Published in final edited form as:

J Acoust Soc Am. 1985 July ; 78(1 Pt 2): 381–388.

Speech perception: Some new directions in research and theory

David B. Pisoni

Speech Research Laboratory, Indiana University, Bloomington, Indiana 47405

Abstract

The perception of speech is one of the most fascinating attributes of human behavior; both the auditory periphery and higher centers help define the parameters of sound perception. In this paper some of the fundamental perceptual problems facing speech sciences are described. The paper focuses on several of the new directions speech perception research is taking to solve these problems. Recent developments suggest that major breakthroughs in research and theory will soon be possible. The current study of segmentation, invariance, and normalization are described. The paper summarizes some of the new techniques used to understand auditory perception of speech signals and their linguistic significance to the human listener.

INTRODUCTION

This is an exciting time for researchers working in the field of speech processing. Many people feel that we are literally at the point of a major breakthrough in research and theory and that many of the old and long-standing problems in the field may be resolved within the next few years (Klatt, 1983). The solution to some of these problems could well have a number of practical implications, such as the development of improved speech recognition systems, more sophisticated aids for the hearing impaired, and a wide range of consumer products using speech technology.

The increased optimism about speech processing has been brought about, in part, by several different but closely related developments that have taken place over the last few years. First, the engineering and artificial intelligence communities realized that a better solution to the speech recognition problem could only be obtained by increased efforts at improving the front end performance of speech recognizers. Improved segmental recognition is one direct way to increase performance well beyond the levels obtainable with the currently available commercial technology.

Second, many investigators realized that our understanding of how human listeners perceive fluent continuous speech was quite impoverished compared to the voluminous research literature on the perception of phonemes in isolated nonsense syllables. Moreover, it has become increasingly clear from the research carried out under the DARPA (Defense Advanced Research Projects Agency) speech understanding project (Klatt, 1977), that a great deal more basic and detailed knowledge is needed about sentence-level phenomena in speech. The numerous acoustic–phonetic changes that are introduced at word boundaries and the operation of phonological rules in language create sources of variability in speech that are still poorly understood even today. These two areas of research on connected speech are assumed to be of great importance in eventually solving the recognition problem.

Finally, new research strategies and techniques have been developed to study the different sources of acoustic–phonetic variability in speech. Some of these efforts involve the use of large databases of natural speech compiled from multiple talkers, which can be used to provide quantitative estimates of the occurrence of various phenomena in speech. These techniques are currently being used to develop new and vastly improved algorithms and

decision strategies for machine recognition of speech (Bush *et al.*, 1983; Chen and Zue, 1983; Huttenlocher and Zue, 1983; Cole and Brennan, 1983). Other novel research efforts have involved more detailed analyses of speech spectrograms (Kuhn, 1983) and new techniques for representing the neural coding of speech as it is processed by the peripheral auditory system (Carlson and Granstrom, 1982; Goldhor, 1983 a,b; Seneff, 1983).

I. BASIC ISSUES IN SPEECH PERCEPTION

The field of speech perception is a very broad interdisciplinary area involving researchers from a number of disciplines including experimental psychology, linguistics, speech and hearing science, electrical engineering, and artificial intelligence. Despite differences in approach and overall goals, there is fairly good agreement among researchers about what the basic issues are in the field today. In this section, the seemingly major theoretical issues in the field will be reviewed briefly first.

A. Lack of acoustic–phonetic invariance and the segmentation problem

The lack of acoustic–phonetic invariance in speech is considered by all speech researchers to be *the* most important problem in the field; this was the major problem uncovered in the late 1940s after the invention of the sound spectrograph; it is still the central problem in the field of speech research today. Despite over 30 years of research on the problem, investigators have been unsuccessful in identifying acoustic segments and properties of the speech waveform that uniquely match the units derived from perceptual analysis. The most obvious case is the failure to find acoustic units that map onto the phonetic segments or phonemes assumed from linguistic analysis of the intended message. A great deal of research has demonstrated that a single acoustic segment of the speech waveform often contains information about several neighboring linguistic segments, and, conversely, that the same linguistic segment is often represented acoustically in the speech waveform in quite different ways depending on the surrounding phonetic context, rate of speaking, talker, and syntactic environment. When compared to words produced in isolation, the acoustic characteristics of individual speech sounds (i.e., phonemes) display even greater variability in connected speech because of the influence of the surrounding phonetic context.

Closely related to the acoustic–phonetic invariance problem is a set of problems associated with segmentation of the speech waveform. The context-conditioned variability of speech and the apparent lack of any direct correspondence between the speech signal and linguistic–perceptual units such as phonemes or words present enormous problems for segmentation of speech into psychologically real and meaningful units that can be used in recognition. Research has demonstrated that it is extremely difficult, if not impossible, to segment speech into acoustically defined units that are independent of adjacent segments and free from the contextual effects that occur in sentence environments. It has been difficult to use strictly physical (i.e., acoustic) criteria to determine where one word ends and another begins in fluent connected speech. Although some segmentation is possible according to acoustic criteria (see Fant, 1962), the number of acoustic segments is typically found to be much greater than the number of phonemes or words in an utterance (see also Cole *et al.*, 1980).

B. Internal representation of speech sounds

Until very recently, there was fairly good agreement among most investigators working on human speech perception that at some stage of perceptual processing the speech signal is represented internally as a sequence of discrete linguistic segments and features (see, for example, Studdert-Kennedy, 1974, 1976; Pisoni, 1981). There was much less agreement, however, about the exact description of these features. Over the years, feature systems have been proposed, based on distinctions in the acoustic domain, the articulatory domain, and a

combination of both. Recently, the trend has shifted to view these traditional feature descriptions with some skepticism, particularly with regard to the precise role these features and units actually play in ongoing speech perception (Klatt, 1977, 1979, 1981). The argument here is that on reexamination, much of the original evidence cited in support of feature- or segment-based processing in perception is ambiguous and equally consistent with more parametric (i.e., acoustic/auditory) representations of the speech signal that do not make any *a priori* assumptions about intermediate or more abstract levels of analysis. Based on these arguments against segmental representations, a number of researchers have begun to look more closely at how speech signals are processed by the peripheral auditory system and what these much more detailed representations may contribute to resolving the issues surrounding the acoustic–phonetic invariance problem.

C. Units in speech perception

A long-standing and closely related issue in speech perception has been the choice of a minimal unit of perceptual analysis. Because of channel capacity limitations in the auditory system, the rich sensory-based neural information output from the peripheral auditory system must be recoded into some abstract and more permanent form that can be used in perception and subsequent decision making. Many researchers have asked whether there is one basic or “natural” coding unit for speech. Over the years, investigators have argued for the primacy of the feature, phoneme, syllable, or word as their candidate for the basic perceptual unit in speech. Other researchers, motivated chiefly by psycho-linguistic processing considerations, have argued for much larger perceptual units such as clauses and sentences (Bever *et al.*, 1969; Miller, 1962). The continuing debate over the choice of a perceptual unit could be resolved, in our view, if a strict distinction were drawn over the level of linguistic analysis under consideration. In understanding spoken language, the size of the processing unit apparently varies from feature to segment to clause as the listener’s attention is directed to different aspects of the incoming linguistic message. When considered in this way, the arguments and continuing debate over the question of whether there is one and only *one* basic or primary unit of perception actually become irrelevant. There are, in fact, many different units used in speech perception and these units become more or less “primary” as a function of the level of processing required by the specific task presented to the listener.

D. Normalization problems in speech

In addition to the problems associated with the lack of acoustic–phonetic invariance and segmentation, another problem arises in connection with perceptual normalization of the speech signal. One aspect of this problem derives from physical and articulatory differences among talkers; specifically, the observation that the length and shape of the vocal tract differ quite substantially among different talkers. Moreover, the articulatory gestures used to produce individual phonemes and the strategies used to realize these in different phonetic environments also differ quite substantially among talkers. One consequence is that substantial acoustic–phonetic differences exist among talkers in the physical correlates of most, if not all, of the phonetic distinctions used in spoken language. While human listeners are able to ignore or minimize talker differences easily through perceptual compensation and talker normalization, currently available speech recognizers have found the same problem much more difficult to overcome.

Another related aspect of this problem concerns time and rate normalization. Research has shown that the durations of individual speech sounds are influenced quite substantially by an individual talker’s speaking rate. Moreover, these durations are also affected by the locations of various syntactic boundaries in connected speech, by syllabic stress, and by the component features of adjacent phonetic segments in words (see Gaitenby, 1965; Klatt,

1975, 1976, 1979; Lehiste, 1970). In addition to these sources of variability, substantial differences have also been observed in the durations of segments in words when they are produced in sentence contexts compared to the same words spoken in isolation. Although human listeners display evidence of perceptual constancy in the face of enormous physical variation, the precise basis for their perceptual compensation is still unknown and remains a topic of intense study (see for example Miller and Liberman, 1979; Miller, 1981; Pisoni *et al.*, 1983).

E. Interaction of knowledge sources in speech

Above and beyond the basic problems inherent in acoustic–phonetic recognition, spoken language understanding involves access to and use of a variety of sources of knowledge that a listener has available as a speaker and hearer of a natural language. In addition to the physical information encoded in the speech waveform, the listener has a great deal of detailed knowledge of the structure of the language, which is actively used with the sensory input to develop a representation of the incoming message. A central issue in the field of speech perception concerns the precise representation of these knowledge sources and the extent to which these different sources of information interact at different stages of perceptual analysis of speech (Reddy, 1976). One of the major points of contention among current theories of auditory word recognition concerns whether words are recognized by an autonomous processing mechanism that is independent of other knowledge sources (e.g., Forster, 1976; Norris, 1982) or whether words are recognized through an interaction of top-down linguistic knowledge and bottom-up sensory information [e.g., Marslen-Wilson *et al.*, 1978; Cole and Jakimik, 1980).

II. AUDITORY MODELING OF SPEECH SIGNALS IN THE PERIPHERY

Over the last three or four years a great deal of new research has been reported on how the peripheral auditory system processes and encodes speech signals (see Carlson and Granstrom, 1982). The research on auditory modeling of speech signals comes from two different directions. On one hand, a number of important physiological studies using animals have been carried out to describe, in fairly precise terms, how simple speech signals are coded in the peripheral auditory system (Kiang, 1980; Delgutte, 1980, 1981, 1982). These studies have examined auditory-nerve activity in response to simple speech signals such as steady-state vowels and stop consonants in CV syllables (see also Moore and Cashin, 1974, 1976). The goal of this work has been to identify reliable and salient properties in the discharge patterns of auditory-nerve fibers that correspond, in some direct way, to the important acoustic properties or attributes of speech sounds (Sachs and Young, 1979, 1980; Young and Sachs, 1979; Miller and Sachs, 1983).

On the other hand, several researchers have begun to develop psychophysically based models of speech processing that explicitly incorporate well-known psychoacoustic data in their descriptions of the filtering that is carried out by the peripheral auditory system (Searle *et al.*, 1979; Zwicker *et al.*, 1979; Kewley-Port, 1980, 1983). The goal of this line of research is to develop representations of the speech signal that take into account known psychophysical facts about hearing such as critical bands, upward spread of masking, and the growth of loudness (Klatt, 1982; Zwicker *et al.*, 1979).

The recent interest and extensive research efforts in developing new and presumably more appropriate and valid representations of speech signals such as “neurograms,” “cochleagrams,” or “neural spectrograms” derives, in part, from the assumption that a more detailed examination of these auditory representations should, in principle, provide researchers with a great deal more relevant information about the distinctive *perceptual* dimensions that underlie speech sounds (Mundie and Moore, 1970; Stevens, 1980). It has

been further assumed that information contained in these so-called neuroacoustic representations will contribute in important ways to finally resolving the acoustic–phonetic invariance problem in speech (see, for example, Goldhor, 1983a,b).

One recent approach has been developed along these lines in our laboratory in terms of representing speech signals with three-dimensional running spectral displays (see Kewley-Port, 1980, 1983; Kewley-Port and Luce, 1984). The interest in using this particular type of display to study the acoustic properties of stop consonants is based, in part, on the assumption that the perceptual dimensions for speech sounds should be represented in terms of what is currently known about transformations of acoustic signals by the peripheral auditory system. Sufficient physiological and psychophysical data are currently available in the literature about the operation of the auditory system to suggest that the peripheral auditory system can be modeled, to a first approximation, as a frequency analyzer with continuously varying neural signals as output (see Klatt, 1979; Plomp, 1964; Schroeder *et al.*, 1979; Zwicker *et al.*, 1979; Siebert, 1968; Searle *et al.*, 1979; Carlson and Granstrom, 1982). In this work, special emphasis has been placed on a detailed examination of precisely how the distribution of spectral energy changes rapidly over time for stop consonants, particularly for the distinctive acoustic correlates of place of articulation in stops (see Kewley-Port, 1983).

In addition to looking at how the distribution of spectral energy changes over time in visual displays of speech, work has also been directed at several different ways of modeling the energy spectra that are derived from the filtering carried out by the auditory system. The spectra typically employed in speech analysis studies are often obtained from linear prediction (LPC) techniques which are used to compute formant frequencies (Atal and Hanauer, 1971; Makhoul, 1975; Markel and Gray, 1976). Because this type of analysis assumes constant bandwidths, the derived spectra are not really appropriate to model how the human auditory system filters these signals. Similarly, in analyses of speech, frequency is often displayed on a linear rather than log-frequency scale despite what is known about how humans encode frequency in the auditory system. Some idea of the differences between these two types of representations can be seen by examining the three-dimensional running spectral displays shown in Fig. 1.

A linear prediction (LPC) running spectrum for the syllable /bu/ is shown with a linear frequency scale in the top panel of Fig. 1; a second running spectrum derived from a set of auditory filters using the Mel scale is shown in the bottom panel of Fig. 1. Examination of the two displays reveals a number of important differences between the two representations. It is well known that the bandwidth of the frequency analysis performed by the auditory system increases with frequency. Various estimates of internal auditory bandwidths have been made in the past from psychophysical data; they range from one-half to one-tenth octave (see Moore and Glasberg, 1983). The ranges most often reported in the literature are shown in the hatched areas in Fig. 2. The bark scale, which is based on the concept of critical bands, is shown as an upper limit (Zwicker *et al.*, 1979), while the one-sixth octave filtering which is similar to Patterson's (1976) proposal is shown as the lower limit. The bandwidths assumed from the linear prediction filtering are shown in the bottom of Fig. 2 by the straight horizontal line. It should be obvious that compared to the other approaches, the LPC analysis which has been used quite extensively in speech processing work may not provide the most appropriate way of characterizing the filtering that takes place in the auditory system.

The long range goal of this work is to develop better auditory representations of the initial sensory encoding of speech at the periphery so that new recognition algorithms can be developed to improve front end performance. While there is good reason to be encouraged

by these efforts, it is also important to continue detailed studies of the central auditory mechanisms that integrate the initial sensory information. And, further research is also needed on the subsequent perceptual–cognitive decision processes that interpret the patterns of sensory stimulation in linguistically relevant ways.

III. SPECTROGRAM READING AND ACOUSTIC–PHONETIC DATABASES

Over the last five years two important methodological developments have dramatically affected the way in which acoustic–phonetic research is carried out. The first development, owed largely to the efforts of Victor Zue at MIT, was the demonstration that speech spectrograms could be analyzed and interpreted at very high levels of accuracy. Through careful study of an “expert spectrogram reader” explicit rules were formulated to describe the strategies used to interpret spectrograms. In a detailed examination of Zue’s abilities, Cole *et al.* (1980) reported that phonetic segments could be reliably identified from acoustic–phonetic information displayed in spectrograms of unknown utterances. Zue’s performance in analyzing spectrograms was shown to be comparable to a group of phoneticians who analyzed the same unknown utterances by careful listening and standard transcription techniques.

While the initial report of Zue’s performance in analyzing spectrograms demonstrates that speech can be recognized from the visual information contained in spectrograms and that reliable cues to phonetic segments can be found in the speech signal, the problems of recognizing acoustic segments in continuous fluent speech still remain to be solved. Nevertheless, these results have already had several important consequences for new research strategies in the field of speech recognition. First, these findings have laid to rest a long-standing and apparently erroneous belief that speech spectrograms, particularly spectrograms of unknown and unfamiliar utterances, simply could not be read or analyzed (Lieberman *et al.*, 1967). As a result, numerous researchers have started to examine spectrograms in very great detail as a rich and potentially valuable source of information about the phonetic content of the linguistic message. What was previously just a simple demonstration exercise in phonetics classes has now become the topic of intense formal study in specialized classes devoted exclusively to learning to interpret spectrographic displays of speech through the application of explicit rules and principles.

A related and perhaps a more far reaching consequence of these efforts has been the development of new feature-based descriptions of the acoustic correlates of speech sounds derived directly from the information contained in spectrographic displays (Cole *et al.*, 1983). The logic behind this approach is that rather than being thought of as impoverished or noisy, the speech signal may be considered as an extremely rich source of reliable information (Elman and McClelland, 1983). This information can be used to identify the acoustic correlates of speech sounds without having to resort to higher sources of knowledge such as morphology, syntax, or semantics or elaborate hypothesis generation and prediction about constraints on vocal tract dynamics. In other words, a great deal of reliable information is present in the speech signal and, as the argument goes, this information has not been used effectively in past efforts at designing front ends of speech recognition systems. Previous work was largely unsuccessful because of the failure to capture the systematic regularities that exist between the phonemes or perceptual units of the linguistic message and their acoustic correlates as represented in spectrograms. Indeed, even the most successful speech understanding systems like Harpy were capable of correctly recognizing only about 45% of the phonemes in the speech signal (Lowerre and Reddy, 1979). Their successful overall performance came about primarily through the creative use of linguistic constraints and powerful search strategies applied to a precompiled knowledge-based network.

The second recent development concerns the construction and systematic use of large databases for research in acoustic–phonetics. The sources of variability in speech and the influence of different contextual factors have been found to be so enormous in scope that it is often difficult, if not literally impossible, to abstract regularities and general principles from a small number of measurements made by hand across a large number of different phonetic environments. This is especially critical when the hand analyses take a great deal of time. In the last few years, several efforts have been made to develop large databases for use in acoustic–phonetic research so that precise quantitative measurements can be made on the effects of various phonetic environments and sources of variability (Shipman, 1983). Using techniques from artificial intelligence, very sophisticated database management procedures have been developed to permit a researcher to search and interrogate the database for specific types of information that would otherwise be difficult or impossible to obtain with the conventional techniques used in acoustic–phonetic research (Shipman, 1982).

Because of the availability of these large databases many research questions that seem trivial today would probably never have been conceived only a few years ago. The ability to generate and test experimental hypotheses at a much faster rate allows speech researchers to make greater advances in accumulating new knowledge about the acoustic–phonetics of a particular language and the effects of different sources of variability. These efforts have been particularly useful in detailing the operation of phonological rules and the effects of local phonetic contexts on the acoustic properties of speech sounds across different environments (Zue, 1976; Lamel, 1983).

Other work using large databases has shown that detailed knowledge about the structural properties of words in the lexicon and their relations in terms of “lexical density” can be used to constrain the potential search space used in word recognition (Eukel, 1980; Landauer and Streeter, 1973; Crystal and House, 1982; Shipman and Zue, 1982). Although the use of specialized databases in acoustic–phonetic research is still at an early stage of development, it seems obvious that use of this detailed information will vastly improve the front end performance of the next generation of speech recognizers (Chen and Zue, 1983). More important, however, these recent developments demonstrate that the sources of variability in speech are finite, and that, at least in principle, the acoustic–phonetic invariance problem may finally be solved in the immediate future.

IV. WORD RECOGNITION AND LEXICAL ACCESS

The study of word recognition and the nature of the lexical representation for words have been long-standing concerns of experimental psychologists (see Bagley, 1900–1901), although these particular topics have not been studied routinely by investigators working in the mainstream of speech research (Cole and Rudnick, 1983). Several reasons for this lack of attention to words can be identified (see Nooteboom, 1981). First, much of what we know about word recognition has been obtained in the visual modality through studies of the reading process. Second, because of the concern for phoneme and feature perception in isolated nonsense syllable contexts, few speech researchers have been willing or interested in venturing out of their own narrow research areas to study word recognition. Moreover, few, if any, efforts have been made by speech researchers to generalize their findings on phoneme and feature perception to the study of the cues used to recognize words in isolation or in connected speech (for example, see reviews by Studdert-Kennedy, 1974, 1976; Darwin, 1976).

Several important problems in speech perception involve the processes of word recognition and lexical access. Moreover, these problems bear directly on issues surrounding the nature of the various types of representations in the mental lexicon and the interaction of

knowledge sources in speech perception. As we noted earlier in connection with the problems surrounding units of perceptual analysis, there has been considerable interest in determining precisely what kinds of representations exist in the mental lexicon. Are words, morphemes, phonemes, or sequences of spectral templates appropriate representations for lexical entries? Are words accessed in terms of their acoustic–phonetic or phonological structure? Why are high-frequency words treated differently than low-frequency words? These are just a few of the questions that investigators are trying to answer at this time (see Pisoni *et al.*, 1985).

One of the central and long-standing problems in speech perception concerns the interaction of the sensory information output from the auditory system with higher-level contextual information. A number of researchers such as Forster (1976) and Massaro and Oden (1980) have maintained that early sensory information is processed independently of context and higher-level knowledge and that the facilitation effects observed for word recognition in context are due to post-perceptual processes associated with readjusting decision criteria. Other researchers such as Marslen-Wilson and Tyler (1980) have argued strongly that context does influence the early sensory analysis of the input signal and that the various knowledge sources used in spoken language comprehension interact freely and quite substantially in supporting recognition (see, also, Reddy, 1976).

V. PERCEPTION OF FLUENT CONNECTED SPEECH

Most of the research in speech perception carried out over the 30 years has been directed almost exclusively at trying to understand the perception of isolated speech sounds. The bulk of this work has examined phoneme perception in isolated contexts using simple nonsense syllables as stimulus materials. Although this approach is quite narrow in scope, the research strategy can be understood when one begins to consider the enormous complexity of spoken language perception and understanding, particularly the perception of fluent connected speech. Relative to the voluminous literature in speech perception on isolated phoneme perception, very little is actually known at this time about how the early sensory-based acoustic–phonetic information output by the peripheral auditory system is used by the higher and more central speech processing mechanisms in tasks that involve word recognition or language comprehension. Very little is known about how changes in the segmental and/or suprasegmental structure of the speech signal affect intelligibility and subsequent comprehension of the linguistic message. Research on the perception of fluent speech obviously represents a sharp departure from the “mainstream” of traditional work in the field of speech perception. Understanding the perception of connected speech is an extremely important area of investigation in speech research because spoken language comprehension is ultimately dependent on the initial sensory and perceptual analysis of the acoustic–phonetic input to the processing system.

Human speech perception and spoken language comprehension appear to take place very rapidly in close-to-real time. A good deal of the perceptual processes and computational mechanisms that support such on-line activities operate automatically and therefore are unavailable to consciousness. Moreover, even under very adverse listening conditions, human observers are able to extract the linguistic message from the speech signal despite the fact that it may be distorted or parts of it entirely obliterated. The questions surrounding the perception of fluent connected speech differ substantially from the issues related to phoneme and feature perception because they inevitably involve the listener’s cognitive system and a consideration of how various sources of linguistic knowledge interact to support perception and comprehension. For example, what is the size of the scanning window over which low-level phonetic decisions are made? What depth of processing is required before a decision can be made about the segmental identity of the input signal? Do all decisions of the speech

processing system take place immediately in real time or are there processing delays at particular analytic levels of the system which wait for further information about the input signal and content of the message? To what extent is the processing “automatic?” How much of the speech perception process is controlled by higher-level linguistic processes? These are just a few of the questions that researchers are working on at this time (see Marslen-Wilson and Tyler, 1980).

Research on speech communication over the last 35 years has shown that the linguistic message contains a great deal of redundant information. One of the goals of communication engineers was to identify the most important information in the speech signal and develop methods to encode this information in efficient ways for transmission. In a similar way, speech researchers have tried to identify the “minimal cues” for phoneme perception in the hope that once these cues could be identified the problems of speech recognition could be solved. Unfortunately, there is a great deal more to speech perception and spoken language understanding than simply discovering the minimal cues to phoneme perception. In studying the perception of fluent continuous speech, the problem becomes one of discovering the “islands of reliability” in the speech signal that listeners use to access various sources of knowledge. Examples of such information include the presence of stressed syllables, the beginnings and endings of words, and the locations of various spectral changes indicating discrete shifts in the vocal tract source function (Bond, 1978; Jakimik and Hunnicutt, 1981; Nooteboom, 1981). Human listeners apparently do not pay equal attention to all aspects of the speech signal; instead they use only those properties or attributes that are informative at any point in time to reach a decision (Marslen-Wilson and Tyler, 1980).

VI. SUMMARY AND CONCLUSIONS

In this brief report we have tried to summarize a number of the major theoretical issues in the field of speech perception and a few of the most promising areas of current research on speech processing. It should be obvious from the present remarks that a great deal of research activity is going on in the field today and that a great deal more is being learned about how humans perceive and understand spoken language. The approaches to studying these problems have become more diverse in the last few years as researchers in several related disciplines apply their knowledge and expertise to the central problems in the field. Moreover, a great deal more communication and interaction are taking place among speech researchers and hearing scientists, particularly on issues related to auditory modeling of the early stages of speech perception.

Although new and important findings will come from continued research on how speech is processed in the auditory periphery, no one should be misled into believing that these new research efforts on processing speech by the auditory nerve will provide all the needed solutions in the field of speech processing (see also Klatt, 1982). On the contrary, a great deal more research is currently needed on central auditory mechanisms involved in pattern recognition and on word recognition and the perception of fluent connected speech. More basic research is needed on spectrogram reading; further efforts should be devoted to the development of large databases that can be used to evaluate new hypotheses about the different sources of variability in speech.

It is clear, however, that an important change in direction and attitude has taken place among many speech researchers. Investigators appear to be directing their research efforts and attention toward much broader theoretical issues than just a few years ago, issues that encompass the study of more meaningful linguistic stimuli in somewhat more naturalistic contexts, which require listeners to use several sources of knowledge to assign a linguistic interpretation to the sensory input. An important shift can be seen in the direction of

research efforts towards a much greater concern for the differential contribution of the surrounding context to the acoustic–phonetic realization of the speech signal. If nothing else, researchers appear to be very optimistic that a final solution to the invariance problem is actually possible. It appears to be only a matter of time and more basic research into the complexity of the speech code before this goal will be realized.

Acknowledgments

Preparation of this paper was supported in part by a NIH grant NS-12179 and in part by contract F33615-83K-501 from the Air Force Office of Scientific Research (AFSC) to Indiana University in Bloomington. I thank Howard Nusbaum and Beth Greene for their help and advice with the manuscript. Special thanks also go to William Yost and Charles Watson for inviting me to present these views at the CHABA meetings in October 1983.

References

- Atal BS, Hanauer SL. Speech analysis and synthesis by linear prediction of the speech wave. *J Acoust Soc Am*. 1971; 50:637–655. [PubMed: 4106390]
- Bagley WC. The apperception of the spoken sentence: A study in the psychology of language. *Am J Psychol*. 1900–1901; 12:80–130.
- Bever TG, Lackner J, Kirk R. The underlying structure sentence is the primary unit of immediate speech processing. *Percept Psychophys*. 1969; 5:225–234.
- Bond, ZS. Listening to elliptical speech: Pay attention to stressed vowels. paper presented at the annual meeting of the Linguistic Society of America; Boston. December; 1978.
- Bush MA, Hamilton M, Hata K. An acoustic analysis of coarticulatory phenomena in connected digit strings. *J Acoust Soc Am Suppl 1*. 1983; 74:S14.
- Carlson, R.; Granstrom, B. Towards an auditory spectrograph. In: Carlson, R.; Granstrom, B., editors. *The Representation of Speech in the Peripheral Auditory System*. Elsevier; New York: 1982. p. 109–114.
- Chen FR, Zue VW. Exploring allophonic and lexical constraints in a continuous recognition system. *J Acoust Soc Am Suppl 1*. 1983; 74:S15.
- Cole, RA., editor. *Perception and Production of Fluent Speech*. Erlbaum; Hillsdale, NJ: 1980.
- Cole RA, Brennan RA. A pretty good formant tracker. *J Acoust Soc Am Suppl 1*. 1983; 74:S15.
- Cole, RA.; Jakimik, J. A model of speech perception. In: Cole, R., editor. *Perception and Production of Fluent Speech*. Erlbaum; Hillsdale, NJ: 1980.
- Cole RA, Rudnicki AI. What's new in speech perception? The research and ideas of William Chandler Bagley, 1874–1946. *Psychol Rev*. 1983; 90:94–101. [PubMed: 6342015]
- Cole, RA.; Rudnicki, AI.; Zue, VW.; Reddy, DR. Speech as patterns on paper. In: Cole, RA., editor. *Perception and Production of Fluent Speech*. Erlbaum; Hillsdale, NJ: 1980.
- Cole, RA.; Stern, RM.; Lasry, MJ. Performing fine phonetic distinctions: Templates vs. features. paper presented at the Symposium on Variability and Invariance of Speech Processes; MIT; 1983.
- Crystal TH, House AS. Segmental durations in connected speech signals: Preliminary results. *J Acoust Soc Am*. 1982; 72:705–716. [PubMed: 7130529]
- Darwin, CJ. The perception of speech. In: Carterette, EC.; Friedman, MP., editors. *Handbook of Perception*. Academic; New York: 1976. p. 175–226.
- Delgutte B. Representation of speechlike sounds in the discharge patterns of auditory-nerve fibers. *J Acoust Soc Am*. 1980; 68:843–857. [PubMed: 7419820]
- Delgutte, B. unpublished doctoral dissertation. MIT; 1981. Representation of speech-like sounds in the discharge patterns of auditory-nerve fibers.
- Delgutte, B. Some correlates of phonetic distinctions at the level of the auditory nerve. In: Carlson, R.; Granstrom, B., editors. *The Representation of Speech in the Peripheral Auditory System*. Elsevier Biomedical; New York: 1982.
- Elman, JL.; McClelland, JL. Exploiting lawful variability in the speech wave. paper presented at the Symposium of Variability and Invariance of Speech Processes; MIT; 1983.

- Eukel B. A phonotactic basis for word frequency effects: Implications for lexical distance metrics. *J Acoust Soc Am Suppl 1*. 1980; 68:S33.
- Fant CGM. Descriptive analysis of the acoustic aspects of speech. *Logos*. 1962; 5:3–17. [PubMed: 13891546]
- Forster, KI. Accessing the mental lexicon. In: Wales, RJ.; Walker, E., editors. *New Approaches to Language Mechanisms*. North Holland; Amsterdam: 1976. p. 257-287.
- Gaitenby JH. The elastic word. *Haskins Lab Stat Rep Speech Res*. 1965; SR-2:3.1–3.12.
- Goldhor R. A speech signal processing system based on a peripheral auditory model. *Proc IEEE, ICASSP-83*. 1983a:1368–1371.
- Goldhor R. The representation of speech signals in a model of the peripheral auditory system. *J Acoust Soc Am Suppl*. 1983b; 173:S4.
- Huttenlocher DP, Zue VW. Exploring phonotactic lexical constraints in word recognition. *J Acoust Soc Am Suppl 1*. 1983; 74:S15.
- Jakimik J, Hunnicutt S. Organizing the lexicon for recognition. *J Acoust Soc Am Suppl 1*. 1981; 69:S41.
- Kewley-Port, D. Res Speech Percept Tech Rep No 3. Speech Research Laboratory, Indiana University; Bloomington, IN: 1980. Representations of spectral change as cues to place of articulation in stop consonants.
- Kewley-Port D. Time-varying features as correlates of place of articulation in stop consonants. *J Acoust Soc Am*. 1983; 73:322–335. [PubMed: 6826902]
- Kewley-Port D, Luce PA. Time-varying features of initial stop consonants in auditory running spectra: A first report. *Percept Psychophys*. 1984; 35:353–360. [PubMed: 6739270]
- Kiang NYS. Processing of speech by the auditory nervous system. *J Acoust Soc Am*. 1980; 68:830–835. [PubMed: 7419818]
- Klatt DH. Vowel lengthening is syntactically determined in a connected discourse. *J Phonet*. 1975; 3:129–140.
- Klatt DH. Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *J Acoust Soc Am*. 1976; 59:1208–1221. [PubMed: 956516]
- Klatt DH. Review of the ARPA speech understanding project. *J Acoust Soc Am*. 1977; 62:1345–1366.
- Klatt DH. Speech perception: A model of acoustic–phonetic analysis and lexical access. *J Phonet*. 1979; 7:279–312.
- Klatt, DH. Lexical representations for speech production and perception. In: Myers, T.; Laver, J.; Anderson, J., editors. *The Cognitive Representation of Speech*. North-Holland; Amsterdam: 1981. p. 11-37.
- Klatt, DH. Speech processing strategies based on auditory models. In: Carlson, R.; Granstrom, B., editors. *The Representation of Speech in the Peripheral Auditory System*. Elsevier; New York: 1982. p. 181-196.
- Klatt, DH. The problem of variability in speech recognition and in models of speech perception. paper presented at the Symposium on Variability and Invariance of Speech Processes; MIT; 1983.
- Kuhn G. Experiments with a color spectrogram. *J Acoust Soc Am Suppl 1*. 1983; 73:S86.
- Lamel LF. The use of phonotactic constraints to determine word boundaries. *J Acoust Soc Am Suppl 1*. 1983; 74:S15.
- Landauer T, Streeter L. Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. *J Verb Learn Verb Behav*. 1973; 12:119–131.
- Lehiste, I. *Suprasegmentals*. MIT; Cambridge, MA: 1970.
- Lieberman AM, Cooper FS, Shankweiler DP, Studdert-Kennedy M. Perception of the speech code. *Psychol Rev*. 1967; 74:431–461. [PubMed: 4170865]
- Lowerre, B.; Reddy, DR. The Harpy speech understanding system. In: Lea, WA., editor. *Trends in Speech Recognition*. Prentice-Hall; Englewood Cliffs, NJ: 1979.
- Makhoul J. Linear prediction: A tutorial review. *Proc, IEEE*. 1975; 63:561–580.
- Markel, JD.; Gray, AH. *Linear Prediction of Speech*. Springer-Verlag; New York: 1976.

- Marslen-Wilson W, Tyler LK. The temporal structure of spoken language understanding. *Cognition*. 1980; 8:1–71. [PubMed: 7363578]
- Marslen-Wilson WD, Welsh A. Processing interactions and lexical access during word recognition in continuous speech. *Cog Psychol*. 1978; 10:29–63.
- Massaro, DW.; Oden, GC. Speech perception: A framework for research and theory. In: Lass, NJ., editor. *Speech and Language: Advances in Basic Research and Practice*. Vol. 3. Academic; New York: 1980.
- Miller GA. Decision units in the perception of speech. *IRE Trans Inf Theory*. 1962; IT-8:81–83.
- Miller, JL. Effects of speaking rate on segmental distinctions. In: Eimas, PD.; Miller, JL., editors. *Perspectives on the Study of Speech*. Erlbaum; Hillsdale, NJ: 1981. p. 39-74.
- Miller JL, Liberman AM. Some effects of later-occurring information on the perception of stop consonants and semi-vowels. *Percept Psychophys*. 1979; 25:457–465. [PubMed: 492910]
- Miller MI, Sachs MB. Representation of stop consonants in the discharge patterns of auditory-nerve fibers. *J Acoust Soc Am*. 1983; 74:502–517. [PubMed: 6619427]
- Moore TJ, Cashin JL. Response patterns of cochlear nucleus neurons to excerpts from sustained vowels. *J Acoust Soc Am*. 1974; 56:1565–1576. [PubMed: 4427027]
- Moore TJ, Cashin JL. Response of cochlear-nucleus neurons to synthetic speech. *J Acoust Soc Am*. 1976; 59:1443–1449. [PubMed: 939877]
- Moore BCJ, Glasberg BR. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *J Acoust Soc Am*. 1983; 74:750–753. [PubMed: 6630731]
- Mundie JR, Moore TJ. Speech analysis using an analog cochlea. *J Acoust Soc Am*. 1970; 48:131.
- Nooteboom, SG. Speech role and segmental perception or the role of words in phoneme identification. In: Myers, T.; Laver, J.; Anderson, J., editors. *The Cognitive Representation of Speech*. North-Holland; Amsterdam: 1981. p. 143-150.
- Norris D. Autonomous processes in comprehension: A reply to Marslen-Wilson and Tyler. *Cognition*. 1982; 11:97–101. [PubMed: 7198955]
- Patterson RD. Auditory filter shapes derived with noise stimuli. *J Acoust Soc Am*. 1976; 59:640–654. [PubMed: 1254791]
- Pisoni DB. In defense of segmental representations in speech processing. *J Acoust Soc Am Suppl 1*. 1981; 69:S32.
- Pisoni DB, Carrell TD, Gans SJ. Perception of the duration of rapid spectrum changes in speech and nonspeech signals. *Percept Psychophys*. 1983; 34:314–322. [PubMed: 6657432]
- Pisani DB, Nusbaum HC, Luce PA, Slowiaczek LM. Speech perception, word recognition and the structure of the lexicon. *Speech Commun*. 1985 in press.
- Plomp R. The ear as a frequency analyzer. *J Acoust Soc Am*. 1964; 36:1628–1636.
- Reddy DR. Speech recognition by machine: A review. *Proc IEEE*. 1976; 64:501–523.
- Sachs MB, Young ED. Encoding of steady-state vowels in the auditory nerve: Representation in terms of discharge rate. *J Acoust Soc Am*. 1979; 66:470–479. [PubMed: 512208]
- Sachs MB, Young ED. Effects of nonlinearities on speech encoding in the auditory nerve. *J Acoust, Soc Am*. 1980; 68:858–875. [PubMed: 7419821]
- Schroeder MR, Atal BS, Hall JL. Optimizing digital speech coders by exploiting masking properties of the human ear. *J Acoust Soc Am*. 1979; 66:1647–1652.
- Searle CL, Jacobson JF, Rayment SG. Stop consonant discrimination based on human audition. *J Acoust Soc Am*. 1979; 65:799–809. [PubMed: 447910]
- Seneff S. Pitch and spectral estimation of speech based on auditory synchrony model. *J Acoust Soc Am Suppl 1*. 1983; 74:S9.
- Shipman DW. Development of speech research software on the MIT lisp machine. *J Acoust Soc Am Suppl 1*. 1982; 71:S103.
- Shipman DW. Spirex: Statistical analysis in the Spire acoustic-phonetic workstation. *Proc IEEE, ICASSP-83*. 1983:1360–1363.
- Shipman DW, Zue VW. Properties of large lexicons: Implications for advanced isolated word recognition systems. *Proc, IEEE, ICASSP-82*. 1982:199–202.

- Siebert, WM. Stimulus transformations in the peripheral auditory system. In: Kolers, PA.; Eden, M., editors. *Recognizing Patterns*. MIT; Cambridge, MA: 1968.
- Stevens KN. Acoustic correlates of some phonetic categories. *J Acoust Soc Am*. 1980; 68:836–842. [PubMed: 7419819]
- Studdert-Kennedy, M. The perception of speech. In: Sebeok, TA., editor. *Current Trends in Linguistics*. Vol. XII. Mouton; The Hague: 1974.
- Studdert-Kennedy, M. Speech perception. In: Lass, NJ., editor. *Contemporary Issues in Experimental Phonetics*. Academic; New York: 1976. p. 243-293.
- Young ED, Sachs MB. Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers. *J Acoust Soc Am*. 1979; 66:1381–1403. [PubMed: 500976]
- Zue, VW. Tech Rep No 523. Lincoln Laboratory, MIT; 1976. Acoustic characteristics of stop consonants: A controlled study.
- Zwicker E, Terhardt E, Paulus E. Automatic speech recognition using psychoacoustic models. *J Acoust Soc Am*. 1979; 65:487–498. [PubMed: 489818]

FIG. 1. Running spectral displays for the syllable /bu/. The top display shows results of the linear predictive analysis with constant bandwidths and the bottom display shows results with auditory filtering using bandwidths that increase with frequency (from Kewley-Port and Luce, 1984).

FIG. 2. Four different filter bandwidths as a function of frequency using the technical Mel scale (from Kewley-Port and Luce, 1984).