Chapter 1

Auditory Processing of Speech

William Ainsworth†
Keele University (deceased)

Steven Greenberg
The Speech Institute

Listening to Speech: An Auditory Perspective

Steven Greenberg and William Ainsworth, editors

Lawrence Erlbaum Associates Publisher

Running head: Auditory processing of speech

Contact Information:

Steven Greenberg, 46 Oxford Drive, Santa Venetia, CA 94903. steveng@cogsci.berkeley.edu

1. INTRODUCTION

Humans are a highly vocal species, and it is our ability to communicate via the spoken word that makes us unique in the animal kingdom (Hauser et al., 2002). Much 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 time (Lieberman, 1990; Wang, 1998).

Despite its importance, our verbal capability is often taken for granted, so seamlessly does it function under so many conditions encountered. The nature of the acoustic background hardly matters – from the chatter of a cocktail party to the roar of a waterfall's descent, humans maintain their ability to verbally interact across a broad spectrum of acoustic environments. Only when our sense of hearing falters does the auditory system's crucial role become apparent. Under such circumstances, the ability to communicate becomes difficult – words "blur," merging with other sounds in the background, and it becomes increasingly difficult to focus on a specific speaker's voice, particularly in noisy and reverberant environments.

The auditory system, in tandem with higher cognitive centers of the brain, performs a remarkable job converting physical pressure variation into a sequence of meaningful elements comprising spoken language. However, the process by

which this transformation occurs is still poorly understood despite decades of scientific investigation.

1.1 The Ear as a Frequency Analyzer

The role of the auditory system has traditionally been viewed as a frequency analyzer (Ohm, 1843; Helmholtz, 1863), of limited precision (Plomp, 1964) providing a faithful spectral representation of the acoustic waveform for higher-level processing. According to Fourier theory, any waveform can be uniquely decomposed into a set of sinusoidal constituents (Lynn & Fuerst, 1998). Fourier analysis makes it possible to describe all speech sounds in terms of energy distributed across frequency and time. From this perspective, the spectrum of a typical vowel is composed 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 vocal tract producing the sound (Fant, 1960).

1.2 Acoustic Theory of Speech Production

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 referred to as “formants” (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 that are potentially important for how they are encoded in the periphery of the auditory pathway. Consonantal segments generally have considerably less energy than vowels. For certain articulatory events, such as articulatory release (accompanying syllable-initial stops) and frication, the energy distribution is rather diffuse, with only a crude delineation of the underlying formant pattern. Some of these segments are voiceless, their waveforms lacking a clear periodic quality that would otherwise reflect the vibration of the vocal folds in the larynx. The amplitude of consonants is typically 30 – 50 dB SPL, which is 20 – 40 dB less intense than their vocalic counterparts (Stevens, 1998). Moreover, the rate of spectral change is generally greater for consonants, and they are often briefer than vocalic segments. Such differences are important for how consonants and vowels are encoded in the auditory system.

1.3 Speech as "Beads on a String"

Within the traditional phonetic framework, words are decomposed into a number of constituent sounds, known as phones (or phonetic segments, and indicated by square brackets "[ ]"). For any given language, the number of distinctive phones is finite, typically between 40 and 60 (Ladefoged and Maddieson, 1996), each with its own distinctive spectral signature. The abstract representation of a phone

is referred to as a "phoneme" (and is usually indicated by back slashes "/ /"). Phonemes are generally used only in abstract descriptions of a language. The phonetic realization of phonemes is usually indicated in terms of phones (i.e., square brackets).

1.4 Articulation Theory

Within the "standard" theory of speech perception, the auditory system's role is viewed primarily as encoding the spectrum, time-frame by time-frame. This encoding is performed in order to provide a faithful representation of the speech signal for conversion into meaningful elements by higher cognitive centers. Within this standard formulation, known as Articulation Theory, speech processing is mainly a matter of frequency analysis (e.g., French & Steinberg, 1947; Fletcher & Gault, 1950; Fletcher, 1953; Pavlovic, Studebaker & Sherbecoe, 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 perspective has been particularly influential in the design of automatic speech recognition systems (see the chapter by Deng & Sheikhzadeh, as well as that of Patterson and colleagues in this volume; for a review see Morgan, Bourlard & Hermansky, 2004), as well as in the development of algorithms for the prosthetic amelioration of sensori-neural hearing loss (see

the chapters by Moore and Faulkner in this volume, as well as Edwards, 2004 and Clark, 2003).

Unfortunately, this view of the ear as a mere frequency analyzer is inadequate for describing the auditory system's ability to process speech. 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.

1.5 Speech Dynamics

Far more is involved in decoding the speech signal than merely computing a conventional frequency analysis (Bronkhorst, 2000). The spectrum of speech changes over time, sometimes slowly, often quickly (Liberman et al. 1956; Kewley-Port 1983; van Wieringen and Pols 1994, 1998, this volume; Kewley-Port and Neel in this volume). These dynamic properties provide information essential for distinguishing among phones (e.g., van Wieringen and Pols 1998, in this volume).

The concept of "time" is important for understanding how speech is processed in the auditory system. Not only the spectrum, but also energy changes over time. It is unusual for a segment's amplitude to remain constant, even over a short span of time. Such energy fluctuation is probably as important as spectral variation (cf.

Van Tassell, Soli, Kirby & Widin, 1987; Drullman, Festen & Plomp, 1994a, 1994b, Drullman in this volume; Kollmeier & Koch 1994; Shannon, Zeng, Kamath & Wygonski, 1995; this volume), for it provides information crucial for segmentation of speech, particularly at the syllabic level (Greenberg, 1996a; Shastri, Chang & Greenberg, 1999).

1.6 Importance of Segmentation

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 considerable energy fluctuation across the acoustic spectrum. Such coarse changes in amplitude may delimit one linguistic unit from the next, irrespective of spectral properties. Smearing the segmentation cues can have a profound impact on speech intelligibility (Drullman et al., 1994a, 1994b; Arai & Greenberg 1998; Greenberg & Arai, 1998), far more so than most forms of spectral distortion (e.g., Licklider 1951; Miller 1951; Blesser 1972). The auditory processes involved in coding syllable-length amplitude fluctuations are likely to play a key role in speech processing (Plomp 1983; Drullman et al. 1994a; Grant & Walden 1996a; Greenberg, 1996a).

1.7 The Many Uses of f_0

Accompanying the 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 & 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 & 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, Manganaro, Peskin & Shriberg, 2002), as well as focusing on a specific talker in a crowded environment (e.g., Brokx & Nooteboom 1982; Cooke & Ellis 2001). In many languages (e.g., Chinese and Thai), pitch (also known as "tone") is used to distinguish among words (Wang, 1972), providing a separate means by which the auditory system plays a key role in the processing of speech.

1.8 The Multiplicity of Speech

Perhaps the most remarkable quality of speech is its multiplicity. Not only are its spectrum, pitch and amplitude constantly changing, but the variation in these properties occur largely independent of each other, and are decoded by the auditory system so effortlessly that rarely are we conscious of the neural

"machinery" underneath the "hood." This multi-tasking capability enables a rich stream of information to be securely transmitted to the higher cognitive centers of the brain (see Greenberg in this volume).

1.9 The Neural Bases of Speech Perception

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 (see the chapters by Sachs and colleagues, Schreiner and colleagues and by Meyer in this volume). The enormous diversity of neuronal response properties in the auditory brainstem, thalamus and cortex (cf. Irvine 1986; Popper & Fay 1992; Oertel, Popper & Fay, 2002; Hackney, in this volume; Adams, in this volume; Budinger & Heil, in this volume) is likely to be relevant for encoding speech and other communicative signals, but the relationship 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 neurophysiology to speech perception. First, it is not yet possible to record from single neuronal elements in the auditory pathway of humans because of the invasive nature of the recording technology. For this reason, current knowledge of hearing is largely limited to

studies in non-human species lacking linguistic capability (Sachs and colleagues, in this volume; Schreiner and colleagues in this volume; Suga, in this volume). Moreover, most of these studies have been performed on anesthesized, non-behaving animals. It is unclear whether such neuronal responses are sufficiently similar to those in the awake preparation for such data to be relevant to speech.

Second, it is difficult to relate neuronal activity in any given region of the auditory pathway with a specific behavior, particularly given the complex nature of spoken language. Many auditory regions are likely to be involved in the analysis and interpretation of sound patterns associated with speech; the relevance of any single anatomical or neuronal site to such complex behavior is necessarily limited.

Ultimately, brain-imaging methods, such as functional magnetic resonance imaging (e.g., Buchsbaum, Hickok & Humphries, 2001) and magnetoencephalography (e.g., Poeppel et al. 1996; Poeppel, 2003), are likely to provide the form of neurological data capable of answering specific questions concerning the relation between speech perception and brain mechanisms. Until the maturation of such technology, much of our knowledge will continue to rely on more indirect methods such as perceptual experiments and modeling studies.

2. AUDITORY NEGLECT IN SPEECH SCIENCE

Historically, the speech signal has been viewed more through the eyes of the vocal tract than the ear. To be sure, biomechanical constraints have played an important role in shaping the acoustic properties of speech. Phylogenetic studies indicate that the human vocal tract has changed dramatically over the course of recent evolutionary history. Comparing the speech production capability of Homo sapiens with our closest biological cousins, the chimpanzee, gorilla and organgutan, makes it clear how quickly human vocal anatomy has changed (Lieberman 1984, 1990, 1998). No ape is capable of speaking, and the vocal repertoire of our closest phylogenetic cousins, the chimpanzees, gorillas and orangutans is impoverished relative to that of humans (Lieberman, 1984). Such rapid changes in vocal anatomy and physiology are probably related to the dramatic expansion of the brain (Wang, 1998). A likely selection factor was the ability to transmit large amounts of information quickly and reliably (Greenberg 1997b).

This dramatic increase in information-transmission capability has been accompanied by relatively small changes in the anatomy and physiology of the auditory system – the structure and function of the auditory pathway does not appear to have changed all that much over the past several million years

(Hackney, in this volume).

Given the conservative design of the auditory system across mammalian species (Fay & Popper, 1994), it is likely that the evolutionary innovations responsible for the evolution of speech were largely shaped by anatomical, physiological and functional constraints imposed by the auditory nervous system in its role as the primary conduit for acoustic information sent to the higher cortical centers of the brain (cf. Ainsworth 1976; Greenberg 1995, 1996a, 1996b, 1997a).

For example, the speed at which we speak (an average of 5 syllables/s) is most likely a consequence of perceptual and information processing constraints, not those imposed by the vocal tract. Although it is possible to speak considerably more slowly than we generally do, listeners usually find such slow speech uncomfortable listening to. Speech can also be produced much faster than the norm, but is often difficult to decode reliably without extreme mental effort. In both instances the vocal tract appears to follow higher-level constraints imposed by the brain as a whole. Such considerations are discussed in two chapters of this book (Todd and colleagues; Greenberg).

3. HOW DOES THE BRAIN GO FROM SOUND TO MEANING?

The process by which the brain proceeds from sound to meaning is not well

understood. Traditionally, speech perception models have assumed that the acoustic signal is decoded segment by segment, analogous to the manner in which words are represented on the printed page (Klatt, 1979; Pisoni & Luce, 1987; Goldinger, Pisoni & Luce, 1996). The segments decoded enable the listener to match the acoustic input to sequence of phonemes stored in the brain's mental lexicon. The auditory system's role is essentially that of a glorified spectrum analyzer.

One problem with this perspective is the enormous acoustic variability observed from one instance of a word to the next. This variability reflects two basic factors – (1) pronunciation variation associated with dialectal and stylistic features (Greenberg, 1999), and (2) environmental forces reflecting reverberation and the acoustic background that shape the speech signal's spectro-temporal properties (Assmann & Summerfield, å2004).

The perceptual invariance associated with a highly variable acoustic signal has intrigued scientists for many years (Perkell & Klatt, 1986) and continues to generate controversy.

The auditory system may hold the key to understanding how the brain so easily appears to understand, given the enormous variability of the physical signal. This volume addresses the issues involved through a detailed consideration

of how hearing functions when listening to speech. Among the questions raised in this book are the following:

(1) What sort of information is conveyed in the acoustic signal? (Chapters 2, 3, 4, 11, 14, 15, 16, 21, 22, 24, 25)

(2) Where is this information located in frequency and time? (Chapters 2, 3, 4, 11, 18, 20, 21, 22, 23, 24, 25)

(3) How is this information encoded in the auditory pathway and other parts of the brain? (Chapters 5, 6, 7, 8, 9, 10, 11, 15, 22, 23, 24)

(4) How are the information-laden portions of the speech signal shielded from the potentially deleterious effects of the acoustic background to ensure reliable and accurate transmission? (Chapters 12, 14, 15, 16, 23)

(5) Which aspects of spoken language are attributable to general auditory processes? (Chapters 11, 12, 13, 22)

(6) Can such knowledge benefit humankind? (Chapters 17, 18, 19, 20)

4. THE SPEECH SIGNAL'S DYNAMIC NATURE

The first two questions are addressed in the section on "Acoustic and Perceptual Cues Germane to the Perception of Speech." Traditionally, the speech signal has been likened to a conveyor belt of segments – "phonemes" – spliced together, as "beads on a string," to form words. The concept, known as "coarticulation,"

dispels this myth. Articulatory gestures often span several segments, and it is not uncommon for vowels and consonants in the same syllable to share certain features in common. In their chapter, van Wieringen and Pols discuss some of these acoustic properties, concentrating on brief (20-ms–50-ms), formant-like transitions typically found in such voiced plosives as [b], [d] and [g], as well as on sinusoidal (i.e., tonal) contours. In particular, they focus on the listener's ability to distinguish among transitions of variable duration and frequency change. Increasing the transition's duration enhances the ability to distinguish the endpoint frequency of a tonal contour, but this precision is not so useful in speech. The ability to perform fine frequency discrimination in speech is not nearly as precise as for simple spectral stimuli like tone glides. This tradeoff between stimulus complexity and perceptual precision may be of particular importance where generalization across variable signal conditions is of particular importance.

The dynamic nature of speech is manifest not only in the spectrum, but also in its amplitude modulation. In the 1930s, Homer Dudley exploited this property for the VOCODER, an early version of the speech synthesizer (Dudley, 1939). Dudley distinguished the "carrier" property of speech from the "modulator" (Dudley, 1940). The carrier is analogous to the modern-day concept of spectrum (or more accurately, the time-domain reflection of the local spectrum), while the

modulator reflects slow fluctuations in energy correlated with syllable-length articulatory gestures. Dudley found that the meaningful components of speech could be reduced to slow modulations below 25 Hz, distributed across as few as 10 distinct frequency channels. This sparse representation of the speech signal was both highly intelligible and natural sounding, resulting in a considerable reduction of data required to synthesize the human voice (Dudley, 1939).

In the 1970s, Houtgast and Steeneken (1973) used the carrier/modulator concept to predict the intelligibility of room auditoria. The modulation spectrum's magnitude between 2 and 10 Hz (called the "speech transmission index" or STI) was shown to correlate highly with the ability to understand speech over a broad range of listening conditions, including background noise and reverberation (Houtgast and Steeneken, 1985). Plomp and his associates adapted the STI for their studies on the loss of intelligibility among the hearing impaired (Plomp and Mimpen, 1979; Plomp, 1983, 2002).

Clearly, slow modulation of energy in the speech signal is important for understanding spoken language. But, what precisely, is its relevance? Drullman addresses this issue in his chapter on "The Significance of Temporal Modulation Frequencies for Speech Intelligibility." By selectively low-pass or high-pass filtering the amplitude fluctuations contained in the signal, he is able to

demonstrate that the modulation spectrum between 2 and 16 Hz is essential for understanding spoken language. Moreover, the higher modulation frequencies, those greater than 8 Hz, are particularly important for distinguishing among certain consonants. His data are consistent with the notion that understanding speech requires multiple time spans of analysis, ranging from the syllable (150-400 ms) to the phonetic segment (60-120 ms).

The importance of learning how to interpret speech dynamics is emphasized in the chapter by Kewley-Port and Neel entitled "Perception of Dynamic Properties of Speech: Peripheral and Central Processes." They find that the discriminability (i.e., difference limens) of formant patterns depends on training and stimulus uncertainty. Listeners can perform extremely well under such circumstances when highly experienced in the task. For this reason, they suggest that higher cognitive processes may play an important role in interpreting the acoustic properties of the speech signal. This is an important message, one that has often been neglected in the study of speech.

5. THE STRUCTURAL AND PHYSIOLOGICAL BASIS OF SPEECH PERCEPTION

The auditory pathway is the primary sensory system associated with spoken language. As noted earlier, mammalian audition has changed relatively little over

the past several million years. Cats, monkeys and humans utilize a common structural plan for processing sounds. This commonality allows us, as speech scientists, to use our knowledge of auditory anatomy and physiology derived from study of non-human species to better understand the relation between hearing function and speech processing.

The section on "Anatomical and Physiological Bases of Speech Perception" examines the structural foundations of hearing in a wide range of mammalian species, but with particular emphasis on humans.

Hackney reviews what is currently known about auditory anatomy in her chapter entitled "From Cochlea to Cortex: a Simple Anatomical Description." The chapter commences with the inner ear, where the transduction process essential to detecting and analyzing sound is firmly established. The frequency-selective properties of hearing originate in the cochlea, where the basilar membrane forms a structural foundation for a spectral (and spatial) gradient that is preserved from the periphery (i.e., auditory nerve) all the way on up to the highest levels of the pathway, in primary auditory cortex. As early as the cochlear nucleus (to which the auditory nerve projects), there is considerable morphological and physiological diversification that is of likely significance for speech processing. Anatomical complexity increases even more in the central portions of the auditory

pathway.

This anatomical complexity forms the focus of Joe Adams' chapter on "Neuroanatomical Considerations of Speech Processing." He is particularly concerned with how the morphological diversity of neurons in the auditory brainstem could be functionally exploited for processing complex sounds, such as speech. The parallel-processing characteristic of spoken language may be reflected in the anatomical diversity observed in the auditory pathway. Adams believes that the precise timing of acoustic events encoded in the auditory brainstem may be of particular importance for reliable speech representation.

Budinger and Heil focus on the intricate circuitry of the highest level of the hearing pathway in their chapter on the "Anatomy of the Auditory Cortex." One of the striking aspects of auditory anatomy is the amount of feedback; there are as many fibers descending from the cortex into the brainstem as afferent projections transmitting information from brainstem to thalamus and cortex. There is also an intricate system of trans-cortical connectivity that is likely to be important for stabilizing the processing of complex signals such as speech. Clearly, we are only beginning to understand the complex morphology of auditory cortex, a region that probably holds the key to solving many puzzles in speech processing.

The physiological basis of speech processing is discussed in two separate

chapters. In the first, Sachs and colleagues consider whether the average firing rate of auditory-nerve fibers can provide the sort of detailed representation required to distinguish among vowels in their chapter entitled "Adequacy of Auditory-nerve Rate Representations of Vowels: Comparison with Behavioral Measures in Cat." They demonstrate that there is, indeed, sufficient dynamic range in the response of auditory-nerve fibers to encode the spectra of vocalic segments with sufficient precision to account for human perceptual capabilities. Thus, it may not be necessary to invoke special temporal or "selective listening" mechanisms to account for vocalic processing in human listeners.

In their chapter on "Temporal Processing in Cat Primary Auditory Cortex: Dynamic Frequency Tuning and Spectro-temporal Representation of Speech Sounds," Schreiner and colleagues examine the representation of speech at the level of the auditory cortex. They demonstrate that there are systematic correlates of phonetic features (e.g., voicing) in cortical responses. However, the neuronal responses are more subtle than those observed in the auditory periphery and brainstem. Discharge rates are considerably lower (just a few spikes per second), and the precise time of discharge may be a particularly important carrier of information.

Georg Meyer reviews the work discussed in the preceding chapters on

anatomy and physiology from a computational perspective in "Anatomical and Physiological Bases of Speech Perception." He discusses a model of cochlear function designed to process frequency and amplitude modulation important for characterizing speech and other complex sounds. Of particular interest is the question of whether speech is truly special, requiring language-specific processing, or can be understood within a general auditory framework capable of processing non-speech sounds as well.

6. NEUROETHOLOGICAL PARALLELS TO SPEECH PROCESSING

Phylogenetically ancient, the mammalian auditory system has changed relatively little over the course of many millions of years. As mentioned in the previous section, the anatomy and physiology of many auditory regions are similar across mammals. It is therefore of interest to examine potential non-human parallels to language in order to gain a broader perspective on the relation between hearing and vocal communication.

Nobuo Suga, in his chapter on "Basic Acoustic Patterns and Neural Mechanisms Shared by Humans and Animals for Auditory Perception" describes in detail parallels between the acoustic signals emitted by certain species of bat and human speech sounds. Suga repeatedly emphasizes the importance of "information-bearing elements" (IBEs) in both forms of vocal communication,

correctly focusing on those aspects of the acoustic signal that provide discriminative potential (what is referred to as [negative] entropy in information theory). In Suga's views the constant-frequency (CF) and frequency-modulated (FM) components of many bat vocalizations are similar to certain components of human speech. He and his colleagues have mapped the representation of such IBEs in the bat auditory cortex; it is likely that a comparable mapping exists for speech sounds in human auditory cortex.

Humans are thought to have evolved in the forests and savannah of equatorial Africa (Tattersall, 1998). The acoustic environments associated with these geographical regions are characterized are highly reverberant and carry a lot of background noise (particularly in the forest). Charles Brown and Joan Sinnott examine the implications of humans' equatorial origins for understanding how the auditory system may have evolved for processing communication systems, including speech. In their view, one of the most important characteristics of speech is its robust nature in a wide range of acoustic environments and may reflect humans' African origins. In the second part of their chapter, Brown and Sinnott compare the frequency and intensity discrimination capabilities of humans and monkeys. Intensity discrimination performance is similar across the two species. In contrast, humans are vastly superior with respect to frequency

discrimination, particularly for spectrally simple signals. The comparative psychophysical data suggest that much of the speech-processing capability in humans may reflect general auditory processes rather than specialized, mechanisms.

Keith Kluender and his colleagues examine the issues discussed in the two preceding chapters from a general learning perspective. In their view, much of the experimental evidence garnered from non-human subjects suggests that many linguistic properties of speech and other vocal communication systems (such as categorical perception) reflect general auditory mechanisms rather than specific-specific feature detectors. In their view, "there is no reason to believe that general principles of learning, whether instantiated in nonhuman animals, or in computational models using covariance matrices or connectionist networks, or in real neural networks will be inadequate to explain the structure of linguistic functional equivalence classes for speech sounds." Human language evolved from other forms of acoustic communication, and the seeds of such a sophisticated system lie in these nonhuman systems.

7. ROBUSTNESS OF SPOKEN LANGUAGE IN ADVERSE ACOUSTIC ENVIRONMENTS AND ITS UTILITY FOR SPEECH RECOGNITION

One of the remarkable properties of spoken language is its durability under a wide range of speaking conditions. Were it not for speech's robustness, language night not have had the evolutionary impact it did. Afterall, if one always needs to converse in a "cone of silence" in order to be understood, the utility of verbal communication is severely compromised. One of the great strengths of spoken language is its "anywhere, anytime" characteristic – under a waterfall, in the middle of Grand Central Station, at a crowded, noisy restaurant, and so on. People enjoy talking without regard to where they are.

The next section of the book on "Robustness of Spoken Language" examines this crucial aspect of speech communication. The first chapter, by Summerfield and colleagues on "The Perception of Speech under Adverse Conditions: Contributions of Spectro-temporal Peaks, Periodicity and Inter-aural timing to Perceptual Robustness" addresses the issue of speech's microstructure. Voicing, associated with the vibration of the vocal folds, plays an important role in allowing listeners to track the speech signal in noisy environments. The tracking utilizes cues based directly on the signal's fundamental frequency as well as

interaural timing information. In addition, voicing helps to enhance the amplitude of key regions of the speech (the formant patterns) signal carrying important linguistic information.

The section's two other chapters examine temporal cues for robustness from a computational perspective. Li Deng and Hamid Sheikhzadeh have developed an elegant model of the cochlea as a front-end processor for an automatic speech recognition system. In their chapter on the "Use of Temporal Codes Computed from a Cochlear Model for Speech Recognition" they examine the utility of neural synchrony for creating noise-robust representations of speech (particularly the voiced parts associated with vowels). As the basis of their model, they use population neural responses, similar to those discussed in the chapter by Murray Sachs and colleagues. They fine-tune the model parameters to conform to the population discharge patterns in the auditory nerve and use this information to derive estimates of the spectrum on a frame-by-frame basis. This method of spectral estimation is extremely effective in noise and improves the performance of automatic speech recognition dramatically.

The importance of binaural processing is examined by Roy Patterson and colleagues in their chapter on "Binaural Auditory Images for Noise-Resistant

Speech Recognition." They demonstrate that cross correlation of timing information from the two ears provides an extremely robust representation of the speech signal in noise, and that this binaural auditory image can be used effectively for phonetic-segment recognition in exceedingly noisy environments. The conclusions echo those of Summerfield and colleagues, who also found significant advantages for binaural information.

8. SPEECH PERCEPTION BY THE HEARING AND LANGUAGE IMPAIRED

For the hearing impaired, noise is a key issue. Most of the communication problems encountered by such persons are associated with noisy and reverberant environments. Can insights garnered from basic research, such as those described in chapters of this book, be used to help the hearing impaired? Three of the four chapters in the section on "Speech Perception by the Hearing and Language Impaired" deal directly with this topic. Brian Moore in his chapter on "Factors Affecting Speech Intelligibility for People with Cochlear Hearing Loss," demonstrates that the decline in intelligibility experienced by the hearing impaired can't be entirely explained by a loss of audibility. There's far more to hearing loss than merely elevation of thresholds for pure tones. In his view, several other factors are equally important – frequency selectivity, loudness recruitment and the

presence of "dead" regions in the cochlea. Frequency selectivity pertains to how well the spectrum of speech is represented. Poor frequency selectivity results in degradation of auditory representation of formant patterns carrying linguistic information. Loudness recruitment is associated with the dynamic range of speech. The hearing impaired typically have a dynamic range on the order of 10-20 dB in affected regions, whereas the speech signal varies by 40-50 dB from softest to loudest sounds. This is one of the reasons why digital compression aids have been able to help many of the hearing impaired. Finally, "dead" regions – those areas of the cochlea where there has been complete degeneration of hair cells – make it difficult to restore spectral information associated with the affected frequencies.

Based on such information, Moore is able to simulate in normal hearing individuals the effect of hearing impairment using sophisticated signal processing methods. Some improvement in speech intelligibility is observed in such listeners when methods are used to reverse some of the frequency-selectivity impairment observed in the hearing impaired. However, such techniques are of limited utility in the presence of considerable background noise.

Andrew Faulkner and Stuart Rosen in their chapter on "Speech Perception and Auditory Impairment: The Roles of Temporal and Spectral Information" raise

some important issues that complement Moore's discussion. Visual cues (a.k.a. "speechreading"), derived from motion of the lips, jaw, teeth and tongue is a crucial adjunct to the acoustic signal, and can be used, in their experience, to boost intelligibility for the hearing impaired, particularly in noisy environments. They also find that voicing can be an extremely useful cue to the hearing impaired for following speech in noisy environments.

Hearing aids are designed for those individuals with some residual degree of auditory capacity. Cochlear implants, on the other hand, are intended for those with little or no hearing capacity due to complete degeneration of hair cells in the inner ear. It is such patients that Robert Shannon and his colleagues study on a daily basis. Over the past twenty years, tremendous strides have been taken in developing technology capable of restoring some semblance of auditory sensation. These cochlear implants allow many patients to converse fluently, even over the telephone (with restricted spectral information and the absence of speechreading cues). Given the primitive nature of the electrical stimulation pattern generated by such implants, how are patients able to use such sparse information to effectively decode the speech signal?

This is the primary issue addressed by Shannon and his colleagues. In their view, speech understanding can be likened to visual image processing. In both domains, the physical stimulus is often a coarse caricature of the original. However, the coarseness does not impair a person's ability to accurately match the input with a stored template. In the case of language, the input is sound and the template is some abstract representation of words. If the acoustic cues for speech sounds and words were not coarse it would be difficult to explain the success of cochlear implant technology.

Shannon describes some elegant experiments in which he and his colleagues attempt to zero in on the defining acoustic information for intelligibility. In their view, the low-frequency modulation of energy is extremely important – just as Rob Drullman's studies suggest they are for normal listening conditions. However, it is not just the amplitude modulation pattern that's essential. This information needs to be organized in such a way that the modulation patterns associated with different portions of the spectrum are preserved in their relation to each other. If these relational cues are disrupted, then intelligibility declines dramatically.

Certain forms of language impairment do not appear to have a cochlear origin. Beverly Wright, in her chapter on "Perceptual Learning of Temporally Based Auditory Skills Known to Be Deficient in Children with Specific Language Impairment," examines a specific population of children who appear normal in most ways, but who exhibit a specific deficit dealing with spoken language. Her studies suggest there is some high-level cognitive impairment pertaining to organizing sequential units of information. The deficit is most apparent in cases where listeners are asked to identify specific sounds that are quickly followed by other sounds. Children with this form of language impairment have difficulty with this identification task even when the interval between the target and masker is relatively long (hundreds of milliseconds). To the extent that the ability to accurately identify sequences of sounds is important for spoken language (see the chapter by Warren in this volume for why this is so), it is not surprising that these children have difficulty with speech.

In the second part of her chapter, Wright addresses the issue of learning. Can individuals be trained to improve their ability to identify sequences of brief sounds. In her view, the answer is clearly "yes." Such training may ultimately be able to help children with specific language impairment.

9. AUDITORY SCENE ANALYSIS AND THE PERCEPTUAL ORGANIZATION OF SPEECH

As the chapters in the previous section attest, there is a lot more to auditory perception than just audibility. In order to make sense of the acoustic world, it is important to identify and order objects relative to their environmental context. The blind perform this remarkable feat during the course of everyday life. The hearing also achieve something akin to auditory scene analysis, though supplemented with visual information.

Richard Warren, in his chapter on "The Relation of Speech Perception to the Perception of Nonverbal Auditory Patterns," makes it clear that time is an extremely important parameter for speech decoding. The ability to temporally order linguistic elements is crucial for understanding spoken language and requires a minimum interval to accurately perform. Still, sequences of short elements can be distinguished from one another even if the order of the units is not specifiable. These "temporal compounds" are analogous in certain ways to syllables, in that listeners often don't decode each phone independently of the others.

There are other ways in which listeners are able to deduce the entire pattern from fragmentary information. Large portions of the spectrum can be discarded;

under the right conditions most of the words spoken are still heard. Extraneous noises can be introduced without any significant impact on intelligibility – often, the listener is completely unaware of the noise itself. Through various forms of "induction" the hearer is able to infer what is said. Warren quite correctly makes the point that this remarkable ability could not occur without very sophisticated information processing going on at a very high level of the brain.

What might some of these higher-level mechanisms look like? Todd and his colleagues propose "A Sensory-motor Theory of Speech Perception" as one possibility. In their view, the phonetic constituents of speech are secondary to the rhythmic properties associated with prosody. In a language such as English, prosody takes two basic forms. One is stress, which pertains to the relative emphasis placed on syllables in an utterance. The other is intonation, which applies to sequences of syllables in a phrase and is used to parse the utterance into coherent units of meaning and nuance.

The time scales of stress and intonation are long relative to those associated with phones. Stress requires two to three syllables to acquire any force – a minimum of 400-600 ms, while intonation generally operates on time scales of a second or longer. In Todd's theory, these long time scales are the result of low-frequency modulation filters. The auditory system (and other parts of the brain)

operates concurrently on many time scales. The prosodic intervals are essentially integration windows that incorporate a lot of phonetic information. They also include information about the motion of the visual articulators, such as the lips and the jaw, which are known to be important for decoding the speech signal. The time constants characterizing hearing and vision are similar, as are those associated with the motor system. Therefore, it is natural to assume that the brain integrates information from the senses and the motor system to generate a global representation of the world. Todd and his colleagues believe that much of the auditory cortex is organized in precisely the manner required to perform this sort of multi-time-scale analysis.

A somewhat different approach is taken by Brown and Wang in their chapter on "Timing is of the Essence: Neural Oscillator Models of Auditory Grouping." They are concerned with how the brain learns to group elements of a scene together as one. In the acoustic realm, time must play an important role in binding auditory objects given the largely sequential nature of modality. In Brown and Wang's view, individual spectral components are bound into a single object (e.g., a voiced speech sound) through the operation of neural oscillators, which inform the brain which temporal properties are shared by various elements. Like the authors of many other chapters in this book, they believe that timing provides a

crucial set of cues for analysis of sound. In this instance, they suggest that the fundamental frequency of speech could be used as a cue for grouping. Cortical oscillators would use such timing information to track elements in the speech signal with common trajectories.

Dan Ellis discusses "Modeling the Auditory Organization of Speech" in his perspective chapter. He points out that the computational approaches described by Todd and colleagues, as well as by Brown and Wang, have their pros and cons. Until more is known about the physiology and anatomy of the auditory pathway it will be difficult to decide which approach makes the most sense. At present, human listeners perform much better than computational models in segregating speech from the acoustic background. Ellis also suggests that we have much to learn from the studies discussed by Warren. So much can be missing from the signal without significant impact on speech understanding – surely there is some powerful computational machinery responsible for this feat.

10. A MULTI-TIER FRAMEWORK FOR UNDERSTANDING SPOKEN LANGUAGE

The material described in this book span the fields of anatomy, physiology, linguistics, psychology, acoustics, information theory, and computer science. To fully comprehend the scope of spoken language it is necessary to traverse all of

these fields. Yet, it is rare for theoretical frameworks to encompass all of these domains.

In the concluding chapter, Greenberg summarizes what is currently known about speech perception and places this knowledge within the framework of a unified theory. The theory describes how listeners go from "sound to meaning" by defining the relevant units of analysis at a variety of levels – acoustic, phonetic, phonological, syllabic, lexical, prosodic and semantic. The key is to delineate how the levels interact with each other to yield something that we all recognize as language.

In Greenberg's view, there has been insufficient emphasis placed on the syllable as a key unit of integration and analysis (but see the chapters by Todd and colleagues and by Warren). Like Todd, Greenberg believes that prosody has been unjustly ignored as a key organizing principle. But he goes farther, by proposing specific strategies the brain could use to relate acoustic and visual patterns to elements commonly referred to as "phones," "syllables" and "words." In Greenberg's view the phone (and phoneme) is not the most accurate means to describe what routinely occurs in spoken language. Prosodic parameters interact with articulatory features (such as place and manner of articulation, and voicing) at the syllabic level to create a pattern that phoneticians have traditionally labeled

as a phonetic sequence. Greenberg's theoretical framework is able to account for many otherwise puzzling properties of speech, including the specific patterns of pronunciation variability observed in casual conversation, historical sound change and the relation between phonological form and words. Although the theory is linguistic in approach and application, it is firmly grounded in the physiology and anatomy of the auditory pathway. Only time will tell whether such a multi-tier perspective provides a truly explanatory framework for spoken language.

11. CONCLUSION

As the chapters in this book attest, speech communication is a highly complex phenomenon, incapable of being fully understood from a single perspective. The speech scientists of tomorrow will likely be trained in a broad array of disciplines, ranging from physics and mathematics to computer science, anatomy, physiology, linguistics and psychology.

The preceding century was an era of remarkable achievement in atomic physics and molecular biology. The twenty-first century is likely to be recalled as "the age of communication," in which the frontiers of computer science and speech technology were repeatedly advanced. Flawless recognition and synthesis of speech will require a level of knowledge and insight far greater than we currently possess, as will hearing aids and auditory implants that fully restore

hearing. Once an objective has been clearly stated, it often takes a relatively short time to accomplish the goal. Hopefully, this is the trajectory of the speech and hearing sciences over the coming decades.

REFERENCES

Ainsworth, W. A. (1976). Mechanisms of speech recognition. Oxford: Pergamon Press.

Ainsworth, W. A. (1986). Pitch change as a cue to syllabification. Journal of Phonetics, 14, 257-264.

Ainsworth, W. A., & Lindsay, D. (1986). Perception of pitch movements on tonic syllables in British English. Journal of the Acoustical Society of America, 79, 472-480.

Allen, J. B. (1994). How do humans process and recognize speech? IEEE Transactions on Speech and Audio Processing, 2, 567-577.

Assmann, P., & Summerfield, A. Q. (2004) The perception of speech under adverse conditions. In S. Greenberg, W. A. Ainsworth, A. N. Popper & R. R. Fay (Eds.), Speech processing in the auditory system (pp. 231-308). New York: Springer-Verlag.

Blesser, B. (1972). Speech perception under conditions of spectral transformation. I. Phonetic characteristics. Journal of Speech and Hearing Research, 15, 5-41.

Brokx, J. P. L., & Nooteboom, S. G. (1982). Intonation and the perceptual separation of simultaneous voices. Journal of Phonetics, 10, 23-36.

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

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

Clark, G. M. (2003). Cochlear implants: Fundamentals and applications. New York: Springer-Verlag.

Cooke, M., & Ellis, D. P. W. (2001). The auditory organization of speech and other sources in listeners and computational models. Speech Communication, 35, 141-177.

Drullman, R., Festen, J. M., & Plomp, R. (1994a). Effect of temporal envelope smearing on speech reception. Journal of the Acoustical Society of America, 95, 1053-1064.

Drullman, R., Festen, J. M., & Plomp, R. (1994b). Effect of reducing slow temporal modulations on speech reception. Journal of the Acoustical Society of America, 95, 2670-2680.

Dudley H (1939) Remaking speech. Journal of the Acoustical Society of America, 11, 169-177.

Dudley, H. (1940). The carrier nature of speech. Bell System Technical Journal, 19, 495-515.

Edwards, B. (2004). Hearing aids and hearing impairment. In S. Greenberg, W. A. Ainsworth, A. N. Popper & R. R. Fay (Eds.), Speech processing in the auditory system (pp. 339-421). New York: Springer-Verlag.

Fant, G. (1960). Acoustic theory of speech production. The Hague: Mouton.

Fay, R. R., & Popper, A. N. (Eds.), (1994). Comparative hearing: Mammals. New York: Springer-Verlag.

Fletcher, H. (1953). Speech and hearing in communication. New York: Van Nostrand.

Fletcher, H., & Gault, R. H. (1950). The perception of speech and its relation to telephony. Journal of the Acoustical Society of America, 22, 89-150.

French, N, R., & Steinberg, J. C. (1947). Factors governing the intelligibility of speech sounds. Journal of the Acoustical Society of America, 19, 90-119.

Goldinger, S. D., Pisoni, D. B., & Luce, P. (1996). Speech perception and spoken word recognition: Research and theory. In N. Lass (Ed.), Principles of experimental phonetics (pp. 277-327). St. Louis: Mosby.

Grant, K. W., & Walden, B. E. (1996a) Spectral distribution of prosodic information. Journal of Speech and Hearing Research, 39, 228-238.

Greenberg, S. (1995). The ears have it: The auditory basis of speech perception. Proceedings of the 13th Internaltional Congress of Phonetic Sciences, 3, 34-41.

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

Greenberg, S. (1996b). Auditory processing of speech. In N. Lass (Ed.), Principles of experimental phonetics (pp. 363-407). St. Louis: Mosby.

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

Greenberg, S. (1997b). On the origins of speech intelligibility in the real world. Proceedings of the 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 Communication, 29, 159-176.

Greenberg, S., & Arai, T. (1998). Speech intelligibility is highly tolerant of cross-channel spectral asynchrony. Proceedings of the Joint Meeting of the Acoustical Society of American and the International Congress on Acoustics, pp. 2677-2678.

Hauser, M. D., 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, H. L. F. 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), translated by A J. Ellis. New York: Dover (reprint of 1897 edition).

Houtgast, T., & Steeneken, H. J. M. (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. Journal of the Acoustical Society of America, 77, 1069-1077.

Irvine, D. R. F. (1986). The auditory brainstem. Berlin: Springer-Verlag.

Kewley-Port, D. (1983). Time-varying features as correlates of place of articulation in stop consonants. Journal of the Acoustical Society of America, 73, 322-335.

Klatt, D. H. (1979). Speech perception: A model of acoustic-phonetic analysis and lexical access. Journal of Phonetics, 7, 279-312.

Kollmeier, B., & Koch, R. (1994). Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction. Journal of the Acoustical Society of America, 95, 1593-1602.

Ladefoged, P., & Maddieson, I. (1996). The sounds of the world's languages. Oxford: Blackwell.

Lehiste, I. (1996). Suprasegmental features of speech. In N. Lass (Ed.), Principles of experimental phonetics (pp. 226-244). St. Louis: Mosby.

Liberman, A. M., Delattre, P. C., Gerstman, L. J., & Cooper, F. S. (1956). Tempo of frequency change as a cue for distinguishing classes of speech sounds. Journal of Experimental Psychology, 52, 127-137.

Licklider, J. C. R. (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.

Lynn, P. A., & Furst, W. (1998). Introductory digital signal processing with computer applications (2nd ed.). New York: John Wiley.

Miller, G. A. (1951). Language and communication. New York: McGraw-Hill.

Morgan, N., Bourlard, H., & Hermansky, H. (2004). Automatic speech recognition: An auditory perspective. In S. Greenberg, W. A. Ainsworth, A. N. Popper & R. R. Fay (Eds.), Speech processing in the auditory system (pp. 309-338). New York: Springer-Verlag.

Oertel, D., Popper, A. N., & Fay, R. R. (Eds.) (2002). Integrative functions in the mammalian auditory system. New York: Springer-Verlag.

Ohm, G. S. (1843). Über die definition des Tones, nebst daran geknupfter Theorie der Sirene und ahnlicher Tonbildener Vorrichtungen. Annalen der Physik, 59, 497-565.

Pavlovic, C.V., Studebaker, G. A., & Sherbecoe, R. L. (1986). An articulation index based procedure for predicting the speech recognition performance of hearing-impaired individuals. Journal of the Acoustical Society of America, 80, 50-57.

Perkell, J. S., & Klatt, D. H. (Eds.) (1986). Invariance and variability in speech processes. Hillsdale, NJ: Lawrence Erlbaum Associates.

Pisoni, D. B., & Luce, P. A. (1987). Acoustic-phonetic representations in word recognition. In U. H. Frauenfelder & L. K. Tyler (Eds), Spoken word recognition (pp. 21-52). Cambridge, MA: MIT Press.

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

Plomp, R. (2002). The intelligent ear: On the nature of sound perception. Mahwah, NJ: Lawrence Erlbaum Associates.

Plomp, R., & Mimpen, A. M. (1979). Improving the reliability of testing the speech reception threshold for sentences. Audiology, 8, 43-52.

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

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: Cerebral lateralization as "asymmetric sampling in time." Speech Communication, 41, 245-255.

Popper, A. N., & Fay, R. R. (1992). The mammalian auditory pathway: Neurophysiology. New York: Springer-Verlag.

Shannon, R. V., Zeng, F. G., Kamath, V., & Wygonski, J. (1995). Speech recognition with primarily temporal cues. Science, 270, 303-304.

Shastri, L., Chang, S., & Greenberg, S. (1999). Syllable detection and segmentation using temporal flow neural networks. Proceedings of the 14th International Congress of Phonetic Sciences, pp. 1721-1724.

Stevens, K. N. (1998). Acoustic phonetics. Cambridge, MA: MIT Press.

Tattersall, I. (1998). Becoming human: Evolution and human uniqueness. Orlando, FL: Harcourt.

Van Tasell, D. J., Soli, S. D., Kirby, V. M., & Widin, G. P. (1987). Speech waveform envelope cues for consonant recognition. Journal of the Acoustical Society of America, 82, 1152-1161.

van Wieringen, A., & Pols, L. C. W. (1994). Frequency and duration discrimination of short first-formant speech-like transitions. Journal of the Acoustical Society of America, 95, 502-511.

van Wieringen, A., & Pols, L. C. W. (1998). Discrimination of short and rapid speechlike transitions. Acta Acustica, 84, 520-528.

Wang, W. S.-Y. (1972). The many uses of f_0. In A. Valdman (Ed.), Papers in linguistics and phonetics dedicated to the memory of Pierre Delattre (pp. 487-503). The Hague: Mouton.

Wang, W. S.-Y. (1998). Language and the evolution of modern humans. In K. Omoto & P. Tobias (Eds.), The origins and past of modern humans (pp 267-282). Singapore: World Scientific.

Weber, F., Manganaro, L., Peskin, B., & Shriberg, E. (2002) Using prosodic and lexical information for speaker identification. Proceedings of the IEEE International Conference on Audio, Speech and Signal Processing.

Williams, C. E., & Stevens, K. N. (1972). Emotions and speech: Some acoustical factors. Journal of the Acoustical Society of America, 52, 1238-1250.