Cortical Representations of the Speech Signal

Neurons do not appear to phase-lock to frequencies above 200-300 Hz above the level of inferior colliculus, implying that spectral information based on timing information in the peripheral and brainstem regions of the auditory pathway is transformed into some other representation in the auditory cortex. Moreover, most auditory cortical neurons respond at very low discharge rates, typically less than 10 spikes/s. It is not uncommon for units at this level of the auditory pathway to respond only once per acoustic event, with the spike associated with stimulus onset.

Shamma and colleagues describe recent work from their laboratory in Chapter 4 that potentially resolves some of the issues discussed earlier in Section 6. Most of the responsiveness observed at this level of the auditory pathway appears to be associated with low-frequency properties of the spectrally filtered waveform envelope, suggesting a neural basis for the perceptual and synthesis studies described in Section 4. In this sense, the cortex appears to be concerned primarily with events occurring over much longer time spans than those of the brainstem and periphery.

7. Functional Properties of Hearing Relevant to Speech

For many applications, such as speech analysis, syntheses and coding, it is useful to know the

perceptual limits pertaining to speech sounds. For example, how accurately do we need to specify the frequency or amplitude of a formant for such applications? Such functional limits can be estimated using psychophysical techniques.

7.1 Audibility and Dynamic Range Sensitivity

The human ear responds to frequencies between 30 and 20,000 Hz, and is most sensitive between 2.5 and 5 kHz (Wiener and Ross 1946). The upper limit of 20 kHz is an average for young adults with normal hearing. As individuals age, sensitivity to high-frequencies diminishes, so much so that by the age of 60, it is unusual for a listener to hear frequencies above 12 kHz. Below 400 Hz sensitivity decreases dramatically. The threshold of detectability at 100 Hz is ca. 30 dB higher (i.e., less sensitive) than at 1 kHz. Above 5 kHz sensitivity declines steeply as well. Most of the energy in the speech signal lies below 2 kHz (Figure 5.1 in Assmann and Summerfield, Chapter 5). The peak in the *average* speech spectrum is ca. 500 Hz, falling off at ca. 6 dB/octave thereafter (Figure 5.1 in Assmann and Summerfield, Chapter 5). There is relatively little energy of informational relevance above 10 kHz in the speech signal. Thus, there is a relatively good match between the spectral energy profile in speech and human audibility. Formant peaks in the very low frequencies are high in magnitude, largely compensating for the decreased sensitivity in this portion of the spectrum. Higher-frequency formants are of lower amplitude but occur in the most sensitive part of the hearing range. Thus, the shape of the speech spectrum is remarkably well adapted to the human audibility curve.

Normal-hearing listeners can generally detect sounds as low as –10 dB SPL in the most sensitive part of the spectrum (ca. 4 kHz) and are capable of withstanding sound pressure levels of 110 dB without experiencing pain. Thus, the human ear is capable of transducing ca. 120-dB (1:1,000,000) dynamic range of sound pressure under normal-hearing conditions. The sound pressure level (SPL) of the most intense speech sounds (usually vowels) generally lies between 70 and 85 dB, while the SPL of certain consonants (e.g., fricatives) can be as low as 35 dB. The dynamic range of speech sounds is therefore ca. 50 dB (this estimate of sound pressure level applies to the entire segment. Prior to initiation of a speech gesture there is little or no energy produced, so the *true* dynamic range of the speech signal from instant to instant is probably ca. 90 dB).

Within this enormous range the ability to discriminate fluctuations in intensity (ΔI) varies. At low sound pressure levels (< 40 dB) the difference limen (DL) lies between 1 and 2 dB (Riesz 1928; Jesteadt et al. 1977; Viemeister 1988). Above this limit the DL can decline appreciably (i.e., discriminability improves) to about half of this value (Greenwood 1994). Thus, within the core range of the speech spectrum, listeners are exceedingly sensitive to variation in intensity. Flanagan (1957) estimated that ΔI for formants in the speech signal to be ca. 2 dB.

7.2 Frequency Discrimination and Speech

Human listeners can distinguish exceedingly fine differences in frequency for sinusoids and other narrow-band signals. At 1 kHz the frequency DL (Δf) for such signals can be as small as 1 to 2 Hz (i.e., 0.1 - 0.2%) (Wier et al. 1977). However, Δf varies as a function of frequency, sound pressure

level and duration. Frequency discriminability is most acute in the range between 500 and 1000 Hz, and falls dramatically at high frequencies (> 4 kHz), particularly when the signal-to-noise ratio is held constant (Dye and Hafter 1980). Thus, discriminability is finest for those parts of the spectrum in which most of the information in the speech spectrum resides. With respect to duration, frequency discriminability is most acute for signals longer than 80-100 ms (at any frequency), and signals greater than 40 dB SPL are generally more finely discriminated in terms of frequency than those of lower intensity (Wier et al. 1977).

The discriminability of broadband signals, such as formants in a speech signal, is not nearly as fine as for narrowband stimuli. In an early study Flanagan (1955) found that Δf ranged between 3 and 5% of the formant frequency for steady-state stimuli. More recent studies indicate that Δf can be as low as 1% when listeners are highly trained (Kewley-Port and Watson 1994). Still, the DL for formant frequency appears to be an order of magnitude greater for formants than for sinusoidal signals.

Of potentially greater relevance for speech perception is discriminability of non-steady-state formants, which possess certain properties analogous to formant transitions interposed between consonants and vowels. Mermelstein (1978) estimated that the DL for formant transitions ranges between 49-70 Hz for F_1 and 171-199 Hz for F_2. A more recent study by van Wieringen and Pols (1994) found that the DL is sensitive to the rate and duration of the transition. For example, the DL is ca. 70 Hz for F_1 when the transition is 20 ms, but decreases (i.e., improves) to 58 Hz when transition duration is increased to 50 ms.

Clearly, the ability to distinguish fine gradations in frequency is much poorer for complex signals, such as speech formants, relative to spectrally simple signals such as sinusoids. At first glance such a relation may appear puzzling, as complex signals provide more "opportunities" for comparing details of the signal than simple ones. However, from an information-theoretic perspective this diminution of frequency discriminability could be of utility for a system that generalizes from signal input to a finite set of classes through a process of learned association, a topic that is discussed further in Section 11.

8. The Relation Between Spectro-temporal Detail and Channel Capacity

It is important for any information-rich system that the information carrier be efficiently and reliably encoded. For this reason a considerable amount of research has been performed over the past century on efficient methods of coding speech (cf. Avendaño et al., Chapter 2). This issue was of particular concern for analog telephone systems in which channel capacity was severely limited (in the era of digital communications, channel capacity is much less of a concern for voice transmission, except for wireless communication – e.g., cell phones). Pioneering studies by Harvey Fletcher and associates at Bell Laboratories,[2] starting in the 1910's, systematically investigated the factors limiting intelligibility as a means of determining how to reduce the bandwidth of the speech signal without compromising the ability to communicate using the telephone (cf. Fletcher 1953).

In essence, Fletcher's studies were directed towards determining the information-laden regions of the spectrum. Although information theory had yet to be mathematically formulated (Shannon's paper on the mathematical foundation of information theory was originally published in the Bell

System Technical Journal, and was issued in book form the following year – Shannon and Weaver 1949), it was clear to Fletcher that the ability to decode the speech signal into constituent sounds could be used as a quantitative means of estimating the amount of information contained. Over a period of twenty years various band-limiting experiments were performed in an effort to ascertain the frequency limits of information contained in speech (Miller 1951; Fletcher 1953; Allen 1994). The results of these studies were used to define the bandwidth of the telephone (300-3400 Hz), a standard still in use today. Although there is information in the frequency spectrum residing outside these limits, Fletcher's studies revealed that its absence did not significantly impair verbal interaction and could therefore be tolerated over the telephone.

More recent work has focused on delineating the location of information contained in both frequency and time. Spectral maxima associated with the three lowest formants are known to carry much of the timbre information associated with vowels and other phonetic classes (e.g., Ladefoged 1967, 2001; Pols et al. 1969). However, studies using "sine-wave" speech, suggest that spectral maxima, *in and of themselves*, are not the ultimate carriers of information in the signal. The speech spectrum can be reduced to a series of three sinusoids, each associated with the center frequency of a formant (Remez et al. 1981, 1994). When played, this stimulus sounds extremely unnatural and is difficult to understand without prior knowledge of the words spoken.[3] In fact, Kakusho and colleagues demonstrated many years ago that for such a sparse spectral representation to sound speech-like and be identified reliably, each spectral component in this sparse representation must be coherently amplitude-modulated at a rate within the voice-pitch range (Kakusho et al. 1971).

This finding is consistent with the notion that the auditory system requires complex spectra, preferably with glottal periodicity, to associate the signal with information relevant to speech (whispered speech lacks a glottal excitation source, yet is comprehensible. However, such speech is extremely fragile, vulnerable to any sort of background noise and is rarely used except in circumstances where secrecy is of paramount concern or vocal pathology has intervened).

Less radical attempts to reduce the spectrum have proven highly successful. For example, smoothing the spectral envelope to minimize fine detail in the spectrum is a common technique used in digital coding of speech (cf. Avendaño et al., Chapter 2), a result consistent with the notion that some property associated with spectral maxima is important, even if it is not the absolute peak by itself (cf. Assmann and Summerfield, Chapter 5). Such spectral envelope smoothing has been successfully applied to automatic speech recognition as a means of reducing extraneous detail for enhanced acoustic-phonetic pattern classification (cf. Davis and Mermelstein 1980; Ainsworth 1988; Hermansky 1990; Morgan et al., Chapter 6). And perceptual studies, in which the depth and detail of the spectral envelope is systematically manipulated, have demonstrated the importance of such information for speech intelligibility both in normal and hearing-impaired individuals (ter Keurs et al. 1992, 1993; Baer and Moore 1993).

Intelligibility can remain high even when much of the spectrum is eliminated in such a manner as to discard many of the spectral peaks in the signal. As few as four band-limited (1/3 octave) channels distributed across the spectrum, *irrespective of the location of spectral maxima*, can provide nearly perfect intelligibility of spoken sentences (Greenberg et al. 1998). Perhaps the

spectral peaks, in and of themselves, are not as important as functional *contrast* across frequency and over time (cf. Lippmann 1996; Müsch and Buus 2001b).

How is such information extracted from the speech signal? Everything we know about speech suggests that the mechanisms responsible for decoding the signal must operate over relatively long intervals of time, between 50 and 1000 ms (if not longer), which are characteristic of cortical rather than brainstem or peripheral processing (Greenberg 1996b). At the cortical level auditory neurons respond relatively infrequently and this response is usually associated with the onset of discrete events (cf. Section 6.4; Palmer and Shamma, Chapter 4). It is as if cortical neurons respond primarily to truly informative features in the signal and otherwise remain silent. A potential analog of cortical speech processing is the highly complex response patterns observed in the auditory cortex of certain echo-locating bats in response to target-ranging or Döppler-shifted signals (Suga et al. 1995; Suga 2003). Many auditory cortical neurons in such species as *Pteronotus parnellii* require specific combinations of spectral components distributed over frequency and/or time in order to fire (Suga et al. 1983). Perhaps comparable "combination-sensitive" neurons function in human auditory cortex (Suga 2003).

If it is mainly at the level of the cortex that information relevant to speech features is extracted, what role is played by more peripheral stations in the auditory pathway?

9. Protecting Information Contained in the Speech Signal

Under many conditions speech (and other communication signals) is transmitted in the presence of background noise and/or reverberation. The sound pressure level of this background can be

considerable and thus poses a considerable challenge to any receiver intent on decoding the message contained in the foreground signal. The problem for the receiver, then, is not just to decode the message, but also to do so in the presence of variable and often unpredictable acoustic environments. In order to accomplish this objective highly sophisticated mechanisms must reside in the brain that effectively "shield" the message in the signal.

This informational shielding is largely performed in the auditory periphery and central brainstem regions. In the periphery are mechanisms that serve to enhance spectral peaks, both in quiet and in noise. Such mechanisms rely on automatic gain control (AGC), as well as mechanical and neural suppression of those portions of spectrum distinct from the peaks (cf. Rhode and Greenberg 1994; Palmer and Shamma, Chapter 4). The functional consequence of such spectral-peak enhancement is the capability of preserving the general shape of the spectrum over a wide range of background conditions and signal-to-noise ratios (SNRs).

In the cochlea are several mechanisms operating to preserve the shape of the spectrum. Mechanical suppression observed in the basilar membrane response to complex signals at high sound pressure levels serves to limit the impact of those portions of the spectrum significantly below the peaks, effectively acting as a peak clipper. This form of suppression appears to be enhanced under noisy conditions (Rhode and Greenberg 1994), and is potentially mediated through the olivo-cochlear bundle (Liberman 1988; Warr 1992; Reiter and Liberman 1995) passing from the brainstem down into the cochlea itself.

A second means with which to encode and preserve the shape of spectrum is through the spatial frequency analysis performed in the cochlea (cf. Greenberg 1996a; Palmer and Shamma, Chapter 4; Section 6 of this chapter). As a consequence of the stiffness gradient of the basilar membrane, its basal portion is most sensitive to high frequencies (>10 kHz) while the apical end is most responsive to frequencies below 500 Hz. Frequencies in between are localized to intermediate positions in the cochlea in a roughly logarithmic manner (for frequencies greater than 1 kHz). In the human cochlea approximately 50% of the 35-mm length of the basilar membrane is devoted to frequencies below 2,000 kHz (Greenwood 1961, 1990) suggesting that the spectrum of the speech signal has been tailored, at least in part, to take advantage of the considerable amount of neural "real estate" devoted to low-frequency signals.

The frequency analysis performed by the cochlea appears to be quantized with a resolution of approximately 1/4-octave. Within this "critical band" (Fletcher 1953; Zwicker et al. 1957) energy is quasi-linearly integrated with respect to loudness summation and masking capability (Scharf 1970). In many ways the frequency analysis performed in the cochlea behaves as if the spectrum is decomposed into separate (and partially independent) channels. This sort of spectral decomposition provides an effective means of protecting the most intense portions of the spectrum from background noise under conditions.

A third mechanism preserving spectral shape is based on neural phase-locking, whose origins arise in the cochlea. The release of neurotransmitter in inner hair cells (IHCs) is temporally modulated by the stimulating (cochlear) waveform and results in a temporal patterning of auditory-

nerve-fiber responses that is “phase-locked” to certain properties of the stimulus. The effectiveness of this response modulation depends on the ratio of the AC (alternating current) to the DC (direct current) components of the inner hair cell (IHC) receptor potential, which begins to diminish for signals greater than 800 Hz. Above 3 kHz, the AC/DC ratio is sufficiently low that the magnitude of phase-locking is negligible (cf. Greenberg 1996a for further details). Phase-locking is thus capable of providing an effective means of temporally coding information pertaining to the first, second and third formants of the speech signal (Young and Sachs 1979). But there is more to phase-locking than mere frequency coding.

Auditory-nerve fibers generally phase-lock to the portion of the local spectrum of greatest magnitude through a combination of automatic gain control (Geisler and Greenberg 1986; Greenberg et al. 1986) and a limited dynamic range of ca. 15 dB (Greenberg et al. 1986; Greenberg 1988). Because ANFs phase-lock poorly (if at all) to noise, signals with a coherent temporal structure (e.g., harmonics) are relatively immune to moderate amounts of background noise. The temporal patterning of the signal insures that peaks in the foreground signal rise well above the average noise level at all but the lowest SNRs. Phase-locking to those peaks riding above the background effectively suppresses the noise (cf. Greenberg 1996a).

Moreover, such phase-locking enhances the effective SNR of the spectral peaks through a separate mechanism which distributes the temporal information across many neural elements. The ANF response is effectively “labeled” with the stimulating frequency by virtue of the temporal properties of the neural discharge. At moderate-to-high sound pressure levels (40-80 dB) the

number of ANFs phase-locked to the first formant grows rapidly, so that it is not just fibers most sensitive to the first formant that respond. Fibers with characteristic (i.e., most sensitive) frequencies as high as several octaves above F_1 may also phase-lock to this frequency region (cf. Young and Sachs 1979; Jenison et al. 1991). In this sense, the auditory periphery is exploiting redundancy in the neural timing pattern distributed across the cochlear partition to robustly encode information associated with spectral peaks. Such a distributed representation renders the information far less vulnerable to background noise (Ghitza 1988; Greenberg 1988) and also provides an indirect measure of peak magnitude via determining the number of auditory channels that are coherently phase-locked to that frequency (cf. Ghitza 1988).

This phase-locked information is preserved to a large degree in the cochlear nucleus and medial superior olive. However, at the level of the inferior colliculus it is rare for neurons to phase-lock to frequencies above 1000 Hz. At this level the temporal information has probably been recoded, perhaps in the form of spatial modulation maps (Langner and Schreiner 1988; Langner 1992).

Phase-locking provides yet a separate means of protecting spectral peak information through binaural cross-correlation. The phase-locked input from each ear meets in the medial superior olive, where it is likely that some form of cross correlational analysis is computed. Additional correlational analyses are performed in the inferior colliculus (and possibly the lateral lemniscus). Such binaural processing provides a separate means of increasing the effective SNR, by weighting

that portion of the spectrum which is binaurally coherent across the two ears (cf. Stern and Trahoitis 1995; Blauert 1996).

Yet a separate means of shielding information in speech is through temporal coding of the signal's fundamental frequency (f_0). Neurons in the auditory periphery and brainstem nuclei can phase-lock to the signal's f_0 under many conditions, thus serving to bind the discharge patterns associated with different regions of the spectrum into a coherent entity, as well as enhance the SNR via phase-locking mechanisms described above. Moreover, fundamental-frequency variation can serve, under appropriate circumstances, as a parsing cue, both at the syllabic and phrasal levels (Brokx and Nooteboom 1982; Ainsworth 1986; Bregman 1990; Darwin and Carlyon 1995; Assmann and Summerfield, Chapter 5). Thus, pitch cues can serve to guide the segmentation of the speech signal, even under relatively low SNRs.

10. When Hearing Fails

The elaborate physiological and biochemical machinery associated with acoustic transduction in the auditory periphery may fail, thus providing a "natural" experiment with which to ascertain the specific role played by various cochlear structures in the encoding of speech. Hearing impairment also provides a method with which to estimate the relative contributions made by "bottom-up" and "top-down" processing for speech understanding (Grant and Walden 1995; Grant and Seitz 1998; Grant et al. 1998).

There are two primary forms of hearing impairment that affect the ability to decode the speech signal.

Conductive hearing loss is usually the result of a mechanical problem in the middle ear, with attendant (and relatively uniform) loss of sensitivity across much of the frequency spectrum. This form of conductive impairment can often be ameliorated through surgical intervention.

Sensori-neural loss originates in the cochlea and has far more serious consequences for speech communication. The problem lies primarily in the outer hair cells (OHCs) which can be permanently damaged as a result of excessive exposure to intense sound (cf. Bohne and Harding 2000; Patuzzi 2002). OHC stereocilia indirectly affect the sensitivity and tuning of IHCs via their articulation with the underside of the tectorial membrane (TM). Their mode of contact directly affects the TM's angle of orientation with respect to the IHC stereocilia and hence can reduce the ability to induce excitation in the IHCs via deflection of their stereocilia (probably through fluid coupling rather than direct physical contact). After exposure to excessive levels of sound the cross-linkages of actin in OHC stereocilia are broken or otherwise damaged, resulting in ciliary floppiness that reduces OHC sensitivity substantially and thereby also reduces sensitivity in the IHCs (cf. Gummer et al. 1996, 2002). In severe trauma the stereocilia of the IHCs are also affected. Over time both the OHCs and IHCs of the affected frequency region are likely to degenerate, making it impossible to stimulate ANFs innervating this portion of the cochlea. Eventually, the ANFs themselves lose their functional capacity and whither, which in turn can result in degeneration of neurons further upstream in the central brainstem pathway and cortex (cf. Gravel and Ruben 1996).

When the degree of sensori-neural impairment is modest it is possible to partially compensate for the damage through the use of a hearing aid (Edwards, Chapter 7). The basic premise of a hearing aid is that audibility has been compromised in selected frequency regions, thus requiring some form of amplification to raise the level of the signal to audible levels (Steinberg and Gardner 1937). However, it is clear from recent studies among the hearing impaired that audibility is not the only problem. Such individuals also manifest under many (but not all) circumstances a significant reduction in frequency and temporal resolving power (cf. Edwards, Chapter 7).

A separate, but related problem concerns a drastic decrease in dynamic range of intensity coding. Because the threshold of neural response is significantly elevated, without an attendant increase in the upper limit of sound pressure transduction, the effective range between the softest and most intense signals is severely compressed. This reduction in dynamic range means that the auditory system is no longer capable of using energy modulation for reliable segmentation in the affected regions of the spectrum, and therefore makes the task of parsing the speech signal far more difficult.

Modern hearing aids attempt to compensate for this dynamic-range reduction through frequency-selective compression. Using sophisticated signal-processing techniques, a 50-dB range in the signal's intensity can be "squeezed" into a 20-dB range as a means of simulating the full dynamic range associated with the speech signal. However, such compression only partially compensates for the hearing impairment, and does not fully restore the patient's ability to understand speech in noisy and reverberant environments (cf. Edwards, Chapter 7).

What other factors may be involved in the hearing-impaired's inability to reliably decode the speech signal? One potential clue is encapsulated in the central paradox of sensori-neural hearing loss. Although most of the energy (and information) in the speech signal lies *below* 2 kHz, most of the impairment in the clinical population is *above* 2 kHz. In quiet, the hearing impaired rarely experience difficulty understanding speech. However, in noisy and reverberant conditions, the ability to comprehend speech completely falls apart (without some form of hearing aid or speechreading cues).

This situation suggests that there is information in the mid- and high-frequency region of the spectrum that is of the utmost importance under acoustic-interference conditions. In quiet, the speech spectrum below 2 kHz can provide sufficient cues to adequately decode the signal. In noise and reverberation the situation changes drastically, since most of the energy produced by such interference is also in the low-frequency range. Thus, the effective signal-to-noise ratio in the portion of the spectrum where hearing function is relatively normal is reduced to the point where information from other regions of the spectrum are required to supplement and disambiguate the speech cues associated with the low-frequency spectrum.

There is some evidence to suggest that normal-hearing individuals do indeed utilize a spectrally adaptive process for decoding speech. Temporal scrambling of the spectrum via desynchonization of narrowband (1/3 octave) channels distributed over the speech range simulates certain properties of reverberation. When the channels are desynchronized by modest amounts, the intelligibility of spoken sentences remains relatively high. As the amount of asynchrony across

channels increases, intelligibility falls. The rate at which intelligibility decreases is consistent with the hypothesis that for small degrees of cross-spectral asynchrony (i.e., weak reverberation) the lower parts of the spectrum (< 1,500 Hz) are responsible for most of the intelligibility performance, while for large amounts of asynchrony (i.e., strong reverberation) it is channels above 1,500 Hz that are most highly correlated with intelligibility performance (Arai and Greenberg 1998; Greenberg and Arai 1998). This result is consistent with the finding that the best single psychoacoustic (non-speech) predictor of speech intelligibility capability in quiet is the pure-tone threshold *below* 2 kHz, while the best predictor of speech intelligibility in noise is the pure-tone threshold *above* 2 kHz (Smoorenburg 1992; but cf. Festen and Plomp 1981 for an alternative perspective).

What sort of information is contained in the high-frequency portion of the spectrum that could account for this otherwise paradoxical result? There are two likely possibilities. The first pertains to articulatory place-of-articulation, information that distinguishes, for example, a [p] from [t] and [k]. The locus of maximum articulatory constriction produces an acoustic “signature” that requires reliable decoding of the entire spectrum between 500 and 3,500 Hz (Stevens and Blumstein 1978, 1981). Place-of-articulation cues are particularly vulnerable to background noise (Miller and Nicely 1955; Wang and Bilger 1973) and removal of any significant portion of the spectrum is likely to degrade the ability to identify consonants on this articulatory dimension. Place-of-articulation is perhaps the single most important acoustic feature dimension for distinguishing among words, particularly at word onset (Rabinowitz et al. 1992; Greenberg and Chang 2000). It

is therefore not surprising that much of the problem the hearing impaired manifest with respect to speech decoding pertains to place-of-articulation cues (Dubno and Dirks 1989; Dubno and Schaefer 1995).

A second property of speech associated with the mid- and high-frequency channels is prosodic in nature. Grant and Walden (1996) have shown the portion of the spectrum above 3 kHz provides the most reliable information concerning the number of syllables in an utterance. It is also likely that these high-frequency channels provide reliable information pertaining to syllable boundaries (Shastri et al. 1999). To the extent that this sort of knowledge is important for decoding the speech signal, the high-frequency channels can provide information that supplements that of the low-frequency spectrum. Clearly, additional research is required to more fully understand the contribution made by each part of the spectrum to the speech-decoding process.

Grant and colleagues (Grant and Walden 1995; Grant and Seitz 1998; Grant et al. 1998) estimate that about two-thirds of the information required to decode spoken material (in this instance sentences) is bottom-up in nature, derived from detailed phonetic and prosodic cues. Top-down information concerned with semantic and grammatical context accounts for perhaps a third of the processing involved. The relative importance of the spectro-temporal detail for understanding spoken language is certainly consistent with the communication handicap experienced by the hearing impaired.

In cases where there is little hearing function left in any portion of the spectrum, a hearing aid is of little use to the patient. Under such circumstances a more drastic solution is required, namely

implantation into the cochlea of an electrode array capable of direct stimulation of the auditory nerve (Clark 2003; Clark, Chapter 8). Over the past twenty-five years the technology associated with cochlear implants has progressed dramatically. Whereas in the early 1980's such implants were rarely capable of providing more than modest amelioration of the communication handicap associated with profound deafness, today there are many thousands who communicate at near normal levels, both in face-to-face interaction and (in the most successful cases) over the telephone (i.e., unaided by visible speechreading cues) using such technology. The technology has been particularly effective for young children who have been able to grow up using spoken language to a degree that would have been unimaginable twenty years ago.

The conceptual basis of cochlear-implant technology is simple (although the surgical and technical implementation is dauntingly difficult to properly execute). An array of ca. 24 electrodes is threaded into the scala media of the cochlea. Generally, the end point of the array reaches into the basal third of the partition, perhaps as far as the 800-1000 Hz portion of the cochlea. Because there is often some residual hearing in the lowest frequencies, this technical limitation is not as serious as it may appear. The 24 electrodes generally span a spectral range between ca. 800 and 6,000 Hz. Not all of the electrodes are active. Rather, the intent is to chose between four and eight electrodes that effectively sample the spectral range. The speech signal is spectrally partitioned so that lower frequencies stimulate the most apical electrodes and the higher frequencies are processed through the more basal ones, in a frequency-graded manner. Thus, the implant performs

a crude form of spatial frequency analysis, analogous to that performed by the normal-functioning cochlea.

Complementing the cochlear place cues imparted by the stimulating electrode array is low-frequency periodicity information associated with the waveform's fundamental frequency. This voice-pitch information is signalled through the periodic nature of the stimulating pulses emanating from each electrode. In addition, coarse amplitude information is transmitted by the overall pulse rate.

Although the representation of the speech signal provided by the implant is a crude one, it enables most patients to verbally interact effectively. Shannon and colleagues have explored the nature of this coarse representation in normal-hearing individuals, demonstrating that only four spectrally discrete channels are required (under ideal listening conditions) to transmit intelligible speech using a noise-like phonation source (Shannon et al. 1995). Thus, it would appear that the success of cochlear implants relies, to a certain extent, on the relatively coarse spectro-temporal representation of information in the speech signal (cf. Section 4.1).

11. The Influence of Learning on Auditory Processing of Speech

Language represents the culmination of the human penchant for communicating vocally and appears to be unique in the animal kingdom (Hauser 1996). Much has been made of the creative aspect of language that enables the communication of ideas virtually without limit (cf. Chomsky 1965; Hauser et al. 2002; Studdert-Kennedy and Goldstein 2003). Chomsky (2000) refers to this singular property as "discrete infinity."

The limitless potential of language is grounded, however, in a vocabulary *with* limits. There are 218,000 word entries listed in the unabridged edition of the Oxford English Dictionary, the gold standard of English lexicography. And estimates of the average individual's *working* vocabulary range from 10,000 to perhaps a decade higher. But a statistical analysis of spontaneous dialogues (Am. English) reveals an interesting fact – 90% of the words used in casual discussions can be covered by less than a thousand distinctive lexical items (Greenberg 1999). The 100 most frequent words from the corpus analyzed by Greenberg (Switchboard) account for two-thirds of the lexical tokens and the ten most common words account for nearly 25% lexical usage (Greenberg 1999). Comparable statistics were compiled by French, Carter and Koenig (1930). Thus, while a speaker may possess the *potential* for producing tens of thousands of different words, in daily conversation this capacity is rarely exercised. Most speakers get by with only a few thousand words most of the time.

This finite property of spoken language is an important one, for it provides a means for the important elements to be learned effectively (if not fully mastered) in order to facilitate rapid and reliable communication. While "discrete infinity" is attractive in principle, it is unlikely to serve as an accurate description of spoken language in the "real world," where speakers are rarely creative or original. Most utterances are composed of common words sequenced in conventional ways (as observed by Skinner (1957) long ago). This stereotypy is characteristic of an overlearned system designed for rapid and reliable communication. With respect to spontaneous speech, Skinner is probably closer to the mark than Chomsky.

11.1 Auditory Processing with an Interpretative Linguistic Framework

Such constraints on lexical usage are important for understanding the role of auditory processing in linguistic communication. Auditory patterns, as processed by the brain, bear no significance except as they are interpretable with respect to the real world. In terms of language this means that the sounds spoken must be associated with specific events, ideas and objects. And given the very large number of prospective situations to describe, some form of structure is required in order that acoustic patterns can be readily associated with meaningful elements.

Such structure is readily discernible in the syntax and grammar of any language, which constrains the order in which words occur relative to each other. On a more basic level, germane to hearing are the constraints imposed on the sound shapes of words and syllables, which enable the auditory system to efficiently decode complex acoustic patterns within a *meaningful* linguistic framework. The examples that follow are designed to illustrate the importance of structure (and constraints implied) for efficiently decoding the speech signal.

The 100 most frequent words in English (accounting for 67% of the lexical instances) tend to contain but a single syllable, and the exceptions contain only two (Greenberg 1999). This subset of spoken English generally consists of the “function” words such as pronouns, articles, and locatives, and is generally of Germanic origin.

Moreover, most of these common words have a simple syllable structure, containing either a consonant followed by a vowel (CV), a consonant followed by a vowel, followed by another consonant (CVC), a vowel followed by a consonant (VC), or just a vowel by itself (V). Together,

these three syllable forms account for more than fourth-fifths of the syllables encountered (Greenberg 1999).

In contrast to function words are the “content” lexemes that provide the specific referential material enabling listeners to decode the message with precision and confidence. Such content words occur less frequently than their function-word counterparts, often contain three or more syllables and are generally nouns, adjectives or adverbs. Moreover, their linguistic origin is often non-Germanic – Latin and Norman French being the most common sources of this lexicon. When the words are of Germanic origin, their syllable structure is often complex (i.e., consonant clusters in either the onset or coda, or both). Listeners appear to be aware of such statistical correlations, however loose they may be.

The point reinforced by these statistical patterns is that spoken forms in language are far from arbitrary, and are highly constrained in their structure. Some of these structural constraints are specific to a language, but many appear to be characteristic of all languages (i.e., universal). Thus, all utterances are composed of syllables, and every syllable contains a nucleus, which is virtually always a vowel. Moreover, syllables can begin with a consonant, and most of them do. And while a syllable can also end with a consonant, this is much less likely to happen. Thus, the structural nature of the syllable is asymmetric – the question arises as to why?

Syllables can begin and end with more than a single consonant in many (but not all) languages. For example, in English, a word can conform to the syllable structure CCCVCCC (“strengths”), but rarely does so. When consonants do occur in sequence within a syllable their order is non-

random, but conforms to certain phonotactic “rules.” And these rules, themselves, are far from arbitrary, but conform to what is known as the “sonority hierarchy” (Clements 1990; Zec 1995), but which is really a cover term for sequencing segments in a quasi-continuous “energy arc” over the syllable.

Syllables begin with gradually increasing energy over time that rises to a crescendo in the nucleus before descending in the coda (or the terminal portion of the nucleus in the absence of a coda segment). This statement is an accurate description only for energy integrated over 25-ms time windows. Certain segments, principally the stops and affricates, begin with a substantial amount of energy that is sustained over a brief (ca. 10-ms) interval of time, which is followed by a more gradual buildup of energy over the following 40-100 ms. Vowels are the most energetic (i.e., intense) of segments, followed by the liquids, and glides (often referred to as “semi-vowels”) and nasals. The least intense segments are the fricatives (particularly of the voiceless variety), the affricates and the stops. It is a relatively straightforward matter of predicting the order of consonant types in onset and coda from the energy-arc principle. More intense segments do not precede less intense ones in the syllable onset building up to the nucleus. And conversely, less intense segments do not precede more intense ones in the coda. And if the manner (mode) of production is correlated with energy level, adjacent segments within the syllable should rarely (if ever) be of the same manner class, which is the case in spontaneous American English (Greenberg et al. 2002).

Moreover, the entropy associated with the syllable onset appears to be considerably greater than in the coda or nucleus. Pronunciation patterns are largely canonical (i.e., of the standard

dictionary form) at onset, with a full range of consonant segments represented. In coda position, three segments – [t], [d] and [n] – account for over 70% of the consonantal forms (Greenberg et al. 2002).

Such constraints serve to reduce the perplexity of constituents within a syllable, thus making "infinity" more finite (and hence more learnable) than would otherwise be the case. More importantly, they provide an auditory-based framework with which to *interpret* auditory patterns within a linguistic framework, reducing the effective entropy associated with many parts of the speech signal to manageable proportions (i.e., much of the entropy is located in the syllable onset which is more likely to evoke neural discharge in the auditory cortex). In the absence of such an interpretive framework auditory patterns could potentially lose all meaning and merely register as sound.

11.2 Visual Information Facilitates Auditory Interpretation

Most verbal interaction occurs face to face, thus providing visual cues with which to supplement and interpret the acoustic component of the speech signal. Normally, visual cues are unconsciously combined with the acoustic signal and are largely taken for granted. However, in noisy environments, such "speechreading" information provides a powerful assist in decoding speech, particularly for the hearing impaired (Sumby and Pollack 1954; Breeuer and Plomp 1984; Massaro 1987; Summerfield 1992; Grant and Walden 1996b; Grant et al. 1998; Assmann and Summerfield, Chapter 5).

Because speech *can* be decoded without visual input much of the time (e.g., over the telephone), the significance of speechreading is seldom fully appreciated. And yet there is substantial evidence that such cues often provide the extra margin of information enabling the hearing-impaired to communicate effectively with others. Grant and Walden (1995) have suggested that the amount of benefit provided by speechreading is comparable to, or even exceeds that of a hearing aid for many of the hearing impaired.

How are such cues combined with the auditory representation of speech? Relatively little is known about the specific mechanisms. Speechreading cues appear to be primarily associated with place-of-articulation information (Grant et al. 1998), while voicing and manner information are derived almost entirely from the acoustic signal.

The importance of the visual modality for place-of-articulation information can be demonstrated through presentation of two different syllables, one using the auditory modality, the other via the visual channel. If the consonant in the acoustic signal is [p] and is [k] in the visual signal (all other phonetic properties of the signals being equal), listeners often report “hearing” [t], which represents a blend of the audio-visual streams with respect to place of articulation (McGurk and McDonald 1976). Although this “McGurk effect” has been studied intensively (cf. Summerfield 1992), the underlying neurological mechanisms remain obscure. Whatever its genesis in the brain, the mechanisms responsible for combining auditory and visual information must lie at fairly abstract level of representation. It is possible for the visual stream to precede the audio by as much as 120-200 ms without an appreciable affect on intelligibility (Grant and

Greenberg 2001). However, if the audio precedes the video, intelligibility falls dramatically for leads as small as 50-100 ms. The basis for this sensory asymmetry in stream asynchrony is the subject of on-going research. Regardless of the specific nature of the neurological mechanisms underlying auditory-visual speech processing, it serves as a powerful example of how the brain is able to interpret auditory processing within a larger context.

11.3 Informational Constrains on Auditory Speech Processing

It is well known that the ability to recognize speech depends on the size of the response set – the smaller the number of linguistic categories involved, the easier it is for listeners to correctly identify words and phonetic segments (Pollack 1959) for any given signal-to-noise ratio. In this sense, the amount of inherent information (often referred to as (negative) “entropy”) associated with a recognition or identification task has a direct impact on performance (cf. Assmann and Summerfield, Chapter 5), accounting to a certain degree, for variation in performance using different kinds of speech material. Thus, at an SNR of 0 dB, spoken digits are likely to be recognized with 100% accuracy, while for words of a much larger response set (in the hundreds or thousands) the recognition score will be ca. 50% or less under comparable conditions.

However, if these words were presented at the same SNR in a connected sentence, the recognition score would rise to ca. 80%. Presentation of spoken material within a grammatical and semantic framework clearly improves the ability to identify words.

The articulation index was originally developed using nonsense syllables devoid of semantic context, on the assumption that the auditory processes involved in this task are comparable to those

operating in a more realistic linguistic context. Hence, a problem decoding the phonetic properties of nonsense material should, in principle, also be manifest in continuous speech. This is the basic premise underlying extensions of the articulation index to meaningful material (e.g., Boothroyd and Nittrouer 1988; cf. Assmann and Summerfield, Chapter 5). However, this assumption has never been fully verified, and therefore the relationship between phonetic-segment identification and decoding continuous speech remains to be clarified.

11.4 Categorical Perception

The importance of learning and generalization in speech decoding is amply illustrated in studies on "categorical perception" (cf. Rosen and Howell 1987). In a typical experiment, a listener will be asked to denote a speech segment as an exemplar of either class A or B. Unbeknownst to the subject, a specific acoustic parameter has been adjusted in fine increments along a continuum. At one end of the continuum virtually all listeners identify the sounds as A, while at the other end, all of the sounds are classified as B. In the middle responses are roughly equally divided between the two. The key test is one in which discrimination functions between two members of the continuum are produced. In instances where one stimulus has been clearly identified as A and the other as B, these signals are accurately distinguished and labeled as "different." In true categorical perception, listeners are only able to reliably *discriminate* between signals *identified* as different phones. Stimuli from within the same labeled class, even though they differ along a specific acoustic dimension, are not reliably distinguished (cf. Liberman et al. 1957).

A number of specific acoustic dimensions have been shown to conform to categorical perception, among them voice onset time (VOT; cf. Lisker and Abramson 1964) and place of articulation. VOT refers to the interval of time separating the articulatory release from glottal vibration (cf. Chapters 2 and 3). For a segment, such as [b], VOT is short, typically less than 20 ms, while for its voiceless counterpart, [p], the interval is generally 40 ms or greater. Using synthetic stimuli it is possible to parametrically vary VOT between 0 and 60 ms, keeping other properties of the signal constant. Stimuli with a VOT between 0 - 20 ms are usually classified as [b], while those with a VOT between 40 and 60 ms are generally labeled as [p]. Stimuli with VOTs between 20 and 40 ms often sound ambiguous, eliciting [p] and [b] responses in varying proportions. The VOT boundary is defined as that interval for which [p] and [b] responses occur in roughly equal proportion. Analogous experiments have been performed for other stop consonants, as well as for segments associated with different manner-of-articulation classes (cf. Liberman et al. 1967; Liberman and Mattingly 1985 for reviews).

Categorical perception provides an illustration of the interaction between auditory perception and speech identification using a highly stylized signal. In this instance listeners are given only two response classes and are forced to choose between them. The inherent entropy associated with the task is low (essentially a single bit of information, given the binary nature of the classification task), unlike speech processing in more natural conditions where the range of choices at any given instant is considerably larger. However, the basic lesson of categorical perception is still valid – that perception can be guided by an *abstraction* based on a learned system, rather than by specific

details of the acoustic signal. Consistent with this perspective are studies in which it is shown that the listener's native language has a marked influence on the location of the category boundary (e.g., Miyawaki et al. 1975).

However, certain studies suggest that categorical perception may not reflect linguistic processing per se, but rather is the product of more general auditory mechanisms. For example, it is possible to shift the VOT boundary by selective adaptation methods, in which the listener is exposed to repeated presentation of the same stimulus (usually an exemplar of one end of the continuum) prior to classification of a test stimulus. Under such conditions the boundary shifts away (usually by 5-10 ms) from the exemplar (Eimas and Corbit 1973; Ganong 1980). The standard interpretation of this result is that VOT detectors in the auditory system have been "fatigued" by the exemplar.

Categorical perception has also been used to investigate the ontogeny of speech processing in the maturing brain. Infants as young as one month are able to discriminate, as measured by recovery from satiation, two stimuli from different acoustic categories more reliably than signals with comparable acoustic distinctions from the same phonetic category (Eimas et al. 1971). Such a result implies that the basic capability for phonetic-feature detection may be "hard-wired" into brain, although exposure to language-specific patterns appears to play an important role as well (Strange and Dittman 1983; Kuhl et al. 1997).

The specific relation between categorical perception and language remains controversial. A number of studies have shown that non-human species, such as chinchilla (Kuhl and Miller 1978),

macaque (Kuhl and Padden 1982) and quail (Kluender 1991) all exhibit behavior comparable in certain respects to categorical perception in humans. Such results suggest that at least some properties of categorical perception are not strictly language-bound but rather reflect the capability of consistent generalization between classes regardless of their linguistic significance (Kluender et al. 2003).

12. Technology, Speech and the Auditory System

Technology can serve as an effective proving ground for ideas generated during the course of scientific research (Greenberg 2003). Algorithms based on models of the auditory system's processing of speech can, in principle, be used in auditory prostheses, as well as for automatic speech recognition systems and other speech applications. To the extent that these auditory-inspired algorithms improve performance of the technology some degree of confidence is gained that the underlying ideas are based on something more than wishful thinking or mathematical elegance. Moreover, careful analysis of the problems encountered in adapting scientific models to real-world applications can provide insight into the limitations of such models as a description of the processes and mechanisms involved (Greenberg 2003).

12.1 Automatic Speech Recognition (Front-end Features)

The historical evolution of automatic speech recognition can be interpreted as a gradually increasing awareness of the specific problems required to be solved (Ainsworth 1988). For example, an early, rather primitive system developed by Davis and colleagues (Davis et al. 1952) achieved a word-recognition score of 98% correct for digits spoken by a single speaker. However,

the recognition score dropped to ca. 50% when the system was tested on other speakers. This particular system measured the zero-crossing rate of the speech signal's pressure waveform after it had been filtered into two discrete frequency channels roughly corresponding to the range associated with the first and second formants. The resulting outputs were cross-correlated with a set of stored templates associated with representative exemplars for each digit. The digit template associated with the highest correlation score was chosen as the recognized word.

This early system's structure – some form of frequency analysis followed by a pattern matcher – persists in contemporary systems, although the nature of the analyses and pattern recognition techniques used in contemporary systems has markedly improved in recent years. Early recognition systems used pattern-matching methods to compare a sequence of incoming feature vectors derived from the speech signal with a set of stored word templates. Recognition error rates for *speaker-dependent* recognizers dropped appreciably when dynamic-time-warping (DTW) techniques were introduced as a means of counteracting durational variability (Velichko and Zagoruyko 1970; Sakoe and Chiba 1978). However the problem associated with *speaker-independent* recognition remained until statistical methods were introduced in the late 1970's.

Over the past twenty-five years statistical approaches have replaced the correlational and DTW approaches of the early ASR systems and are embedded within a mathematical framework known as hidden Markov models (HMMs) (e.g., Jelinek 1976, 1997). HMMs are used to represent each word and sub-word (usually phoneme) unit involved in the recognition task. Associated with each

HMM state is a probability score associated with the likelihood of a particular unit occurring in that specific context given the training data used to develop the system.

One of the key problems that a speech recognizer must address is how to efficiently reduce the amount of data representing the speech signal without compromising recognition performance. Can principles of auditory function can be used to achieve this objective as well as to enhance automatic speech recognition performance, particularly in background noise?

Speech technology provides an interesting opportunity to test many of the assumptions that underlie contemporary theories of hearing (cf. Hermansky 1998; Morgan et al., Chapter 6). For example, the principles underlying the spectral representation used in automatic speech recognition (ASR) systems are directly based on perceptual studies of speech and other acoustic signals. In contrast to Fourier analysis, which samples the frequency spectrum linearly (in terms of Hz units), modern approaches (Mel frequency cepstral coefficients - Davis and Mermelstein 1980; Perceptual linear prediction - Hermansky 1990) warp the spectral representation, giving greater weight to frequencies below 2 kHz. The spatial-frequency mapping is logarithmic above 800 Hz (Avendaño et al., Chapter 2; Morgan et al., Chapter 6), in a manner comparable to what has been observed in both perceptual and physiological studies. Moreover, the granularity of the spectral representation is much coarser than the Fast Fourier Transform (FFT), and is comparable to the critical-band analysis performed in the cochlea (Section 9). The representation of the spectrum is highly smoothed, simulating integrative processes in both the periphery and central regions of the auditory pathway. In addition, the representation of spectral magnitude is not in terms of decibels

(a physical measure), but rather in units analogous to sones, a perceptual measure of loudness rooted in the compressive nature of transduction in the cochlea and beyond (cf. Zwicker 1975; Moore 1997). This sort of transformation has the effect of compressing the variation in peak magnitude across the spectrum, thereby providing a parsimonious and effective method of preserving the *shape* of the spectral envelope across a wide variety of environmental conditions.

RASTA is yet another example of auditory-inspired signal processing that has proven useful in ASR systems. Its conceptual roots lie in the sensory and neural adaptation observed in the cochlea and other parts of the auditory pathway. Auditory neurons adapt their response level to the acoustic context in such a manner that a continuous signal evokes a lower level of activity during most of its duration than at stimulus onset (Smith 1977). This reduction in responsiveness may last for hundreds or even thousands of milliseconds after cessation of the signal, and can produce an auditory "negative afterimage" in which a phantom pitch is "heard" in the region of the spectrum close to that of the original signal (Zwicker 1964). Summerfield et al. (1987) demonstrated that such an afterimage could be generated using a steady-state vowel in a background of noise. Once the original vowel was turned off subjects faintly perceived a second vowel whose spectral properties were the inverse of the first.

This type of phenomenon implies that the auditory system should be most responsive to signals whose spectral properties evade the depressive consequences of adaptation through constant movement at rates that lie outside the time constants characteristic of sensori-neural adaptation. The formant transitions in the speech signal move at such rates over much of their course, and are

therefore likely to evoke a relatively high level of neural discharge across a tonotopically organized population of auditory neurons. The rate of this formant movement can be modeled as a temporal filter with a specific time constant (ca. 160 ms) and used to process the speech signal in such a manner as to provide a representation that weights the spectrally dynamic portions of the signal much more highly than the steady-state components. This is the essence of RASTA, a technique that has been used to shield the speech spectrum against the potential distortion associated with microphones and other sources of extraneous acoustic energy (Hermansky and Morgan 1993; Morgan et al., Chapter 6).

12.2 Speech Synthesis

Computational simulation of the speech-production process, known as speech synthesis, affords yet a separate opportunity to evaluate the efficacy of auditory-inspired algorithms. Synthesis techniques have focused on three broad issues: (1) intelligibility, (2) quality (i.e., naturalness), and (3) computational efficiency. Simulating the human voice in a realistic manner requires knowledge of the speech production process, as well as insight into how the auditory system interprets the acoustic signal.

Over the years two basic approaches have been used, one modeling the vocal production of speech, the other focusing on spectro-temporal manipulation of the acoustic signal. The vocal tract method was extensively investigated by Flanagan and colleagues at Bell Labs (Flanagan 1972) and by Klatt at MIT (Klatt 1987). The entire speech production process is simulated, from the flow of the air stream through the glottis into the oral cavity and out of the mouth, to the movement of the

tongue, lips, velum and jaw. These serve as control parameters governing the acoustic resonance patterns and mode of vocal excitation. The advantage of this method is representational parsimony – a production-based model generally contains between 30 and 50 parameters updated 100 times per second. Because many of the control states do not change from frame to frame, it is possible to specify an utterance with perhaps a thousand different parameters (or less) per second. In principle, any utterance, from any language, can be generated from such a model, as long as the relation between the control parameters and the linguistic input are known. Although such vocal tract synthesizers are generally intelligible, they are typically judged as sounding unnatural by human listeners. The voice quality has a metallic edge to it, and the durational properties of the signal are not quite what a human would produce.

The alternative approach to synthesis starts with a recording of the human voice. In an early version of this method, as exemplified by the Vocoder (cf. Section 4.1), the granularity of the speech signal was substantially reduced both in frequency and in time, thereby compressing the amount of information required to produce intelligible speech. This synthesis technique is essentially a form of recoding the signal, as it requires a recording of the utterance to be made in advance. It does not provide a principled method for extrapolating from the recording to novel utterances.

Concantenative synthesis attempts to fill this gap by generating continuous speech from several hours of pre-recorded material. Instead of simulating the vocal production process, it assumes that the elements of any and all utterances that might ever be spoken are contained in a finite sample of

recorded speech. Thus, it is a matter of splicing the appropriate intervals of speech together in the correct order. The “art” involved in this technique pertains to the algorithms used to determine the length of the spliced segments and the precise context from which they come. At its best, concantenative synthesis sounds remarkably natural and is highly understandable. For these reasons, most contemporary commercial text-to-speech applications are based on this technology. However, there are two significant limitations. First, synthesis requires many hours of material to be recorded from each speaker used in the system. The technology does not provide a principled method of generating voices other than those previously recorded. Second, the concantenative approach does not, in fact, handle all instances of vocal stitching well. Every so often such systems produce unintelligible utterances in circumstances where the material to be spoken lies outside the range of verbal contexts recorded.

A new form of synthesis, known as “STRAIGHT,” has the potential to rectify the problems associated with production-based models and concantenative approaches. STRAIGHT is essentially a highly granular Vocoder, melded with sophisticated signal-processing algorithms that enable flexible and realistic alteration of the formant patterns and fundamental frequency contours of the speech signal (Kawahara et al. 1999). Although the synthesis method uses pre-recorded material, it is capable of altering the voice quality in almost unlimited ways, thereby circumventing the most serious limitation of concantenative synthesis. Moreover, it can adapt the spectro-temporal properties of the speech waveform to any specifiable target. STRAIGHT requires ca. 1,000 separate channels to fully capture the natural quality of the human voice, 100 times as many

channels as used by the original Vocoder of the 1930's. Such a dense sampling of the spectrum is consistent with the innervation density of the human cochlea – 3,000 inner hair cells projecting to 30,000 auditory-nerve fibers – and suggests that under-sampling of spectral information may be a major factor in the shortcomings of current-generation hearing aids in rendering sound to the ear.

12.3 Auditory Prostheses

Hearing-aid technology stands to benefit enormously from insights into the auditory processing of speech and other communication signals. A certain amount of knowledge, pertaining to spectral resolution and loudness compression, has already been incorporated into many aids (e.g., Villchur 1987; cf. Edwards, Chapter 7). However, such aids do not entirely compensate for the functional deficit associated with sensori-neural damage (cf. Section 10). The most sophisticated hearing aids incorporate up to 64 channels of quasi-independent processing, with four to eight different compression settings specifiable over the audio range. Given the spectral-granularity capability of the normal ear (cf. Sections 7 and 9) it is conceivable that hearing aids would need to provide a much finer-grained spectral representation of the speech signal in order to provide the sort of natural quality characteristic of the human voice. On the other hand, it is not entirely clear whether the damaged ear would be capable of exploiting such fine spectral detail.

One of the most significant problems with current-generation hearing aids is the difficulty encountered processing speech in noisy backgrounds. Because the basic premise underlying hearing-aid technology is amplification ("power to the ear!"), boosting the signal level per se also increases the noise background. The key is to enhance the speech signal and other foreground

signals while suppressing the background. To date, hearing-aid technology has not been able to solve this problem despite some promising innovations. One method, called the "Voice Activity Detector," adjusts the compression parameters in the presence (or absence) of speech, based on algorithms similar in spirit to RASTA. Modulation of energy at rates between 3 and 10 Hz are interpreted as speech, with attendant adjustment of the compression parameters. Unfortunately, this form of quasi-dynamic range adjustment is not sufficient to ameliorate the acoustic interference problem. Other techniques, based on deeper insight into auditory processes, will be required (cf. Section 13).

12.4 Automatic Speech Recognition (Lexical Decoding)

There is far more to decoding speech than mere extraction of relevant information from the acoustic signal. It is for this reason that ASR systems focus much of their computational power on associating spectro-temporal features gleaned from the "front end" with meaningful linguistic units such as phones, syllables and words.

Most current-generation ASR systems use the phoneme as the basic decoding unit (cf. Morgan et al., Chapter 6). Words are represented as linear sequences of phonemic elements, which are associated with spectro-temporal cues in the acoustic signal via acoustic models trained on context-dependent phone models. The problem with this approach is the enormous amount of pronunciation variation characteristic of speech spoken in the real world. Much of this variation is inherent to the speaking process and reflects dialectal, gender, emotional, socio-economic and

stylistic factors. The phonetic properties can vary significantly from one context to the next, even for the same speaker (Section 2).

Speech recognition systems currently do well only in circumstances where they have been trained on extensive amounts of data representative of the task domain and where the words spoken (and their order) are known in advance. For this reason, ASR systems tend to perform best on prompted speech, where there is a limited set of lexical alternatives (e.g., an airline reservation system), or where the spoken material is read a careful manner (and hence the amount of pronunciation variation is limited). Thus, current ASR systems function essentially as sophisticated decoders rather than as true open-set recognition devices. For this reason automatic speech recognition is expensive, time-consuming technology to develop and is not easily adaptable to novel task domains.

13. Future Trends in Auditory Research Germane to Speech

Spoken language is based on processes of enormous complexity, involving many different regions of the brain, including those responsible for hearing, seeing, remembering and interpreting. This volume focuses on just one of these systems, hearing, and attempts to relate specific properties of the auditory system to the structure and function of speech. In coming years our knowledge of auditory function is likely to increase substantially and in ways potentially capable of having a direct impact on our understanding of the speech decoding process.

It is becoming increasing clear that the auditory pathway interacts either directly or indirectly with many other parts of the brain. For example, visual input can directly affect the response

properties of neurons in the auditory cortex (Sharma et al. 2000), and there are instances where even somatosensory input can affect auditory processing (Gao and Suga 2000). It is thus becoming increasingly evident that auditory function cannot be entirely understood without taking such cross-modal interactions into consideration. Moreover, the auditory system functions as part of an integrated behavioral system where, in many circumstances, it may provide only a small part of the information required to perform a task. Many properties of hearing can only be fully appreciated within such a “holistic” framework. Spoken language is perhaps the most elaborate manifestation of such integrated behavior and thus provides a fertile framework in which to investigate the interaction among various brain regions involved in the execution of complex behavioral tasks.

Future research pertaining to auditory function and speech is likely to focus on several broad areas.

Human brain-imaging technology has improved significantly over the past decade, so that is it now possible to visualize neural activation associated with specific behavioral tasks with a degree of spatial and temporal resolution undreamt of in the recent past. Such techniques as functional magnetic resonance imaging (fMRI) and magnetoencephalography (MEG) will ultimately provide (at least in principle) the capability of answering many of the “where” and “when” questions posed in this chapter. Dramatic discoveries are likely to be made using such imaging methods over the next decade, particularly with respect to delineating the interaction and synergy among various neurological systems involved in processing spoken language.

Language is a highly learned behavior, and it is increasing clear that learning plays an important role in auditory function (Recanzone et al. 1993; Wright et al. 1997) germane to speech processing. How does the auditory system adapt to experience with specific forms of acoustic input? Do sensory maps of fundamental auditory features change over time in response to such acoustic experience (as has been demonstrated in the barn owl, cf. Knudsen 2002)? What is the role of attentional processes in the development of auditory representations and the ability to reliably extract behaviorally relevant information? Do human listeners process sounds differently depending on exposure to specific acoustic signals? Are certain language-related disorders the result of a poor connection between the auditory and learning systems? These and related issues are likely to form the basis of much hearing-based research over the next twenty years.

Technology has historically served as a "forcing function," driving the pace of innovation in many fields of scientific endeavor. This technology-driven research paradigm is likely to play an ever-increasing role in the domains of speech and auditory function.

For example, hearing aids do not currently provide a truly effective means of shielding speech information in background noise, nor are automatic speech recognition systems fully capable of decoding speech under even moderately noisy conditions. For either technology to evolve the noise-robustness problem needs to be solved, both from an engineering and (more importantly) a scientific perspective. And because of this issue's strategic importance for speech technology it is likely that a considerable amount of research will focus on this topic over the next decade.

Intelligibility remains perhaps the single most important issue for auditory prostheses. The hearing-impaired wish to communicate easily with others, and the auditory modality provides the most effective means to do so. To date, conventional hearing aids have not radically improved the ability to understand spoken language except in terms of enhanced audibility. Digital compression aids provide some degree of improvement with respect to noise robustness and comfort, but a true breakthrough in terms of speech comprehension awaits advances in the technology. One of the obstacles to achieving such an advance is our limited knowledge of the primary cues in the speech signal required for a high degree of intelligibility (cf. Greenberg et al. 1998; Greenberg and Arai 2001; Müsch and Buus 2001a, 2001b). Without such insight it is difficult to design algorithms capable of significantly enhancing speech understanding. Thus, it is likely that a more concerted effort will be made over the next few years to develop accurate speech intelligibility metrics germane to a broad range of acoustic-environment conditions representative of the real world (and which are more accurate than the AI and STI).

Related to this effort will be advances in cochlear implant design that provide a more natural-sounding input to the auditory pathway than current devices afford. Such devices are likely to incorporate a more fine-grained representation of the speech spectrum than is currently provided, as well as using frequency-modulation techniques in tandem with those based on amplitude modulation to simulate much of the speech signal's spectro-temporal detail.

The fine detail of the speech signal is also important for speech synthesis applications, where a natural-sounding voice is often of paramount importance. Currently, the only practical means of

imparting a natural quality to the speech is by pre-recording the materials with a human speaker. However, this method ("concantenative synthesis") limits voice quality and speaking styles to the range recorded. In the future, new synthesis techniques (such as STRAIGHT, cf. Kawahara et al. 1999), will enable life-like voices to be created, speaking in virtually any style and tone imaginable (and for a wide range of languages). Moreover, the acoustic signal will be melded with a visual display of a talking avatar simulating the look and feel of a human speaker. Achieving such an ambitious objective will require far more detailed knowledge of the auditory (and visual) processing of the speech stream, as well as keen insight into the functional significance of the spectro-temporal detail embedded in the speech signal.

Automatic speech recognition is gaining increasing commercial acceptance and is now commonly deployed for limited verbal interactions over the telephone. Airplane flight and arrival information, credit card and telephone account information, stock quotations and the like are now often mediated by speaker-independent, constrained-vocabulary ASR systems in various locations in North America, Europe and Asia. This trend is likely to continue, as companies learn how to exploit such technology (often combined with speech synthesis) to simulate many of the functions previously performed by human operators.

However, much of automatic speech recognition's true potential lies beyond the limits of current technology. Currently, ASR systems perform well only in highly constrained, linguistically prompted contexts, where very specific information is elicited through the use of pinpoint questions (e.g., Gorin et al. 1997). This form of interaction is highly unnatural and customers

quickly tire of its repetitive, tedious nature. Truly robust ASR would be capable of providing the illusion of speaking to a real human operator, an objective that lies many years in the future. The knowledge required to accomplish this objective is immense and highly variegated. Detailed information about spoken language structure and its encoding in the auditory system is also required before speech recognition systems achieve the level of sophistication required to successfully simulate human dialogue.

Advances in speech recognition and synthesis technology may ultimately advance the state of auditory prostheses. The hearing aid and cochlear implant of the future are likely to utilize such technology as a means of providing a more intelligible and life-like signal to the brain. Adapting the auditory information provided, depending on the nature of the interaction context (e.g., the presence of speechreading cues and/or background noise) will be commonplace.

Language learning is yet another sector likely to advance as a consequence of increasing knowledge of spoken language and the auditory system. Current methods of teaching pronunciation of foreign languages are often unsuccessful, focusing on the articulation of phonetic segments in isolation, rather than as an integrated whole organized prosodically. Methods for providing accurate, production-based feedback based on sophisticated phonetic and prosodic classifiers could significantly improve pronunciation skills of the language student. Moreover, such technology could also be used in remedial training regimes for children with specific articulation disorders.

Language is what makes humans unique in the animal kingdom. Our ability to communicate via the spoken word is likely to be associated with the enormous expansion of the frontal regions of the human cortex over the course of recent evolutionary history and probably laid the behavioral groundwork for development of complex societies and their attendant cultural achievements. A richer knowledge of this crucial behavioral trait depends in large part on deeper insight into the auditory foundations of speech communication.

Acknowledgements

The authors would like to thank the chapter authors for their hard work and diligence in preparing the material that appears in this book. We are also grateful to Arthur Popper and Richard Fay, series editors for the Springer Handbook of Auditory Research, for their encouragement and patience throughout this volume's lengthy gestation period. One of the authors of this chapter, Bill Ainsworth, died unexpectedly in January of 2002 (cf. In Memoriam at the front of this volume).

Notes

1. However, it is unlikely that speech evolved *de novo*, but rather represents an elaboration of a more primitive form of acoustic communication utilized by our primate forebears (cf. Hauser 1996). Many of the selection pressures shaping these non-human communication systems, such as robust transmission under uncertain acoustic conditions (cf. Chapter 5 in this volume), apply to speech as well.

2. Fletcher began his speech research at Western Electric, which manufactured telephone equipment for AT&T and other telephone companies. In 1925 Western Electric was merged with AT&T, and Bell Laboratories established. Fletcher directed the acoustics research division at Bell Labs for many years before his retirement from AT&T in 1951.

3. Remez and associates would disagree with this statement, claiming in their paper and in subsequent publications and presentations that sine-wave speech is indeed intelligible. The authors of this chapter (and many others in the speech community) respectfully disagree with their assertion.

List of Abbreviations

AC	alternating current
AGC	automatic gain control
AI	articulation index
ALSR	average localized synchronized rate
ANF	auditory nerve fiber
ASR	automatic speech recognition
CF	characteristic frequency
CV	consonant - vowel
CVC	consonant-vowel-consonant
Δf	frequency DL
ΔI	intensity DL
DC	direct current
DL	difference limen
DTW	dynamic time warping
F_1	first formant
F_2	second formant

F_3	third formant
FFT	fast Fourier transform
fMRI	functional magnetic resonance imaging
f_0	fundamental frequency
FTC	frequency threshold curve
HMM	hidden Markov model
IHC	inner hair cell
MEG	magnetoencephalography
OHC	outer hair cell
PLP	perceptual linear prediction
SNR	signal-to-noise ratio
SR	spontaneous rate
STI	speech transmission index
TM	tectorial membrane
V	vowel
VC	vowel-consonant
VOT	voice onset time

References

Ainsworth WA (1976) Mechanisms of Speech Recognition. Oxford: Pergamon Press.

Ainsworth WA (1986) Pitch change as a cue to syllabification. J Phonetics 14: 257-264.

Ainsworth W A (1988) Speech Recognition by Machine. Stevenage (UK): Peter Peregrinus.

Ainsworth WA, Lindsay D (1986) Perception of pitch movements on tonic syllables in British English. J Acoust Soc Am 79: 472-480.

Allen JB (1994) How do humans process and recognize speech? IEEE Trans Speech Audio Proc 2: 567-577.

Anderson DJ, Rose JE, Brugge JF (1971) Temporal position of discharges in single auditory nerve fibers within the cycle of a sine-wave stimulus: Frequency and intensity effects. J Acoust Soc Am 49: 1131-1139.

Arai T, Greenberg S (1998) Speech intelligibility in the presence of cross-channel spectral asynchrony. Proc IEEE Int Conf Acoust Speech Sig Proc (ICASSP-98), pp. 933-936.

Baer T, Moore BCJ (1993) Effects of spectral smearing on the intelligibility of sentences in noise. J Acoust Soc Am 94: 1229-1241.

Blackburn CC, Sachs MB (1990) The representation of the steady-state vowel sound [ε] in the discharge patterns of cat anteroventral cochlear nucleus neurons. J Neurophysiol 63: 1191-1212.

Blauert J (1996) Spatial Hearing: The Psychophysics of Human Sound Localization, 2nd ed. Cambridge, MA: MIT Press.

Blesser B (1972) Speech perception under conditions of spectral transformation. I. Phonetic characteristics. J Speech Hear Res 15: 5-41.

Bohne BA, Harding GW (2000) Degeneration in the cochlea after noise damage: Primary versus secondary events. Am J Otol 21: 505-509.

Bolinger D (1986) Intonation and Its Parts: Melody in Spoken English. Stanford: Stanford University Press.

Bolinger D (1989) Intonation and Its Uses: Melody in Grammar and Discourse. Stanford: Stanford University Press.

Boothroyd A, Nittrouer S (1988) Mathematical treatment of context effects in phoneme and word recognition. J Acoust Soc Am 84 101-114.

Boubana S, Maeda S (1998) Multi-pulse LPC modeling of articulatory movements. Speech Comm 24: 227-248.

Breeuer M, Plomp R (1984) Speechreading supplemented with frequency-selective sound-pressure information. J Acoust Soc Am 76: 686-691.

Bregman AS (1990) Auditory Scene Analysis. Cambridge, MA: MIT Press.

Brokx JPL, Nooteboom SG (1982) Intonation and the perceptual separation of simultaneous voices. J Phon 10: 23-36.

Bronkhorst AW (2000) The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. Acustica 86: 117-128.

Brown GJ, Cooke MP (1994), Computational auditory scene analysis. Comp Speech Lang 8: 297-336.

Buchsbaum BR, Hickok G, Humphries C (2001) Role of left posterior superior temporal gyrus in phonological processing for speech perception and production. Cognitive Sci 25: 663-678.

Carlson R, Granstrom B (eds) (1982) The Representation of Speech in the Peripheral Auditory System. Amsterdam: Elsevier.

Carré R, Mrayati M (1995) Vowel transitions, vowel systems and the Distinctive Region Model. In: Sorin C, Méloni H, Schoentingen J (eds). Levels in Speech Communication: Relations and Interactions. Amsterdam: Elsevier, pp. 73-89.

Chistovich LA (1985) Central auditory processing of peripheral vowel spectra. J Acoust Soc Am 77: 789-805.

Chomsky N (1965) Aspects of the Theory of Syntax. Cambridge, MA: MIT Press.

Chomsky N (2000) New Horizons in the Study of Language and Mind. Cambridge: Cambridge University Press.

Clark GM (2003) Cochlear Implants: Fundamentals and Applications. New York: Springer-Verlag.

Clements GN (1990) The role of the sonority cycle in core syllabification. In: Kingston J and Beckman M (eds) Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech. Cambridge: Cambridge University Press, pp. 283-325.

Cooke MP (1993) Modelling Auditory Processing and Organisation. Cambridge: Cambridge University Press.

Cooke M, Ellis DPW (2001) The auditory organization of speech and other sources in listeners and computational models. Speech Comm 35: 141-177.

Darwin CJ (1981) Perceptual grouping of speech components different in fundamental frequency and onset-time. Q J Exp Psychol 3(A): 185-207.

Darwin CJ, Carlyon RP (1995) Auditory grouping. In: Moore BCJ (ed) The Handbook of Perception and Cognition, Vol. 6, Hearing. London: Academic Press, pp. 387-424.

Davis K, Biddulph R, Balashek S (1952) Automatic recognition of spoken digits. J Acoust Soc Am 24: 637-642.

Davis SB, Mermelstein P (1980) Comparison of parametric representation for monosyllabic word representation in continuously spoken sentences. IEEE Trans Acoust Speech Sig Proc 28: 357-366.

Delgutte B, Kiang NY-S (1984) Speech coding in the auditory nerve: IV. Sounds with consonant-like dynamic characteristics. J Acoust Soc Am 75: 897-907.

Deng L, Geisler CD, Greenberg S (1988) A composite model of the auditory periphery for the processing of speech. J Phonetics 16: 93-108.

Drullman R, Festen JM, Plomp R (1994a) Effect of temporal envelope smearing on speech reception. J Acoust Soc Am 95: 1053-1064.

Drullman R, Festen JM, Plomp R (1994b) Effect of reducing slow temporal modulations on speech reception. J Acoust Soc Am 95: 2670-2680.

Drullman R (2003) The significance of temporal modulation frequencies for speech intelligibility. In: Greenberg S, Ainsworth WA (eds) Listening to Speech: An Auditory Perspective. Hillsdale, NJ: Erlbaum, pp. xxx-xxx.

Dubno JR, Dirks DD (1989) Auditory filter characteristics and consonant recognition for hearing-impaired listeners. J Acoust Soc Am 85: 1666-1675.

Dubno JR, Schaefer AB (1995) Frequency selectivity and consonant recognition for hearing-impaired and normal-hearing listeners with equivalent masked thresholds. J Acoust Soc Am 97: 1165-1174.

Dudley H (1939) Remaking speech. J Acoust Soc Am 11: 169-177.

Dye RH, Hafter ER (1980) Just-noticeable differences of frequency for masked tones. J Acoust Soc Am 67: 1746-1753.

Eimas PD, Corbit JD (1973) Selective adaptation of linguistic feature detectors. Cognitive Psychol 4: 99-109.

Eimas PD, Siqueland ER, Jusczyk P, Vigorito J (1971) Speech perception in infants. Science 171: 303-306.

Fant G (1960) Acoustic Theory of Speech Production. The Hague: Mouton.

Fay RR, Popper AN (1994) Comparative Hearing: Mammals. New York: Springer-Verlag.

Festen JM, Plomp R (1981) Relations between auditory functions in normal hearing. J Acoust Soc Am 70: 356-369.

Flanagan JL (1955) A difference limen for vowel formant frequency. J Acoust Soc Am 27: 613-617.

Flanagan JL (1957) Estimates of the maximum precision necessary in quantizing certain "dimensions" of vowel sounds. J Acoust Soc Am 29: 533-534.

Flanagan JL (1972) Speech Analysis, Synthesis and Perception, 2nd ed. Berlin: Springer-Verlag.

Fletcher H (1953) Speech and Hearing in Communication. New York: Van Nostrand.

Fletcher H, Gault RH (1950) The perception of speech and its relation to telephony. J Acoust Soc Am 22: 89-150.

Fourcin AJ (1975). Language development in the absence of expressive speech. In: Lenneberg EH, Lenneberg E (eds). Foundations of Language Development, Vol. 2. New York: Academic Press, pp. 263-268.

Fowler C (1986) An event approach to the study of speech perception from a direct-realist perspective. J Phon 14: 3-28.

Fowler CA (1996) Listeners do hear sounds, not tongues. J Acoust Soc Am 99: 1730-1741.

French, NR, Carter CW, Koenig W. (1930) The words and sounds of telephone conversations. Bell System Tech J 9: 290-324.

French NR, Steinberg JC (1947) Factors governing the intelligibility of speech sounds. J Acoust Soc Am 19: 90-119.

Fujimura O, Lindqvist J (1971) Sweep-tone measurements of vocal tract characteristics. J Acoust Soc Am 49: 541-558.

Ganong WF (1980) Phonetic categorization in auditory word recognition. J Exp Psych (HPPP): 6: 110-125.

Gao E, Suga N (2000) Experience-dependent plasticity in the auditory cortex and the inferior colliculus of bats: Role of the corticofugal system. Proc Nat Acad Sci 97: 8081-8085.

Geisler CD, Greenberg S (1986) A two-stage automatic gain control model predicts the temporal responses to two-tone signals. J Acoust Soc Am 80: 1359-1363.

Ghitza O (1988) Temporal non-place information in the auditory-nerve firing patterns as a front-end for speech recognition in a noisy environment. J Phon 16: 109-123.

Gibson JJ (1966) The Senses Considered as Perceptual Systems. Boston: Houghton Miflin.

Gibson JJ (1979) The Ecological Approach to Visual Perception. Boston: Houghton Miflin.

Goldinger SD, Pisoni DB, Luce P (1996) Speech perception and spoken word recognition: Research and theory. In: Lass N (ed) Principles of Experimental Phonetics. St. Louis: Mosby, pp. 277-327.

Goldstein JL, Srulovicz P (1977) Auditory nerve spike intervals as an adequate basis for aural spectrum analysis. In: Evans EF, Wilson JP (eds) Psychophysics and Physiology of Hearing. London: Academic Press, pp. 337-346.

Gorin AL, Riccardi G, Wright JH (1997) How may I help you? Speech Comm 23: 113-127.

Grant K, Greenberg S (2001) Speech intelligibility derived from asynchronous processing of auditory-visual information. Proc Workshop Audio-Visual Speech Proc (AVSP-2001), pp. 132-137.

Grant KW, Seitz PF (1998) Measures of auditory-visual integration in nonsense syllables and sentences. J Acoust Soc Am 104: 2438-2450.

Grant KW, Walden BE (1995) Predicting auditory-visual speech recognition in hearing-impaired listeners. Proc XIIIth Int Cong Phon Sci, Vol. 3, pp. 122-125.

Grant KW, Walden BE (1996a) Spectral distribution of prosodic information. J Speech Hearing Res 39: 228-238.

Grant, KW, Walden BE (1996b) Evaluating the articulation index for auditory-visual consonant recognition. J Acoust Soc Am 100: 2415-2424.

Grant KW, Walden BE, Seitz PF (1998) Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. J Acoust Soc Am 103: 2677-2690.

Gravel JS, Ruben RJ (1996) Auditory deprivation and its consequences: From animal models to humans. In: Van De Water TR, Popper AN, Fay RR (eds) Clinical Aspects of Hearing. New York: Springer-Verlag, pp. 86-115.

Greenberg S (1988) The ear as a speech analyzer. J Phon 16: 139-150.

Greenberg S (1995) The ears have it: The auditory basis of speech perception. Proc 13th Int Cong Phon Sci, Vol. 3, pp. 34-41.

Greenberg S (1996a) Auditory processing of speech. In: Lass N (ed) Principles of Experimental Phonetics. St. Louis: Mosby, pp. 362-407.

Greenberg S (1996b) Understanding speech understanding – Towards a unified theory of speech perception. Proc ESCA Tutorial and Advanced Research Workshop on the Auditory Basis of Speech Perception, pp. 1-8.

Greenberg S (1997a) Auditory function. In: Crocker M (ed) Encyclopedia of Acoustics. New York: John Wiley, pp. 1301-1323.

Greenberg S (1997b) On the origins of speech intelligibility in the real world. Proc ESCA Workshop on Robust Speech Recognition in Unknown Communication Channels, pp. 23-32.

Greenberg S (1999) Speaking in shorthand – A syllable-centric perspective for understanding pronunciation variation. Speech Comm 29: 159-176.

Greenberg S (2003) From here to utility – Melding phonetic insight with speech technology. In: Barry W, Domelen W (eds) Integrating Phonetic Knowledge with Speech Technology, Dordrecht: Kluwer, pp. xxx-xxx.

Greenberg S, Ainsworth WA (2003) Listening to Speech: An Auditory Perspective. Hillsdale, NJ: Erlbaum.

Greenberg S, Arai T (1998) Speech intelligibility is highly tolerant of cross-channel spectral asynchrony. Proc Joint Meeting Acoust Soc Am and Int Cong Acoust, pp. 2677-2678.

Greenberg S, Arai T (2001) The relation between speech intelligibility and the complex modulation spectrum. Proc 7th European Conf Speech Comm Tech (Eurospeech-2001), pp. 473-476.

Greenberg S, Arai T, Silipo R (1998) Speech intelligibility derived from exceedingly sparse spectral information, Proc 5th Int Conf Spoken Lang Proc, pp. 74-77.

Greenberg S, Carvey HM, Hitchcock L, Chang S (2002) Beyond the phoneme – A juncture-accent model for spoken language. Proc Human Language Technology Conference, pp. xxx-xxx.

Greenberg S, Chang S (2000) Linguistic dissection of switchboard-corpus automatic speech recognition systems. Proc ISCA Workshop on Automatic Speech Recognition: Challenges for the New Millennium, pp. 195-202.

Greenberg S, Geisler CD, Deng L (1986) Frequency selectivity of single cochlear nerve fibers based on the temporal response patterns to two-tone signals. J Acoust Soc Am 79: 1010-1019.

Greenwood DD (1961) Critical bandwidth and the frequency coordinates of the basilar membrane. J Acoust Soc Am 33: 1344-1356.

Greenwood DD (1990) A cochlear frequency-position function for several species – 29 years later. J Acoust Soc Am 87: 2592-2650.

Greenwood DD (1994) The intensitive DL of tones: Dependence of signal/masker ratio on tone level and spectrum of added noise. Hearing Res 65: 1-39.

Gummer AW, Hemmert W, Zenner HP (1996) Resonant tectorial membrane motion in the inner ear: Its crucial role in frequency tuning. Proc Natl Acad Sci 93: 8727-8732.

Gummer AW, Meyer J, Frank G, Scherer MP, Preyer S (2002) Mechanical transduction in outer hair cells. Audiol Neurootol 7: 13-16.

Halliday MAK (1967) Intonation and Grammar in British English. The Hague: Mouton.

Hauser MD (1996) The Evolution of Communication. Cambridge, MA: MIT Press.

Hauser MD, Chomsky N, Fitch H (2002) The faculty of language: What is it, who has it, and how did it evolve? Science 298: 1569-1579.

Helmholtz HLF. von (1863) Die Lehre von Tonemfindungen als Physiologie Grundlage dur die Theorie der Musik. Braunschweige: F. Vieweg und Sohn. [translated as: On the Sensations of

Tone as a Physiological Basis for the Theory of Music (4th ed., 1897), trans. by A J. Ellis. New York: Dover (reprint of 1897 edition).

Hermansky H (1990) Perceptual linear predictive (PLP) analysis of speech. J Acoust Soc Am 87: 1738-1752.

Hermansky H (1998) Should recognizers have ears? Speech Comm 25: 3-27.

Hermansky H, Morgan N (1994) RASTA processing of speech. IEEE Trans Speech and Audio 2: 578-589.

Houtgast T, Steeneken HJM (1973) The modulation transfer function in room acoustics as a predictor of speech intelligibility. Acustica 28: 66-73.

Houtgast T, Steeneken H (1985) A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. J Acoust Soc Am 77: 1069-1077.

Huggins WH (1952) A phase principle for complex-frequency analysis and its implications in auditory theory. J Acoust Soc Am 24: 582-589.

Humes LE, Dirks DD, Bell TS, Ahlstrom C, Kincaid GE (1986) Application of the Articulation Index and the Speech Transmission Index to the recognition of speech by normal-hearing and hearing-impaired listeners. J Speech Hear Res 29: 447-462.

Irvine DRF (1986) The Auditory Brainstem. Berlin: Springer-Verlag.

Ivry RB, Justus TC (2001) A neural instantiation of the motor theory of speech perception. Trends Neurosci 24: 513-515.

Jakobson R, Fant G, Halle M (1952/1963) Preliminaries to Speech Analysis. Tech Rep 13. Cambridge, MA: Massachusetts Institute of Technology [reprinted by MIT Press, 1963].

Jelinek F (1976) Continuous speech recognition by statistical methods. Proc IEEE 64: 532-556.

Jelinek F (1997) Statistical Methods for Speech Recognition. Cambridge, MA: MIT Press.

Jenison R, Greenberg S, Kluender K, Rhode WS (1991) A composite model of the auditory periphery for the processing of speech based on the filter response functions of single auditory-nerve fibers. J Acoust Soc Am 90: 773-786.

Jesteadt W, Wier C, Green D (1977) Intensity discrimination as a function of frequency and sensation level. J Acoust Soc Am 61: 169-177.

Kakusho O, Hirato H, Kato K, Kobayashi T (1971) Some experiments of vowel perception by harmonic synthesizer. Acustica 24: 179-190.

Kawahara H, Masuda-Katsuse I, de Cheveigné A (1999) Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f_0 extraction: Possible role of a repetitive structure in sounds. Speech Comm 27: 187-207.

Kewley-Port D (1983) Time-varying features as correlates of place of articulation in stop consonants. J Acoust Soc Am 73: 322-335.

Kewley-Port D, Neel A (2003) Perception of dynamic properties of speech: Peripheral and central processes. In: Greenberg S, Ainsworth WA (eds) Listening to Speech: An Auditory Perspective. Hillsdale, NJ: Erlbaum, pp. xxx-xxx.

Kewley-Port D, Watson CS (1994) Formant-frequency discrimination for isolated English vowels. J Acoust Soc Am 95: 485-496.

Kitzes LM, Gibson MM, Rose JE, Hind JE (1978) Initial discharge latency and threshold considerations for some neurons in cochlear nucleus complex of the cat. J Neurophysiol 41: 1165-1182.

Klatt DH (1979) Speech perception: A model of acoustic-phonetic analysis and lexical access. J Phon 7: 279-312.

Klatt DH (1982) Speech processing strategies based on auditory models. In: Carlson R, Granstrom B (eds) The Representation of Speech in the Peripheral Auditory System. Amsterdam: Elsevier.

Klatt D (1987) Review of text-to-speech conversion for English. J Acoust Soc Am 82: 737-793.

Kluender KR (1991) Effects of first formant onset properties on voicing judgments resultfrom processes not specific to humans. J Acoust Soc Am 90: 83-96.

Kluender KK, Greenberg S (1989) A specialization for speech perception? Science 244: 1530 (L).

Kluender KR, Jenison RL (1992) Effects of glide slope, noise intensity, and noise duration on the extrapolation of FM glides though noise. Percept Psychophys 51: 231-238.

Kluender KR, Lotto AJ, Holt LL (2003) Contributions of nonhuman animal models to understanding human speech perception. In: Greenberg S, Ainsworth WA (eds) Listening to Speech: An Auditory Perspective. Hillsdale, NJ: Erlbaum, pp. xxx-xxx.

Knudsen EI (2002) Instructed learning in the auditory localization pathway of the barn owl. Nature 417: 322-328.

Kollmeier B, Koch R (1994) Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction. J Acoust Soc Am 95: 1593-1602.

Kuhl PK, Andruski JE, Chistovich IA, Chistovich LA, Kozhevnikova EV, Ryskina VL, Stolyarova EI, Sundberg U, Lacerda F (1997) Cross-language analysis of phonetic units in language addressed to infants. Science 277: 684-686.

Kuhl PK, Miller JD (1978) Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. J Acoust Soc Am 63: 905-917.

Kuhl PK, Padden DM (1982) Enhanced discriminability at the phonetic boundaries for the voicing feature in Macaques. Percept Psychophys 32: 542-550.

Ladefoged P (1967) Three Areas of Experimental Phonetics. Oxford: Oxford University Press.

Ladefoged P (1971) Preliminaries to Linguistic Phonetics. Chicago: University of Chicago Press.

Ladefoged P (2001) A Course in Phonetics, 4th ed. New York: Harcourt.

Ladefoged P, Maddieson I (1996) The Sounds of the World's Languages. Oxford: Blackwell.

Langer G (1992) Periodicity coding in the auditory system. Hearing Res 60: 115-142

Langner G, Schreiner CE (1988) Periodicity coding in the inferior colliculus of the cat. I. Neuronal mechanisms. J Neurophys 60: 1799-1822.

Lehiste, I. (1996) Suprasegmental features of speech. In: Lass N (ed) Principles of Experimental Phonetics. St. Louis: Mosby, pp. 226-244.

Lenneberg EH (1962) Understanding language without ability to speak: A case report. J Abnormal Soc Psychol 65: 419-425.

Liberman AM, Cooper FS, Shankweiler DS, Studdert-Kennedy M (1967) Perception of the speech code. Psychol Rev 74: 431-461.

Liberman AM, Delattre PC, Gerstman LJ, Cooper FS (1956) Tempo of frequency change as a cue for distinguishing classes of speech sounds. J Exp Psych 52: 127-137.

Liberman AM, Harris KS, Hoffman HS, Griffith BC (1957) The discrimination of speech sounds within and across phoneme boundaries. J Exp Psychol 53: 358-368.

Liberman AM, Mattingly IG (1985). The motor theory of speech perception revised. Cognition, 21: 1-36.

Liberman AM, Mattingly IG (1989) A specialization for speech perception. Science 243: 489-494.

Liberman MC (1988) Response properties of cochlear efferent neurons: Monaural vs. binaural stimulation and the effects of noise. J Neurophys 60: 1779-1798.

Licklider JCR (1951) A duplex theory of pitch perception. Experientia 7: 128-133.

Lieberman P (1984) The Biology and Evolution of Language. Cambridge, MA: Harvard University Press.

Lieberman P (1990) Uniquely Human: The Evolution of Speech, Thought and Selfless Behavior. Cambridge, MA: Harvard University Press.

Lieberman P (1998) Eve Spoke: Human Language and Human Evolution. New York: Norton.

Liljencrants J, Lindblom B (1972) Numerical simulation of vowel quality systems: The role of perceptual contrast. Lang 48: 839-862.

Lindblom B (1983) Economy of speech gestures. In: MacNeilage PF (ed) Speech Production. New York: Springer-Verlag, pp. 217-245.

Lindblom B (1990) Explaining phonetic variation: A sketch of the H & H theory. In: Hardcastle W, Marchal A (eds). Speech Production and Speech Modeling. Dordrecht: Kluwer, pp. 403-439.

Lippmann RP (1996) Accurate consonant perception without mid-frequency speech energy. IEEE Trans Speech Audio Proc 4: 66-69.

Lisker L, Abramson A (1964) A cross-language study of voicing in initial stops: Acoustical measurements. Word 20: 384-422.

Lynn PA, Furst W (1998) Introductory Digital Signal Processing with Computer Applications, 2nd ed. New York: John Wiley.

Lyon R, Shamma SA (1996) Auditory representations of timbre and pitch. In: Hawkins H, Popper, AN, Fay RR (eds) Auditory Computation. New York: Springer-Verlag, pp. 221-270.

McGurk H, MacDonald J (1976) Hearing lips and seeing voices. Nature 264: 746-78.

Massaro DM (1987) Speech Perception by Ear and by Eye. Hillsdale, NJ: Erlbaum.

Mermelstein P (1978) Difference limens for formant frequencies of steady-state and consonant-bound vowels. J Acoust Soc Am 63: 572-580.

Miller GA (1951) Language and Communication. New York: McGraw-Hill.

Miller GA, Nicely PE (1955) An analysis of perceptual confusions among some English consonants. J Acoust Soc Am 27: 338-352.

Miller MI, Sachs MB (1983) Representation of stop consonants in the discharge patterns of auditory-nerve fibers. J Acoust Soc Am 74: 502-517.

Miyawaki K, Strange W, Verbrugge R, Liberman AM, Jenkins JJ, Fujimura O (1975) An effect of linguistic experience: The discrimination of [r] and [l] of Japanese and English. Percept Psychophys 18: 331-340.

Moore BCJ (1997) An Introduction to the Psychology of Hearing, 4th ed. London: Academic Press.

Mozziconacci SJL (1995) Pitch variations and emotions in speech. Proc 13th Intern Cong Phon Sci Vol. 1, pp. 178-181.

Müsch H, Buus S (2001a). Using statistical decision theory to predict speech intelligibility. I. Model structure. J Acoust Soc Am 109: 2896-2909.

Müsch H, Buus S (2001b). Using statistical decision theory to predict speech intelligibility. II. Measurement and prediction of consonant-discrimination performance. J Acoust Soc Am 109: 2910-2920.

Oertel D, Popper AN, Fay RR (2002) Integrative Functions in the Mammalian Auditory System. New York: Springer-Verlag.

Ohala JJ (1983) The origin of sound patterns in vocal tract constraints. In: MacNeilage P (ed) The Production of Speech. New York: Springer-Verlag, pp. 189-216.

Ohala JJ (1994) Speech perception is hearing sounds, not tongues. J Acoust Soc Am 99: 1718-1725.

Ohm GS 1843 Über die definition des Tones, nebst daran geknupfter Theorie der Sirene und ahnlicher Tonbildener Vorrichtungen. Ann D Phys 59: 497-565.

Patuzzi R (2002) Non-linear aspects of outer hair cell transduction and the temporary threshold shifts after acoustic trauma. Audiol Neurootol 7: 17-20.

Pavlovic CV, Studebaker GA, Sherbecoe RL (1986) An articulation index based procedure for predicting the speech recognition performance of hearing-impaired individuals. J Acoust Soc Am 80: 50-57.

Pickett JM (1980) The Sounds of Speech Communication. Baltimore: University Park Press.

Pisoni DB, Luce PA (1987) Acoustic-phonetic representations in word recognition. In: Frauenfelder UH, Tyler LK (eds). Spoken Word Recognition. Cambridge, MA: MIT Press, pp. 21-52.

Plomp R (1964) The ear as a frequency analyzer. J Acoust Soc Am 36: 1628-1636.

Plomp R (1983) The role of modulation in hearing. In: Klinke R (ed) Hearing: Physiological Bases and Psychophysics. Heidelberg: Springer-Verlag, pp. 270-275.

Poeppel D, Yellin E, Phillips C, Roberts TPL, Rowley H., Wexler K, Marantz A (1996) Task-induced asymmetry of the auditory evoked M100 neuromagnetic field elicited by speech sounds. Cognitive Brain Res 4: 231-242.

Pollack I (1959) Message uncertainty and message reception. J Acoust Soc Am 31: 1500-1508.

Pols LCW, van der Kamp LJT, Plomp R (1969) Perceptual and physical space of vowel sounds. J Acoust Soc Am 46: 458-467.

Pols LCW, van Son RJJH (1993) Acoustics and perception of dynamic vowel segments. Speech Comm 13: 135-147.

Popper AN, Fay RR (1992) The Mammalian Auditory Pathway: Neurophysiology. New York: Springer-Verlag.

Proakis JG, Manolakis DG (1996) Digital Signal Processing: Principles, Algorithms and Applications. New York: Macmillan.

Rabinowitz WM, Eddington DK, Delhorne LA, Cuneo PA (1992) Relations among different measure of speech reception in subjects using a cochlear implant. J Acoust Soc Am 92: 1869-1881.

Recanzone GH, Schreiner CE, Merzenich MM (1993) Plasticity of frequency representation in the primary auditory cortex following discrimination training in adult owl monkeys. J Neurosci 13: 87-103.

Reiter ER, Liberman MC (1995) Efferent-mediated protection from acoustic overexposure: Relation to slow effects of olivocochlear stimulation. J Neurophysiol 73: 506-514.

Remez RE, Rubin PE, Berns SM, Pardo JS, Lang JM (1994) On the perceptual organization of speech. Psychol Rev 101: 129-156.

Remez RE, Rubin PE, Pisoni DB, Carrell TD (1981) Speech perception without traditional speech cues. Science 212: 947-950.

Rhode WS, Greenberg S (1994) Lateral suppression and inhibition in the cochlear nucleus of the cat. J Neurophys 71: 493-519.

Rhode WS, Kettner RE (1987) Physiological study of neurons in the dorsal and posteroventral cochlear nucleus of the unanesthesized cat. J Neurophysiol 57: 414-442.

Riesz RR (1928) Differential intensity sensitivity of the ear for pure tones. Phys Rev 31: 867-875.

Rose JE, Brugge JF, Anderson DJ, Hind JE (1967) Phase-locked response to low-frequency tones in single auditory nerve fibers of the squirrel monkey. J Neurophysiol 30: 769-793.

Rosen S, Howell P (1987) Auditory, articulatory, and learning explanations of categorical perception in speech. In: Harnad S (ed) Categorical Perception. Cambridge: Cambridge University Press, pp. 113-160.

Sachs MB, Blackburn CC, Young ED (1988) Rate-place and temporal-place representations of vowels in the auditory nerve and anteroventral cochlear nucleus. J Phon 16: 37-53.

Sachs MB, Young ED (1980) Effects of nonlinearities on speech encoding in the auditory nerve. J Acoust Soc Am 68: 858-875.

Sakoe H, Chiba S (1978) Dynamic programming algorithms optimization for spoken word recognition. IEEE Trans Acoust Speech Sig Proc 26:43-49.

Schalk TB, Sachs MB (1980) Nonlinearities in auditory-nerve responses to bandlimited noise. J Acoust Soc Am 67: 903-913.

Scharf B (1970) Critical bands. In: Tobias JV (ed) Foundations of Modern Auditory Theory, Vol 1. New York: Academic Press, pp. 157-202.

Schreiner CE, Urbas JV (1988) Representation of amplitude modulation in the auditory cortex of the cat. I. The anterior auditory field (AAF). Hearing Res 21: 227-241.

Shamma SA (1985a) Speech processing in the auditory system I: The representation of speech sounds in the responses of the auditory nerve. J Acoust Soc Am 78: 1612-1621.

Shamma SA (1985b) Speech processing in the auditory system II: Lateral inhibition and central processing of speech evoked activity in the auditory nerve. J Acoust Soc Am 78: 1622-1632.

Shamma SA (1988) The acoustic features of speech sounds in a model of auditory processing: Vowels and voiceless fricatives. J Phon 16: 77-91.

Shannon CE, Weaver W (1949) A Mathematical Theory of Communication. Urbana: University of Illinois Press.

Shannon RV, Zeng FG, Kamath V, Wygonski J. (1995) Speech recognition with primarily temporal cues. Science 270: 303-304.

Sharma J, Angelucci A, Sur M (2000) Induction of visual orientation modules in auditory cortex. Nature 404: 841-847.

Shastri L, Chang S, Greenberg S (1999) Syllable detection and segmentation using temporal flow neural networks. Proc 14th Int Cong Phon Sci, pp. 1721-1724.

Shattuck R (1980) The Forbidden Experiment: The Story of the Wild Boy of Aveyron. New York: Farrar Straus Giroux.

Sinex DG, Geisler CD (1983) Responses of auditory-nerve fibers to consonant-vowel syllables. J Acoust Soc Am 73: 602-615.

Skinner BF (1957) Verbal behavior. New York: Appleton-Century-Crofts.

Smith RL (1977) Short-term adaptation in single auditory nerve fibers: Some poststimulatory effects. J Neurophys 40: 1098-1111.

Smoorenburg GF (1992) Speech reception in quiet and in noisy conditions by individuals with noise-induced hearing loss in relation to their tone audiogram. J Acoust Soc Am 91: 421-437.

Sokolowski BHA, Sachs MB, Goldstein JL (1989) Auditory nerve rate-level functions for two-tone stimuli: Possible relation to basilar membrane nonlinearity. Hearing Res 41: 115-124.

Srulovicz P, Goldstein JL (1983) A central spectrum model: A synthesis of auditory-nerve timing and place cues in monaural communication of frequency spectrum. J Acoust Soc Am 73: 1266-1275.

Steinberg JC, Gardner MB (1937) The dependence of hearing impairment on sound intensity. J Acoust Soc Am 9: 11-23.

Stern RM, Trahiotis C (1995) Models of binaural interaction. In: Moore BCJ (ed) Hearing: Handbook of Perception and Cognition. San Diego: Academic Press, pp. 347-386.

Stevens KN (1972) The quantal nature of speech: Evidence from articulatory-acoustic data. In: David EE, Denes PB (eds) Human Communication: A Unified View. New York: McGraw-Hill, pp. 51-66.

Stevens KN (1989) On the quantal nature of speech. J Phon 17: 3-45.

Stevens KN (1998) Acoustic Phonetics. Cambridge, MA: MIT Press.

Stevens KN, Blumstein SE (1978) Invariant cues for place of articulation in stop consonants. J Acoust Soc Am 64: 1358-1368.

Stevens KN, Blumstein SE (1981) The search for invariant acoustic correlates of phonetic features. In: Eimas PD, Miller JL (eds) Perspectives on the Study of Speech. Hillsdale, NJ: Erlbaum, pp. 1-38.

Strange W, Dittman S (1984) Effects of discrimination training on the perception of /r-l/ by Japanese adults learning English. Percept Psychophys 36:131-145.

Studdert-Kennedy M (2002) Mirror neurons, vocal imitation, and the evolution of particulate speech. In: Stamenov M, Gallese V (eds) Mirror Neurons and the Evolution of Brain and Language. Amsterdam: Benjamins John Publishing, pp. xxx-xxx.

Studdert-Kennedy M, Goldstein L (2003) Launching language: The gestural origin of discrete infinity. In: Christiansen M, Kirby S (eds) Language Evolution: The States of the Art. Oxford: Oxford University Press, pp. xxx-xxx.

Suga N (2003) Basic acoustic patterns and neural mechanisms shared by humans and animals for auditory perception. In: Greenberg S, Ainsworth WA (eds) Listening to Speech: An Auditory Perspective. Hillsdale, NJ: Erlbaum, pp. xxx-xxx.

Suga N, Butman JA, Teng H, Yan J, Olsen JF (1995) Neural processing of target-distance information in the mustached bat. In: Flock A, Ottoson D, Ulfendahl E (eds) Active Hearing. Oxford: Pergamon Press, pp. 13-30.

Suga N, O'Neill WE, Kujirai K, Manabe T (1983) Specificity of combination-sensitive neurons for processing of complex biosonar signals in the auditory cortex of the mustached bat. J Neurophysiol 49: 1573-1626.

Sumby WH, Pollack I (1954) Visual contribution to speech intelligibility in noise. J Acoust Soc Am 26: 212-215.

Summerfield Q (1992) Lipreading and audio-visual speech perception. In: Bruce V, Cowey A, Ellis AW, Perrett DI (eds) Processing the Facial Image (eds). Oxford: Oxford University Press, pp. 71-78.

Summerfield AQ, Sidwell A, Nelson T (1987) Auditory enhancement of changes in spectral amplitude. J Acoust Soc Am 81: 700-708.

Sussman, HM, McCaffrey HAL, Matthews SA (1991) An investigation of locus equations as a source of relational invariance for stop place categorization. J Acoust Soc Am 90: 1309-1325.

ter Keurs M, Festen JM, Plomp R (1992) Effect of spectral envelope smearing on speech reception. I. J Acoust Soc Am 91: 2872-2880.

ter Keurs M, Festen JM, Plomp R (1993) Effect of spectral envelope smearing on speech reception. II. J Acoust Soc Am 93: 1547-1552.

Van Tasell DJ, Soli SD, Kirby VM, Widin GP (1987) Speech waveform envelope cues for consonant recognition. J Acoust Soc Am 82: 1152-1161.

van Wieringen A, Pols LCW (1994) Frequency and duration discrimination of short first-formant speech-like transitions. J Acoust Soc Am 95: 502-511.

van Wieringen A, Pols LCW (1998) Discrimination of short and rapid speechlike transitions. Acta Acustica 84: 520-528.

van Wieringen A, Pols LCW (2003) Perception of highly dynamic properties of speech. In: Greenberg S, Ainsworth WA (eds) Listening to Speech: An Auditory Perspective. Hillsdale, NJ: Erlbaum, pp. xxx-xxx.

Velichko VM, Zagoruyko NG (1970) Automatic recognition of 200 words. Int J Man-Machine Studies 2: 223-234.

Viemeister NF (1979) Temporal modulation transfer functions based upon modulation thresholds. J Acoust Soc Am 66: 1364-1380.

Viemeister NF (1988) Psychophysical aspects of auditory intensity coding. In: Edelman G, Gall W and Cowan W (eds) Auditory Function. New York: Wiley, pp. 213-241.

Villchur E (1987) Multichannel compression for profound deafness. J Rehabil Res Dev 24: 135-148.

von Marlsburg C, Schneider W (1986) A neural cocktail-party processor. Biol Cybern 54: 29-40.

Wang MD, Bilger RC (1973) Consonant confusions in noise: A study of perceptual features. J Acoust Soc Am 54: 1248-1266.

Wang WS-Y. (1972) The many uses of f_0. In: Valdman A (ed) Papers in Linguistics and Phonetics Dedicated to the Memory of Pierre Delattre. The Hague: Mouton, pp. 487-503.

Wang WS-Y (1998) Language and the evolution of modern humans. In: Omoto K, Tobias PV (eds) The Origins and Past of Modern Humans. Singapore: World Scientific, pp 267-282.

Warr WB (1992) Organization of olivocochlear efferent systems in mammals. In: Webster DB, Popper AN, Fay RR (eds) The Mammalian Auditory Pathway: Neuroanatomy. New York: Springer-Verlag, pp. 410-448.

Warren RM (2003) The relation of speech perception to the perception of nonverbal auditory patterns. In: Greenberg S, Ainsworth WA (eds) Listening to Speech: An Auditory Perspective. Hillsdale, NJ: Erlbaum, pp. xxx-xxx.

Weber F, Manganaro L, Peskin B, Shriberg E (2002) Using prosodic and lexical information for speaker identification. Proc IEEE Int Conf Audio Speech Sig Proc, pp. xxx-xxx.

Wiener FM, Ross DA (1946) The pressure distribution in the auditory canal in a progressive sound field. J Acoust Soc Am 18: 401-408.

Wier CC, Jestaedt W, Green DM (1977) Frequency discrimination as a function of frequency and sensation level. J Acoust Soc Am 61: 178-184.

Williams CE, Stevens KN (1972) Emotions and speech: Some acoustical factors. J Acoust Soc Am 52: 1238-1250.

Wong S, Schreiner CE (2003) Representation of stop-consonants in cat primary auditory cortex: Intensity dependence. Speech Comm xx: xxx-xxx.

Wright BA, Buonomano DV, Mahncke HW, Merzenich MM (1997) Learning and generalization of auditory temporal-interval discrimination in humans. J Neurosci 17: 3956-3963.

Young ED, Sachs MB (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.

Zec D (1995) Sonority constraints on syllable structure. Phonology 12: 85-129.

Zwicker E (1964) “Negative afterimage” in hearing. J Acoust Soc Am 36: 2413-2415.

Zwicker E (1975) Scaling. In: Keidel W, Neff WD Handbook of Sensory Physiology V. Hearing. Heidelberg: Springer-Verlag, pp. 401-448.

Zwicker E, Flottorp G, Stevens SS (1957) Critical bandwidth in loudness summation. J Acoust Soc Am 29: 548-557.

Figure Legend

1.1 A temporal perspective of speech processing in the auditory system. The time scale associated with each component of auditory and linguistic analysis is shown, along with the presumed anatomical locus of processing. The auditory periphery and brainstem is presumed to engage solely in prelinguistic analysis relevant for spectral analysis, noise robustness and source segregation. The neural firing rates at this level of the auditory pathway are relatively high (100-800 spikes/s). Phonetic and prosodic analyses are probably the product of auditory cortical processing given the relatively long time intervals required for evaluation and interpretation at this linguistic level. Lexical processing probably occurs beyond the level of the auditory cortex, involves both memory and learning. The higher-level analyses germane to syntax and semantics (i.e., meaning) is probably a product of many different regions of the brain and requires hundreds to thousands of milliseconds to complete.

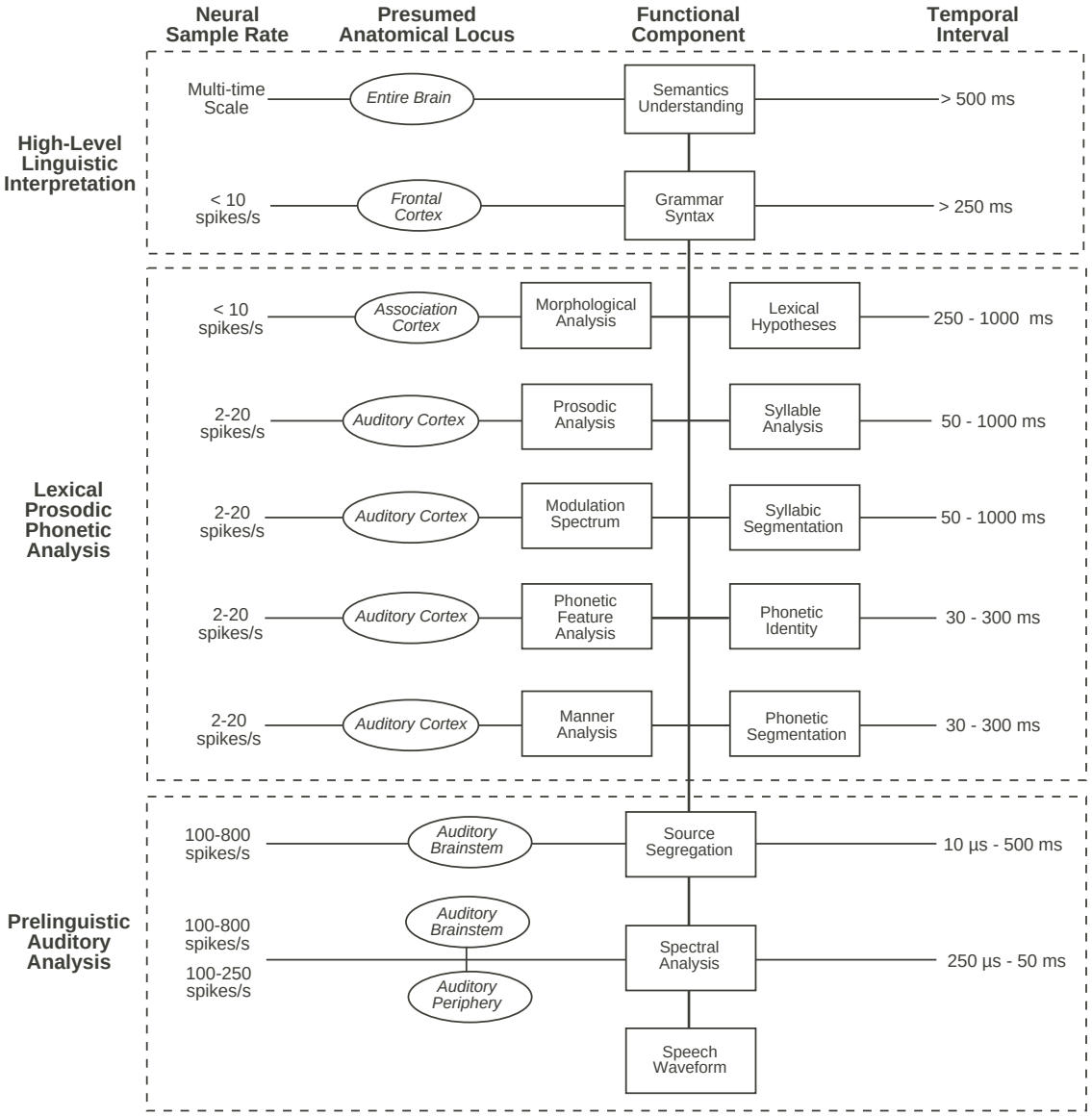

Figure 1.1