

RESEARCH ON SPOKEN LANGUAGE PROCESSING

Progress Report No. 27 (2005)

Indiana University

**Some Observations on Representations and Representational Specificity in
Speech Perception and Spoken Word Recognition ¹**

David B. Pisoni and Susannah V. Levi

*Speech Research Laboratory
Department of Psychological and Brain Sciences
Indiana University
Bloomington, Indiana 47405*

¹ Preparation of this chapter was supported by grants from the National Institutes of Health to Indiana University (NIH-NIDCD T32 Training Grant DC-00012 and NIH-NIDCD Research Grant R01 DC-00111). We wish to thank Cynthia Clopper, Daniel Dinnsen, Robert Goldstone, Vsevolod Kapatsinski, Conor McLennan, Robert Port, and Steve Winters for their useful discussion and comments on this chapter.

Some Observations on Representations and Representational Specificity in Speech Perception and Spoken Word Recognition

Abstract. The conventional view of speech perception and spoken word recognition relies on discrete, abstract symbolic units. This conventional view faces several problems which led to the development of new approaches to representing speech. In particular, recent exemplar and episodic models can account for both the robustness of speech perception and the effects of indexical information on speech processing. Instead of only representing speech with conventional abstract symbolic representations, the evidence reviewed here suggests that highly detailed information is encoded and stored in memory as well.

Introduction

The field of speech perception and spoken word recognition has undergone rapid change over the last few years as researchers have begun to realize that many of the properties of speech that are responsible for its perceptual robustness, such as speed, fluency, automaticity, perceptual learning and adaptation, and errorful recovery, reflect general properties shared by other self-organizing systems in physics, biology, and neuroscience (Grossberg, 2003; McNellis & Blumstein, 2001; Sporns, 2003). Theoretical developments in cognitive science and brain modeling, as well as new computational tools, have led to a reconceptualization of the major theoretical problems in speech perception and spoken word recognition. Several new exemplar-based approaches to the study of speech perception and spoken word recognition have also emerged from independent developments in categorization (Kruschke, 1992; Nosofsky, 1986) and frequency-based phonology (Pierrehumbert, 2001; Bybee, 2001). These alternatives offer fresh ideas and new insights into old problems and issues related to variability and invariance (Goldinger, 1998; Goldinger & Azuma, 2003; Johnson, 1997). Although many of the basic research questions in speech perception remain the same, the answers to these questions have changed in fundamental ways reflecting new theoretical and methodological developments (Pardo & Remez, in press). These questions deal with the nature of phonological and lexical knowledge and representation, processing of stimulus variability, perceptual learning and adaptation and individual differences in linguistic performance (see Pisoni & Remez, 2005).

When compared to research in other areas of cognitive and neural science, speech perception is unique because of the close coupling and synchrony between speech production and perception. Speech exists simultaneously in several different domains: the acoustic and optical, the articulatory-motor and the perceptual. While the relations among these domains are complex, they are not arbitrary. The sound patterns used in a particular language function within a common linguistic system of contrast that is used in both production and perception. Thus, the phonetic contrasts generated in speech production by the vocal tract are precisely the same acoustic differences that serve a distinctive function in perceptual analysis by the listener (Stevens, 1972). As a result, any theoretical account of speech perception must also consider aspects of speech production and acoustics as well as optics. The articulatory spaces mapped out in speech production are closely coupled with the perceptual spaces used in speech perception and spoken word recognition (Fowler & Balantucci, 2005).

The fundamental problem in speech perception and spoken language processing is to describe how the listener recovers the talker's intended message. This complex problem has been typically broken down into several more specific subquestions: What stages of perceptual analysis intervene between the presentation of the speech signal and recognition of the intended message? What types of processing operations occur at each stage? What are the primary perceptual processing units and what is the nature

and content of the neural representations of speech in memory? Finally, what specific perceptual, cognitive, and linguistic mechanisms are used in speech perception and spoken language processing?

In this chapter, we provide an overview of some recent developments that have been underway in the field as they bear directly on issues surrounding representation and representational specificity in speech perception and spoken word recognition. Because of space limitations, our presentation is selective and is not meant to be an exhaustive survey of the field (see Pisoni & Remez, 2005). It is important to emphasize here, however, our strong belief that the changes that have occurred recently will very likely have profound and long-lasting effects on research, theory and clinical application in the years to come. Put in a slightly different way, there is a revolution going on in the field and it is important to understand the reasons for these changes in thinking and the consequences for the future.

Conventional View of Speech

Background

Different disciplines approach the study of speech perception and spoken language processing in fundamentally different ways reflecting their diverse interests, goals and theoretical assumptions. Linguists have one set of goals in mind while psycholinguists, cognitive scientists and neuroscientists have another set of goals. Historically, generative linguists adopted a formalist view and focused their research on two related problems: describing the linguistic knowledge that native speakers have about their language (their so-called linguistic competence) and explaining the systematic regularities and patterns that natural languages display. To accomplish these goals, linguists made several foundational assumptions about speech which embrace a strong abstractionist, symbol-processing approach. The linguistic approach to speech assumes that speech is structured in systematic ways and that the linguistically significant information in the speech signal can be represented efficiently and economically as a linear sequence of abstract, idealized, discrete symbols using an alphabet of conventional phonetic symbols. Linguists also assumed that the regularities and patterns observed within and between languages could be described adequately by sets of formal rules that operate on these abstract symbols. The segmental representations of speech that linguists constructed were assumed to be idealized and redundancy-free because they were designed to code only the linguistically significant differences in meaning between minimal pairs of words in the language (Twaddell 1952). These representations excluded other redundant or accidental information that may be present in the speech signal, but which is not linguistically contrastive. Two examples of this conventional view are given below.

“. . . there is so much evidence that speech is basically a sequence of discrete elements that it seems reasonable to limit consideration to mechanisms that break the stream of speech down into elements and identify each element as a member, or as probably a member, of one or another of a finite number of sets.” (Licklider, 1952, p. 590)

“The basic problem of interest to the linguist might be formulated as follows: What are the rules that would make it possible to go from the continuous acoustic signal that impinges on the ear to the symbolization of the utterance in terms of discrete units, e.g., phonemes or the letters of our alphabet? There can be no doubt that speech is a sequence of discrete entities, since in writing we perform the kind of symbolization just mentioned, while in reading aloud we execute the inverse of this operation; that is, we go from a discrete symbolization to a continuous acoustic signal.” (Halle, 1956, p. 510)

The conventional segmental view of speech as a linear sequence of abstract, idealized, discrete symbols has been the primary method used for coding and representing the linguistic structure of spoken words in language. This approach to speech has been adopted across a wide range of related scientific disciplines that study speech processing such as speech and hearing sciences, psycholinguistics, cognitive and neural sciences and engineering (Peterson, 1952). The theoretical motivation for this approach goes back many years to the early Panini grammarians and it has become an inextricable part of all linguistic theories. Words have an internal structure and they differ from each other in systematic ways reflecting the phonological contrasts and morphology of a particular language. Although not often made explicit, several important basic theoretical assumptions are made in this particular view of speech that are worth mentioning because they bear directly on several broader theoretical issues related to the nature and content of lexical representations.

First, the conventional linguistic approach to the representation of speech assumes that a set of discrete and linear symbols can be used to represent what is essentially continuous, parametric and gradient information in the speech signal (Pierrehumbert & Pierrehumbert, 1990). Second, it is universally assumed by almost all linguists that the symbols representing phonetic segments or phonemes in speech are abstract, static, invariant, and context-free having combinatory properties like the individual letters used in alphabetic writing systems. Although speech can be considered as a good example of the "particulate principle of self-diversifying systems," (Ablar, 1989) a property of natural systems like genetics and chemical interaction that make "infinite use out of finite media," ambiguity and some degree of uncertainty still remain in the minds of some linguists and speech scientists about precisely what the elemental primitives of speech actually are even after many years of basic and applied research on speech. Are the basic building blocks of speech acoustic segments or features that emerge from speech perception or are they the underlying sensory-motor articulatory gestures used in speech production or are they both or something else?

Finally, the conventional view of speech relies heavily on some set of psychological processes that function to "normalize" acoustically different speech signals and to make them functionally equivalent in perception (Joos, 1948). It is generally assumed by both linguists and speech scientists that perceptual normalization is needed in speech perception in order to reduce acoustic-phonetic variability in the speech signal making physically different signals perceptually equivalent by bringing them into conformity with some common standard or referent (see Pisoni, 1997).

Problems with the Conventional View of Speech Perception

The fundamental problems in speech perception today are the same set of basic problems that have eluded definitive solution for more than four and a half decades (Fant, 1973; Stevens, 1998). Although the intractability of these long-standing problems has led to a voluminous body of literature on the production and perception of speech, researchers are still hard-pressed to describe and explain precisely how listeners perceive speech. Indeed, not only are speech scientists still unsure about the exact nature of the linguistic units arrived at in perceptual processing of speech, but little attention has been directed towards how perceptual analysis of the speech waveform makes contact with representations of words in the lexicon or how these representations are used to support spoken language understanding and comprehension. The acoustic consequences of coarticulation and other sources of contextually conditioned variability result in the failure of the acoustic signal to meet two formal conditions, linearity and invariance, which in turn give rise to a third related problem, the absence of segmentation into discrete units (first discussed by Chomsky & Miller, 1963).

Linearity of the Speech Signal. One fundamental problem facing the conventional view is linearity. The linearity condition states that for each phoneme in the message there must be a corresponding stretch of sound in the utterance (Chomsky & Miller, 1963). Furthermore, if phoneme X is followed by phoneme Y in the phonemic representation, the stretch of sound corresponding to phoneme X must precede the stretch of sound corresponding to phoneme Y in the physical signal. The linearity condition is clearly not met in the acoustic signal. Because of coarticulation and other contextual effects, acoustic features for adjacent phonemes are often “smeared” across individual phonemes in the speech waveform and a clear acoustic division between “adjacent” phonemes is rarely observed (Liberman, Delattres, & Cooper, 1952). Although segmentation is possible according to strictly acoustic criteria (see Fant, 1962), the number of acoustic “segments” is typically greater than the number of phonemes in the utterance. This smearing, or “parallel transmission” of acoustic features, results in stretches of the speech waveform in which acoustic features of more than one phoneme are present (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). For this reason, Liberman and his colleagues at Haskins Laboratories have argued that speech is not a simple cipher or alphabet, but is, instead, a complex code in which “speech sounds represent a very considerable restructuring of the phonemic ‘message’” (p.4).

Acoustic-Phonetic Invariance. Another condition that the speech signal fails to satisfy is the principle of invariance (Chomsky & Miller, 1963). This condition states that for each phoneme X, there must be a specific set of critical acoustic attributes or features associated with X in all contexts. These “defining features” must be present whenever X or some variant of X occurs and they must be absent whenever some other phoneme occurs in the representation (Estes, 1994; Smith & Medin, 1981; Murphy, 2002). Because of coarticulatory effects in speech production, the acoustic features of a particular speech sound routinely vary as a function of the phonetic environment in which it is produced. For example, the formant transitions for syllable-initial stop consonants which provide cues to place of articulation (e.g., /b/ vs. /d/ vs. /g/) vary considerably depending on properties of the following vowel (Liberman, Delattre, Cooper & Gerstman, 1954). These transitions do not uniquely specify place of articulation across all vowels. If formant transitions are the primary cues to the perception of place of articulation for stop consonants, they are highly context-dependent. In short, the problem of acoustic-phonetic invariance is one of explaining how perceptual constancy for speech sounds is achieved and maintained when reliable acoustic correlates for individual phonemes in the speech waveform are absent (Blumstein & Stevens, 1981; Studdert-Kennedy, 1974).

Not only is invariance rarely observed for a specific segment across different phonetic environments within a talker, it is also absent for a particular segment in a particular context across speakers. For example, men, women, and children with different vocal tract lengths exhibit large differences in their absolute formant values (Peterson & Barney, 1952).

Speech Segmentation. The combination of non-linearity of the speech signal and context-conditioned variability leads to a third problem in speech perception, namely, the segmentation of the speech waveform into higher-order units of linguistic analysis such as syllables and words. Because of the lack of linearity and acoustic-phonetic invariance, the speech signal cannot be reliably segmented into acoustically defined units that are independent of adjacent segments and free from the conditioned effects of sentence-level contexts. For example, in fluent speech it is difficult to identify by means of simple acoustic criteria where one word ends and another begins.

Assumptions about segmentation and word recognition are probably not independent from assumptions made about the structure and organization of words in the lexicon (see Bradley & Forster, 1987; Luce, 1986). Precisely how the continuous speech signal is mapped on to discrete symbolic

representations by the listener has been and continues to be one of the most important and challenging problems to solve. In speech perception this is what is referred to as the “mapping problem.”

The description of the mapping problem in speech was first characterized by Charles Hockett in his well-known Easter-egg analogy.

“Imagine a row of Easter eggs carried along a moving belt; the eggs are of various sizes, and variously colored, but not boiled. At a certain point the belt carries the row of eggs between the two rollers of a wringer, which quite effectively smash them and rub them more or less into each other. The flow of eggs before the wringer represents the series of impulses from the phoneme source; the mess that emerges from the wringer represents the output of the speech transmitter. At a subsequent point, we have an inspector whose task it is to examine the passing mess and decide, on the basis of the broken and unbroken yolks, the variously spread out albumen, and the variously colored bits of shell, the nature of the flow of eggs which previously arrived at the wringer.”
(Hockett, 1955, p. 210)

In the field of human speech perception and spoken word recognition, the basic mapping problem has simply been ignored by speech researchers who simply assumed that the continuous speech signal could be represented and encoded as a sequence of discrete symbols and that any further processing by the nervous system used these symbolic representations (Licklider, 1952; Peterson, 1952).

Indeed, a major stumbling block is that the conventional view has routinely assumed a bottom-up approach to speech perception and spoken word recognition where phonemes are first recognized and then parsed into words (Lindgren, 1965). An alternative view of speech perception that we discuss does not suffer from this problem because it allows for a top-down approach where words are recognized as whole units first, and then segmentation into phonemes follows as a natural consequence as required by the specific behavioral task and processing demands on the listener.

New Approaches to Speech Perception and Spoken Word Recognition

Views of the mental lexicon have changed significantly in recent years (Goldinger & Azuma 2003; Goldinger, 1998; Elman, 2004; Johnson, 1997). While traditional theories of word recognition and lexical access assumed that the mental lexicon consisted of a single canonical entry for each word (Oldfield, 1966; Marslen-Wilson, 1984; Morton, 1979), recent episodic approaches to the lexicon have adopted ideas from “multiple-trace” theories of human memory which propose that multiple entries for each word are encoded and stored in lexical memory in the form of detailed perceptual traces that preserve fine phonetic detail of the original articulatory event (Goldinger, 1996; Goldinger, 1998). In contrast to the conventional abstractionist, symbol-processing views of the lexicon, current episodic approaches to spoken word recognition and lexical access emphasize the continuity and tight coupling between speech perception, speech production, and memory in language processing (Goldinger, 1996, 1997, 1998).

Nonanalytic Cognition

Over the last twenty years, a large number of studies in cognitive psychology on categorization and memory have provided evidence for the encoding and retention of episodic or “instance-specific”

information (Jacoby & Brooks, 1984; Brooks, 1978; Tulving & Schacter, 1990; Schacter, 1990, 1992). According to this nonanalytic approach to cognition, stimulus variability is viewed as "lawful" and informative in perceptual analysis (Elman & McClelland, 1986). Memory involves encoding specific perceptual episodes, as well as the processing operations used during recognition (Kolers, 1973; Kolers, 1976). The major emphasis of this view of cognition is the focus on particulars and specific instances, rather than on abstract generalizations or symbolic coding of the stimulus input into idealized categories. Thus, the intractable problem of variability found in speech perception can be approached in fundamentally different ways by nonanalytic accounts of perception and memory.

We believe that the findings from studies on nonanalytic cognition are directly relevant to several long-standing theoretical questions about the nature of perception and memory for speech. When the criteria used for postulating episodic or nonanalytic representations are examined carefully (Brooks, 1978), it becomes obvious that speech displays a number of distinctive properties that make it amenable to this approach (Jacoby & Brooks, 1984). Several properties that encourage a nonanalytic processing strategy are high stimulus variability, complex stimulus-category relations, classifying inputs under incomplete information, and classifying inputs of structures with high analytic difficulty. These criteria are summarized briefly below.

High Stimulus Variability. Stimuli with a high degree of acoustic-phonetic variability are compatible with nonanalytic representations. Speech signals display a great deal of physical variability primarily because of factors associated with the production of spoken language. Among these factors are within- and between-talker variability, such as changes in speaking rate and dialect, differences in social contexts, syntactic, semantic and pragmatic effects and emotional state, as well as a wide variety of context effects due to the ambient environment such as background noise, reverberation and microphone characteristics (Klatt, 1986). These diverse sources of variability produce large changes in the acoustic-phonetic properties of speech and they need to be accommodated in theoretical accounts of the categorization process in speech perception. Variability is an inherent property of all biological systems including speech and it cannot be ignored, designed out of experimental protocols, or simply thought of as an undesirable source of noise in the system. Variability has to be taken seriously and approached directly.

Complex Stimulus-Category Relations. Complex relations between stimuli and their category membership can also be captured by nonanalytic processing strategies. In speech, the relation between the physical acoustic stimulus and its categorization as a string of symbols is complex because of the large amount of variability within a particular speaker across different phonetic contexts and the enormous variability across speakers. Despite these differences, categorization is reliable and robust (Twaddell, 1952). The conventional use of phonemes as perceptual units in speech perception entails a set of complex assumptions about category membership. These assumptions are based on linguistic criteria involving principles such as complementary distribution, free variation and phonetic similarity. In traditional linguistics, for example, the concept of a phoneme as a basic primitive of speech is used in a number of quite different ways. Gleason (1961), for example, characterizes the phoneme as a minimal unit of contrast, as the set of allophones of a phoneme, and as a non-acoustic abstract unit of a language. Thus, like other category domains studied by cognitive psychologists, speech sounds display complex stimulus-category relations that place strong constraints on the class of categorization models that can account for these operating principles.

Classifying Stimuli with Incomplete Information. Classifying incomplete or degraded stimuli is also consistent with nonanalytic analysis. Speech is a system that allows classification under highly degraded or incomplete information, such as silent-center vowels (Jenkins, Strange, & Trent, 1999),

speech processed through a cochlear implant simulator (Shannon et al., 1995), speech mixed with noise (Miller, Heise, & Lichten, 1951), or sinewave speech (Remez, Rubin, Pisoni, & Carrell, 1981). Correct classification of speech under these impoverished conditions is possible because speech is a highly redundant system which has evolved to maximize the transmission of linguistic information. In the case of speech perception, numerous studies have demonstrated the existence of multiple speech cues for almost every phonetic contrast (Raphael, 2005). While these speech cues are for the most part highly context-dependent, they also provide reliable information that can facilitate recognition of the intended message even when the signal is presented under poor listening conditions. This feature of speech perception permits very high rates of information transmission using sparsely-coded and broadly-specified categories (Pollack, 1952, 1953).

Classification of Stimuli with High Analytic Difficulty. Stimuli with high analytic difficulty are those which differ along one or more dimensions that are difficult to quantify or describe. Because of the complexity of speech and its high acoustic-phonetic variability, the category structure of speech is not amenable to simple hypothesis testing. As a result, it has been extremely difficult to construct a set of explicit formal rules that can successfully map multiple speech cues onto discrete phoneme categories. Moreover, the perceptual units of speech are also highly automatized; the underlying category structure of a language is learned in a tacit and incidental way by young children.

Episodic Approaches to Speech Perception

Not only has the focus in speech perception changed in recent years, but the conceptions of the mental lexicon and the nature of the representation of words in lexical memory have also undergone substantial revisions and development based on new findings and theoretical proposals from several different disciplines. The recent episodic approaches to the lexicon considered here assume that spoken words are represented in lexical memory as a collection of specific individual perceptual episodes or tokens rather than the conventional abstract symbolic word types that have been universally assumed in the past. Evidence supporting episodic exemplar-based approaches to the mental lexicon has accumulated over the last few years as researchers from a number of related disciplines recognize the potential theoretical power and utility of this conceptual framework. Recent studies on the processing of stimulus variability provide evidence for episodic models of speech perception and spoken word recognition.

According to episodic views of perception and memory, listeners encode “particulars,” that is, specific instances or perceptual episodes, rather than generalities or abstractions (Kruschke, 1992; Nosofsky, 1986). Abstraction “emerges” from computational processes at the time of retrieval (Nosofsky, 1986; Estes, 1994). A series of studies carried out in our lab has shown that “indexical” information about a talker's voice and face and detailed information about speaking rate are encoded into memory and become part of the long-term representational knowledge that a listener has about the words of his/her language (Pisoni, 1997). Rather than discarding talker-specific details of speech in favor of highly abstract representations, these studies have shown that human listeners encode and retain very fine episodic details of the perceptual event (Pisoni, 1997).

In acquiring the sound system of a language, children not only learn to develop abilities to discriminate and identify sounds, they also learn to control the motor mechanisms used in speech articulation to generate precisely the same phonetic contrasts in speech production that they have become attuned to in perception. One reason that the developing perceptual system might preserve very fine episodic phonetic details of speech, as well as the specific characteristics of the talker's voice, would be to allow young children to accurately imitate and reproduce speech patterns heard in their surrounding language learning environment (Studdert-Kennedy, 1983). Imitation skills of this kind would provide

children with an enormous benefit in rapidly acquiring the phonology of the local dialect from speakers they are exposed to early in life.

In contrast to the conventional, abstractionist approach, episodic models assume that listeners store a very large number of specific instances and then use them in an analogical rather than analytic way to categorize novel stimuli (Brooks, 1978; Whittlesea, 1987). Recent findings showing that some sources of variability disrupt language processing and that familiarity with the details of the voice benefit language processing provide converging support for the claim that very detailed, instance-specific information about speech is encoded, represented and stored in memory.

Evidence for Detailed Episodic Representations

Over the last 15 years, we have been carrying out a research program on different sources of variability in speech, specifically, variability from different talkers, speaking rates and speaking modes, to determine how these factors affect spoken word recognition. Our findings suggest that many long-standing theoretical assumptions held by researchers about basic perceptual units of speech such as features, phonemes, and syllables need to be substantially revised. In particular, assuming the existence of only abstract symbolic representations of speech cannot account for the new results showing that variability matters in speech perception and that detailed episodic information affects language processing and memory.

Encoding and Storage of Variability in Speech Perception A number of studies from our research group have explored the effects of different sources of variability on speech perception and spoken word recognition. Instead of reducing or eliminating variability in the stimulus materials, as most speech researchers have routinely done over the years, in a series of novel studies we specifically introduced variability from different talkers and different speaking rates to directly study the effects of these variables on perception (Pisoni, 1993).

Our research on this problem first began with several observations of Mullennix, Pisoni and Martin (1989) who found that the intelligibility of isolated spoken words presented in noise was affected by the number of talkers that were used to generate the test words in the stimulus ensemble. In one condition, all the words in a test list were produced by a single talker; in another condition, the words were produced by 15 different talkers, which included both male and female voices. Across three different signal-to-noise ratios, identification performance was always better for words produced by a single talker than words produced by multiple talkers. Trial-to-trial variability in the speaker's voice affected recognition performance. These findings replicated results originally reported by Peters (1955) and Creelman (1957) and suggested to us that the perceptual system is highly sensitive to talker variability and therefore must engage in some form of "recalibration" each time a novel voice is encountered during the set of test trials using multiple voices.

In a second set of experiments, Mullennix et al. (1989) measured naming latencies to the same set of words presented under single and multiple-talker two test conditions. They found that subjects were not only slower to name words presented in multiple-talker lists but they were also less accurate when their performance was compared to words from single-talker lists. Both sets of findings were surprising in light of the conventional view of speech perception because all the test words used in the experiment were highly intelligible when presented in the quiet. The intelligibility and naming data from this study immediately raised a number of additional questions about how the different perceptual dimensions of the speech signal are processed and encoded by the human listener. At the time, we followed the conventional view of speech that assumed that the acoustic attributes of the talker's voice

were processed independently of the linguistic properties of the signal, although no one had ever tested this assumption directly.

In another series of experiments, Mullennix and Pisoni (1990) used a speeded classification task to assess whether attributes of a talker's voice are perceived independently of the phonetic form of the words. Subjects were required to attend selectively to one stimulus dimension (e.g., talker voice) while simultaneously ignoring another stimulus dimension (e.g., phoneme). Across all conditions, Mullennix and Pisoni found increases in interference from both perceptual dimensions when the subjects were required to attend selectively to only one of the stimulus dimensions. The pattern of results suggested that words and voices were processed as integral dimensions; that is, the perception of one dimension (e.g., phoneme) affects classification of the other dimension (e.g., voice) and vice versa. Subjects could not selectively ignore irrelevant variation in the non-attended dimension. If both perceptual dimensions were processed separately, as we originally assumed, interference from the non-attended dimension should not have been observed. Not only did we find mutual interference between the two dimensions suggesting that the perceptual dimensions were perceived in a mutually-dependent manner, but we also found that the pattern of interference was asymmetrical. It was easier for subjects to ignore irrelevant variation in the phoneme dimension when their task was to classify the voice dimension than it was for them to ignore the voice dimension when they had to classify the phonemes.

The results from these novel perceptual experiments were surprising to us at the time given our original assumption that the indexical and linguistic properties of speech are perceived independently. To study this problem further, we carried out a series of memory experiments to assess the mental representation of speech in long-term memory. Experiments on serial recall of lists of spoken words by Martin, Mullennix, Pisoni, and Summers (1989) and Goldinger, Pisoni, and Logan (1991) demonstrated that specific details of a talker's voice are not lost or discarded during early perceptual analysis but are perceived and encoded in long-term memory along with item information. Using a continuous recognition memory procedure, Palmeri, Goldinger, and Pisoni (1993) found that detailed episodic information about a talker's voice is also encoded in memory and is available for explicit judgments even when a great deal of competition from other voices is present in the test sequence.

In a series of other recognition memory experiments, Goldinger (1998) found strong evidence of implicit memory for attributes of a talker's voice which persists for a relatively long period of time (up to a week) after perceptual analysis has been completed. Moreover, he also showed that the degree of perceptual similarity between voices affects the magnitude of repetition priming effects, suggesting that the fine phonetic details are not lost and the perceptual system encodes very detailed talker-specific information about spoken words in episodic memory representations (see Goldinger, 1997).

Other experiments were carried out to examine the effects of speaking rate on perception and memory. These studies, which were designed to parallel the earlier experiments on talker variability, also found that the perceptual details associated with differences in speaking rate were not lost as a result of perceptual analysis. In one experiment, Sommers, Nygaard, and Pisoni (1992) found that words were identified more poorly when speaking rate was varied (i.e., fast, medium and slow), than when the same words produced at a single speaking rate. These results were compared to another condition in which differences in amplitude were varied randomly from trial to trial in the test sequences. In this case, identification performance was not affected by variability in overall signal level.

The effects of speaking rate variability have also been observed in experiments on serial recall. Nygaard, Sommers, and Pisoni (1992) found that subjects recalled words from lists produced at a single speaking rate better than the same words produced at several different speaking rates. Interestingly, the

differences appeared in the primacy portion of the serial position curve suggesting greater difficulty in the transfer of items into long-term memory. The effects of differences in speaking rate, like those observed for talker variability in our earlier experiments, suggested that perceptual encoding and rehearsal processes, which are typically thought to operate on only abstract symbolic representations, are also influenced by low-level perceptual sources of variability. If these sources of variability were automatically filtered out or normalized by the perceptual system at early stages of analysis, differences in recall performance would not be expected in memory tasks like the ones used in these experiments. Taken together, the findings on variability and speaking rate suggest that details of the early perceptual analysis of spoken words are not lost as a result of early perceptual analysis. Detailed perceptual information becomes an integral part of the mental representation of spoken words in memory. In fact, in some cases, increased stimulus variability in an experiment may actually help listeners to encode items into long-term memory because variability helps to keep individual items in memory more distinct and discriminable, thereby reducing confusability and increasing the probability of correct recall (Goldinger, Pisoni, & Logan, 1991; Nygaard, Sommers, & Pisoni, 1992). Listeners encode speech signals along many perceptual dimensions and the memory system apparently preserves these perceptual details much more reliably than researchers believed in the past.

Reinstatement in Speech Perception and Spoken Word Recognition. Our findings on the perception of talker variability and speaking rate encouraged us to examine perceptual learning in speech more carefully, specifically, the rapid tuning or perceptual adaptation that occurs when a listener becomes familiar with the voice of a specific talker (Nygaard, Sommers & Pisoni, 1994). This particular problem has not received very much attention in the field of human speech perception despite its obvious relevance to problems of speaker normalization, acoustic-phonetic invariance and the potential application to automatic speech recognition and speaker identification (Fowler, 1990; Kakehi, 1992; Bricker & Pruzansky, 1976). An extensive search of the research literature on talker adaptation by human listeners revealed only a small number of behavioral studies on this topic and all of them appeared in obscure technical reports from the late 1940s and early 1950s (Mason, 1946; Miller, Wiener, & Stevens, 1946; Peters, 1955).

To determine how familiarity with a talker's voice affects the perception of spoken words, Nygaard, Sommers, and Pisoni (1994) had two groups of listeners learn to explicitly identify a set of ten unfamiliar voices over a nine-day period using common names (i.e., Bill, Joe, Sue, Mary). After this initial learning period, subjects participated in a word recognition experiment designed to measure speech intelligibility. Subjects were presented with a set of novel words mixed in noise at several signal-to-noise ratios. One group of listeners heard the words produced by talkers that they were previously trained on, and the other group heard the same words produced by a new set of unfamiliar talkers. In the word recognition task, subjects were required to identify the words rather than recognize the voices, as they had done in the first phase of the experiment.

The results of the speech intelligibility experiment showed that the subjects who had heard novel words produced by familiar voices were able to recognize the novel words more accurately than subjects who received the same novel words produced by unfamiliar voices. Differences in inherent intelligibility between the two sets of words was not a confounding factor. An additional study with two new sets of untrained listeners confirmed that both sets of voices were equally intelligible, indicating that the difference in performance found in the original study was due to training.

The findings from this voice learning experiment demonstrate that exposure to a talker's voice facilitates subsequent perceptual processing of novel words produced by a familiar talker. Thus, speech

perception and spoken word recognition draw on highly specific perceptual knowledge about a talker's voice that was obtained in an entirely different experimental task.

What kind of perceptual knowledge does a listener acquire when he listens to a speaker's voice and is required to carry out an explicit name recognition task like our subjects did in this experiment? One possibility is that the procedures or perceptual operations (Kolers, 1973) used to recognize the voices are encoded and retained in some type of "procedural memory" and these perceptual analysis routines are invoked again when the same voice is encountered in a subsequent intelligibility test. This kind of procedural knowledge might increase the efficiency of the perceptual analysis for novel words produced by familiar talkers because detailed analysis of the speaker's voice would not have to be carried out over and over again as each new word was encountered. Another possibility is that specific instances - perceptual episodes or exemplars of each talker's voice are encoded and stored in memory and then later retrieved during the process of word recognition when new tokens from a familiar talker are encountered (Jacoby & Brooks, 1984).

Whatever the exact nature of this procedural knowledge turns out to be, the important point to emphasize here is that prior exposure to a talker's voice facilitates subsequent recognition of novel words produced by the same talkers. Such findings demonstrate a form of source memory for a talker's voice that is distinct from the retention of the individual items used and the specific task that was originally employed to familiarize the listeners with the voices (Glanzer, Hilford, & Kim, 2004; Roediger, 1990; Schacter, 1992). These findings provide additional support for the view that the internal representation of spoken words encompasses both a phonetic description of the utterance, as well as information about the structural description of the source characteristics of the specific talker. The results of these studies suggest that speech perception is carried out in a "talker-contingent" manner; the indexical and linguistic properties of the speech signal are closely coupled and are not dissociated into separate, independent channels in perceptual analysis.

Anti-Representationalist Approaches

Another more radical approach to representations has been proposed recently by a group of artificial intelligence (AI) researchers working on behavior-based autonomous robotics and biological intelligence (Brooks, 1991a,b; Beer, 2000; Clark, 1999). According to this perspective, called "embodied cognition," mind, body and world are linked together as a "coupled" dynamic system (Beer, 2000; Clark, 1999). Conventional abstract symbolic representations and information processing involving the manipulation of abstract symbols are not needed to link perception and action directly in real-world tasks, such as navigating around in novel unpredictable environments. Modest degrees of intelligent behavior have been achieved in robots without computation and without complex knowledge structures representing models of the world (Brooks, 1991a,b). Intelligent adaptive behavior reflects the operation of the whole system working together in synchrony without control by a dedicated central executive that is needed to access and manipulate abstract symbolic representations and guide behavior based on internal models of the world.

These are strong claims and important criticisms to raise about the most central and basic foundational assumptions of classical cognition and traditional information processing approaches to perception, memory, learning and language. While the bulk of the research efforts on embodied and situated cognition has come from the field of AI and is related to constructing autonomous robots and establishing links between perception and action in simple sensory-motor systems, the recent arguments against conventional abstract symbolic representations and the mainstream symbol-processing views of cognition and intelligence have raised a number of challenging theoretical issues that are directly relevant

to current theoretical assumptions about representations and processes in speech perception and spoken word recognition. With regard to the problems of representations in speech perception and spoken word recognition, these issues are concerned directly with questions about “representational specificity” and the nature of lexical representations assumed in spoken word recognition and comprehension. Such an anti-representation view of spoken language has been proposed recently by Port who argues that discrete representations are not needed for real-time human speech perception (Port & Leary, 2005).

Although the anti-representation theorists working in AI have argued that it is not necessary to postulate conventional symbolic representations or even to assume complex mediating states corresponding to internal models of the world for the relatively simple sensory and motor domains they have explored so far, there are several reasons to believe that their global criticisms of the conventional symbol-processing approach to cognition may not generalize gracefully to more complex knowledge-based cognitive domains (Markman & Dietrich, 2000). Compared to the simple sensory-motor systems and navigational behaviors studied by researchers working on autonomous robotics, there is good consensus that speech perception and spoken language processing are “informationally-rich” and “representationally-hungry” knowledge-based domains (Clark, 1997) that shares computational properties with a small number of other complex self-diversifying systems. These are systems like language, genetics, and chemistry that have a number of highly distinctive powerful combinatorial properties that set them apart and make them uniquely different from other natural complex systems that have been studied in the past.

Several years ago, William Abler (1989) examined the properties of self-diversifying systems and drew several important parallels with speech and spoken language. He argued that human language displays structural properties that are consistent with other “particulate systems” such as genetics and chemical interaction. All of these systems have a small number of basic “particles” such as genes or atoms that can be combined and recombined to create infinite variety and unbounded diversity without blending of the individual components or loss of perceptual distinctiveness of the new patterns created by the system.

It is hard to imagine that any of the anti-representationalists would seriously argue or even try to maintain that speech and spoken language are non-representational or non-symbolic in nature. The mere existence of reading, orthographies and alphabetic writing systems can be taken as strong evidence and serve as an existence proof that some aspects of speech and spoken language can be represented discretely and efficiently by a linear sequence of abstract symbols. Looking at several selected aspects of speech and the way spoken languages work, it is obvious that spoken language can be offered as the prototypical example for a symbol-processing system. Indeed, this is one of the major “design features” of human language (Hockett, 1960).

Evidence for Symbolic Representations in Speech Perception

For a number of years, there has been an on-going debate concerning the role of segmental representations in speech perception and spoken word recognition. Several theorists have totally abandoned an intermediate segmental level of representation in favor of direct access models of spoken word recognition (Gaskell & Marslen-Wilson, 1997; Klatt, 1979). In these models, words are recognized without an analysis of their “internal structure” into units like phones, allophones, phonemes, diphones, or demisyllables. In this section, we present arguments against this position and summarize evidence from several different areas supporting the existence of discrete segmental units in speech perception and spoken word recognition.

The first general line of evidence we offer in support of segmental representations in speech perception comes from linguistics. One of the fundamental assumptions of linguistic analysis is that the continuously varying speech waveform can be represented as a sequence of discrete units such as features, phones, allophones, phonemes, and morphemes. This assumption is central to all current conceptions of language as a system of rules that governs the sound patterns and sequences used to encode meanings (Chomsky & Halle, 1968). The widespread existence of a range of phonological phenomena such as alternation, systematic regularity, and diachronic and synchronic sound changes require, ipso facto, that some type of segmental level be postulated in order to capture significant linguistic generalizations that exist within and between languages. In describing the sound structure of a given language, then, a level of segmental representation is required in order to account for the idiosyncratic and predictable regularities of the sound patterns of that language (see Kenstowicz & Kisseberth, 1979). Whether these segmental units are actually used by human listeners in the real-time analysis of spoken language is another matter.

The second general line of evidence in support of the segmental representations in speech perception is psychological in nature. One source of evidence comes from observations of speakers of languages with no orthography who are attempting to develop writing systems. In his well-known article on the psychological reality of phonemes, Edward Sapir (1933) cites several examples of conscious awareness of the phonological structure language. Read (1971) also described a number of examples of children who have invented their own orthographies spontaneously. The children's initial encounters with print show a systematic awareness of the segmental structure of language, thereby demonstrating an ability to analyze spoken language into representations with discrete segments. Several theorists have also proposed that young children's ability to learn to read an alphabetic writing system like English orthography is highly dependent on the development of phonemic analysis skills, that is, perceptual and linguistic skills that permit the child to consciously analyze speech into segmental units (Lieberman, Shankweiler, Fischer, & Carter, 1974; Rozin & Gleitman, 1977; Treiman, 1980).

The existence of language games based on insertion of a sound sequence, movement of a sound sequence, or deletion of a sound sequence all provide additional support for the existence of segmental units in the internal structure of words (see Treiman, 1983, 1985). The presence of rhymes and the metrical structure of poetry also entail an awareness that words have an internal structure and organization and that this structure can be represented as linear sequence of discrete symbolic units distributed in time.

An examination of errors in speech production also provides additional evidence that words are represented in the lexicon in terms of segments. The high frequency of single segment speech errors such as substitutions and exchanges provide evidence of the phonological structure of the language (Fromkin, 1973, 1980; Garrett, 1976, 1980; Shattuck-Hufnagel & Klatt, 1979; Stemberger, 1982). It is difficult, if not impossible, to explain these kinds of error patterns without assuming some kind of segmental representation in the organization of the lexicon used for speech production.

Other evidence comes from studies of the phoneme-restoration effect (Samuel, 1981 a,b; Warren, 1970), a phenomenon demonstrating the on-line synthesis of the segmental properties of fluent speech. Many studies have also been carried out using the phoneme monitoring task in which subjects are required to detect the presence of a specified target phoneme while listening to sentences or short utterances (see Foss, Hawood, & Blank, 1980). Although some earlier findings suggested that listeners first recognize the word and then carry out an analysis of the segments within the word (Foss and Swinney, 1973; Morton and Long, 1976), other studies indicate that subjects can detect phonemes in nonwords that are not present in their lexicon (Foss & Blank, 1980; Foss & Gernsbacher, 1983). Thus,

subjects can detect phonemes based on two sources of knowledge: information from the sensory input and information constructed from their knowledge of the phonological structure of the language (Dell & Newman, 1980).

Finally, in terms of perceptual data, there is a growing body of data on misperceptions of fluent speech (Bond & Garnes, 1980; Bond & Robey, 1983; Bond, 2005). The errors collected in these studies also suggest that a very large portion of the misperceptions involve segments rather than whole words or features.

Recognizing the existence of segments in the representation of phonology does not imply that this is the only information included in the representation. Indeed, this is precisely what we argued against earlier. Indexical properties of speech also contribute to the representation and processing of speech and language, especially under highly degraded listening conditions when multiple sources of knowledge are routinely used to perceive and interpret highly impoverished, partially-specified speech signals.

Representational Specificity

In our view, the current debate that emerges from the embodied cognition criticisms of conventional, symbolic representations is not about whether spoken language is a symbol processing system or whether there are representations and internal states. In the case of language, the evidence is pretty clear; low-dimensional segmental units exist at multiple levels of language. The principal theoretical issues revolve around describing more precisely the exact nature of the phonetic, phonological and lexical representations used in speech perception, production and spoken language processing and the degree of representational specificity these representations preserve.

Two major questions emerge: (a) how much detail of the original speech signal is encoded by the brain and nervous system in order to support language processing and (b) how much detail can be discarded as a consequence of phonological and lexical analysis? Some sources of information in speech are clearly more important and linguistically significant than others and understanding these particular properties of the speech signal may provide new insights into both representation and process and may help to resolve many of the long-standing issues in the field. Moreover, the results from numerous perceptual studies with human listeners over the last 50 years indicate that the distinctive properties of speech vary with the specific task demands placed on the listener as well as properties of the talker. Thus, there may not be one basic unit of perception or only one common representational format in speech perception and spoken word recognition. It is very likely there are multiple units and several different representations that are used in parallel (see Pisoni & Luce, 1987).

Interface Between Speech Perception and Spoken Word Recognition

The conventional symbol-processing approach to speech has a long history dating back to the early days of telephone communications (Allen, 1994, 2005; Fletcher, 1953). The principal assumption of this bottom-up approach to spoken language processing is that spoken word recognition is logically based on prior phoneme identification and that spoken words are recognized by recovering and identifying sequences of phonemes from the acoustic-phonetic information present in the speech waveform. In the early days of speech research, the basic building blocks of speech—the perceptual primitives, were universally assumed to be the discrete segments and symbols—phones or phonemes that were derived from linguistic analysis of speech (Fano, 1950; Licklider, 1952; Peterson, 1952).

According to this conventional approach, speech perception is equivalent to phoneme perception. As the thinking went at the time, if a listener could recognize and recover the phonemes from the speech waveform like reading discrete letters on the printed page, he/she would be successful in perceiving the component words and understanding the talker's intended message (Allen, 2005). This bottom-up reductionist approach to speech perception was readily embraced and universally adopted by engineers, psychologists, and linguists and this view of speech perception is still widely accepted in the field of speech science even today despite the technical and conceptual difficulties that have been encountered over the last 50 years in trying to identify reliable discrete physical units in the speech waveform that correspond uniquely to the component sound segments of the linguistic message resulting from perceptual analysis. The primary problem of this bottom-up approach is its inability to deal with the enormous amount of acoustic-phonetic variability that exists in the speech waveform.

The conventional bottom up "segmental view" of speech perception and spoken language processing was significantly transformed and recast in a fundamentally different way in the early 1980's by Marslen-Wilson (Marslen-Wilson & Welsh, 1978). He argued convincingly that the primary objective of the human language comprehension system is the recognition of spoken words rather than the identification of individual phonemes in the speech waveform (see also Blesser, 1972). Marslen-Wilson proposed that the level at which lexical processing and word recognition is carried out in language comprehension should be viewed as the functional locus of the interactions between the initial bottom-up sensory input in the speech signal and the listener's contextual-linguistic knowledge of the structure of language. Thus, spoken word recognition was elevated to a special and privileged status within the conceptual framework of the Cohort Theory of spoken language processing developed by Marslen-Wilson and his colleagues (Marslen-Wilson, 1984). Speech perception is thus no longer simply phoneme perception, but it is also the process of recognizing spoken words and understanding sentences.

Cohort theory has been extremely influential in bringing together research scientists working in what at the time were two quite independent fields of research on spoken language processing—speech and hearing scientists who were studying speech cues and speech sound perception and psycholinguists who were investigating spoken word recognition, lexical access and language comprehension. The theoretical assumptions and strong claims of cohort theory served to focus and solidify research efforts on common problems that were specific to speech perception and spoken language processing as well as a set of new issues surrounding the organization of words in the mental lexicon (Grosjean & Frauenfelder, 1997). Segments and phonemes "emerge" from the process of lexical recognition and selection rather than the other way around. Lexical segmentation, then, may actually be viewed as a natural by-product of the primary lexical recognition process itself (Reddy, 1975).

Closely related to Cohort Theory is the Neighborhood Activation Model (NAM) developed by Luce and Pisoni (1998). NAM confronts the acoustic-phonetic invariance problem more directly by assuming that a listener recognizes a word "relationally" in terms of oppositions and contrasts with phonologically similar words. Like the Cohort Model, the focus on spoken word recognition in NAM avoids the long-standing problem of recognizing individual phonemes and features of words directly by locating and identifying invariant acoustic-phonetic properties. A key methodological tool of the NAM has been the use of a simple similarity metric for estimating phonological distances of words using a one-phoneme substitution rule (Greenberg & Jenkins, 1964; Pisoni, Nusbaum, Luce, & Slowiaczek, 1985). This computational method provided an efficient way of quantifying the "perceptual similarity" between words in terms of phonological contrasts among minimal pairs.

As Luce and McLennan (2005) have recently noted in their discussion of the challenges of variation in speech perception and language processing, all contemporary models of the spoken word

recognition assume that speech signals are represented in memory using conventional abstract representational formats consisting of discrete features, phones, allophones or phonemes. Current models of spoken word recognition also routinely assume that individual words are represented discretely and are organized in the mental lexicon. All of the current models also assume that the mental lexicon contains abstract idealized word “types” that have been normalized and made equivalent to some standard representation. None of the current models encode or store specific instances of individual word “tokens” or detailed perceptual episodes of speech (but see Kapatsinski, forthcoming for an alternative). Not only are the segments and features of individual words abstract, but the lexical representations of words and possible nonwords, are assumed to consist of abstract types, not specific experienced tokens. The only exception to this general pattern of thinking about speech as a sequence of abstract symbols was the LAFS model proposed by Klatt (1979). The LAFS model assumed that words were represented in the mental lexicon as sequences of power spectra in a large multidimensional acoustic space without postulating intermediate phonetic representations or abstract symbols (also see Treisman 1978a, 1978b). The recognition process in LAFS was carried out directly by mapping the power spectra of sound patterns onto words without traditional linguistic features or an intermediate level of analysis corresponding to discrete segments or features. In many ways, LAFS was ahead of its time in terms of its radical assumptions that intermediate segmental representations are not needed in spoken word recognition and that an optimal system does not discard potentially useful information.

Frequency and Usage-Based Views from Linguistics

Concerns about the inadequacies of the conventional, abstractionist representations of speech have also been expressed recently by a small group of linguists who have been promoting frequency- and usage-based accounts of a range of phenomena in phonetics, phonology, and morphology. For example, Pierrehumbert (1999) argued that conventional accounts of language are unable to capture several generalizations about phonological regularity and change in language. Instead, she argues that a probabilistic or stochastic approach deals better with language-particular phonetic targets (e.g., location of cardinal vowels in the vowel space or VOT differences), phonotactics (e.g., new generalizations about word-internal consonant clusters in relation to the probability of the individual parts occurring in word-initial or final position), and morphological alternations (e.g. vowel changes like in *serene/serenity*).

Bybee has also recently suggested that fine phonetic details of specific instances of speech are retained in phonological representations (Bybee, 2005). In Bybee's model, individual tokens/exemplars are stored in memory and the frequency of these tokens accounts for resistance to morphological leveling (e.g., *keep/kept*~**keeped* versus *weep/wept*~*weeped*), phonetic reduction (e.g., the frequent “I don't know”), and grammaticalization (e.g., *gonna* < “going to” from the general motion verb construction “journeying to”, “returning to”, “going to”, etc.) (Bybee 1998, 1999, 2005). Even Donca Steriade, who has carried out extensive research in phonology within the formalist tradition has suggested recently that acoustic-phonetic variability in speech needs to be captured and represented in some fashion in linguistic representations and analysis that reflect actual experience with specific instances an individual tokens of speech (Steriade, 2001 a,b).

Conclusions

Evidence from a wide variety of studies suggests that speech is not initially perceived and transformed into idealized abstract context-independent symbolic representations like sequences of letters on the printed page. Instead, highly detailed perceptual traces representing both the “medium” (detailed source information) and the “message” (content of the utterance) are encoded and stored in memory for later retrieval in the service of word recognition, lexical access and spoken language

comprehension. A record of the processing operations and procedures used in perceptual analysis and recognition remains after the primary recognition process has been completed and this residual information is used again when the same source information is encountered in another utterance. Speech is not simply transformed or recoded into an abstract idealized symbolic code like the linear sequence of discrete segments and features resulting from a linguist's phonetic transcription. The fine phonetic details of the individual talker's articulation in production of speech are not lost or discarded as a result of early perceptual processing; instead, human listeners retain dynamic information about the sensory-motor procedures and the perceptual operations and these sources of information become an integral part of the neural and cognitive representation of speech in long-term lexical memory. The representation of speech is not an either/or phenomenon where abstraction and detailed instance-specific exemplars are mutually exclusive; evidence for both detailed episodic traces and abstract segments exist and both must be represented in memory.

The most important and distinctive property of speech perception is its perceptual robustness in the face of diverse physical stimulation over a wide range of environmental conditions that produce large changes and transformations in the acoustic signal. Listeners adapt very quickly and effortlessly to changes in speaker, dialect, speaking rate and speaking style and are able to adjust rapidly to acoustic degradations and transformations such as noise, filtering, and reverberation that introduce significant physical perturbations to the speech signal without apparent loss of performance. Investigating these remarkable perceptual, cognitive and linguistic abilities and understanding how the human listener recognizes spoken words so quickly and efficiently despite enormous variability in the physical signal and listening conditions is the major challenge for future research in speech perception and spoken word recognition.

References

- Abler, W.L. (1989). On the particulate principle of self-diversifying systems. *Journal of Social Biological Structure*, 12, 1-13.
- Allen, J.B. (1994). How do humans process and recognize speech? *IEEE Trans. Speech audio*, 2, 567-577.
- Allen, J.B. (2005). *Articulation and intelligibility*. Morgan and Claypool Publishers, San Rafael.
- Beer, R.D. (2000). Dynamical approaches to cognitive science. *Trends in Cognitive Sciences*, 4, 91-99.
- Blessner, B. (1972). Speech perception under conditions of spectral transformations: I. Phonetic characteristics. *Journal of Speech and Hearing Research*, 15, 5-41.
- Blumstein, S.E. & Stevens, K.N. (1981). Phonetic features and acoustic invariance in speech. *Cognition*, 10, 25-32.
- Bond, Z.S. (2005). Slips of the ear. In D.B. Pisoni & R.E. Remez, (eds.), *The handbook of speech perception*, pp. 290-310. Blackwell Publishing Ltd, Oxford, UK.
- Bond, Z.S. & Garnes, S. (1980). Misperceptions of fluent speech. In R.A. Cole (ed.), *Perception and production of fluent speech*. pp. 115-132. Erlbaum, Hillsdale, NJ.
- Bond, Z.S. & Robey, R.R. (1983). The phonetic structure of errors in the perception of fluent speech. In N.J.U. Lass, ed. *Speech and language: Advances in basic research and practice* (Vol. 9). pp. 249-283. Academic Press, New York.
- Bradley, D.C. & Forster, K.I. (1987). A reader's view of listening. *Cognition*, 25, 103-134.
- Bricker, P.D. & Pruzansky, S. (1976). Speaker Recognition. In N.J. Lass (ed.), *Contemporary Issues in Experimental Phonetics*, pp. 295-326. Academic Press, New York.
- Brooks, L. (1978). Non-analytic concept formation and memory for instances. In E. Rosch & B. Lloyd (eds.), *Cognition and categorization*, 169-211. Erlbaum, Hillsdale, NJ.
- Brooks, R.A. (1991a). New approaches to robotics. *Science*, 253(5025), 1227-1232.

- Brooks, R.A. (1991b). Intelligence without representation. *Artificial Intelligence*, 47, 139-159.
- Bybee, J.L. (1998). The emergent lexicon. *Chicago Linguistic Society*, 34, 421-435.
- Bybee, J.L. (1999). Usage-based phonology. In M. Darnell, E. Moravcsik, F. Newmeyer, M. Noonan, & K. Wheatley (eds.), *Functionalism and Formalism in Linguistics, Volume I: General papers; Volume II: Case studies*, pp. 211-242. John Benjamins, Amsterdam, Netherlands.
- Bybee, J. (2001). *Phonology and Language Use*. Cambridge University Press, Cambridge.
- Bybee, J.L. (2005). *The impact of use on representation: grammar is usage and usage is grammar*. Presidential address, Annual Meeting of the Linguistic Society of America, Oakland.
- Chomsky, N. & Halle, M. (1968). *The sound pattern of English*. Harper and Row, New York.
- Chomsky, N. & Miller, G.A. (1963). Introduction to the formal analysis of natural languages. In R.D. Luce, R. Bush, & E. Galanter (ed.), *Handbook of mathematical psychology (Vol 2)*, pp. 269-321. John Wiley & Sons, New York.
- Clark, A. (1997) *Being There: Putting Brain, Body, and World Together Again*. MIT Press, Cambridge, MA.
- Clark, A. (1999). An embodied cognitive science? *Trends in Cognitive Sciences*, 3, 345-351.
- Creelman, C.D. (1957). Case of the unknown talker. *Journal of the Acoustical Society of America*, 29, 655.
- Dell, G.S. & Newman, J.E. (1980). Detecting phonemes in fluent speech. *Journal of Verbal Learning and Verbal Behavior*, 19, 608-623.
- Elman, J.L. (2004). An alternative view of the mental lexicon. *TRENDS in Cognitive Sciences*, 8, 301-306.
- Elman, J.L. & McClelland, J.L. (1986). Exploiting lawful variability in the speech waveform. In J.S. Perkell & D.H. Klatt (eds.), *Invariance and Variability in Speech Processing*. pp. 360-385. Erlbaum, Hillsdale, NJ.
- Estes, W.K. (1994) *Classification and Cognition*. Oxford psychology series, No. 22. Oxford University Press, New York, NY.
- Fano, R.M. (1950). The information theory point of view in speech communication. *Journal of the Acoustical Society of America*, 22, 691-696.
- Fant, G. (1973). *Speech sounds and features*. The MIT Press, Cambridge, MA.
- Fant, G. (1962). Descriptive analysis of the acoustic aspects of speech. *Logos*, 5, 3-17.
- Fletcher, H. (1953). *Speech and Hearing in Communication*. Robert E. Krieger Publishers: Huntington, NY.
- Foss, D.J. & Blank, M.A. (1980). Identifying the speech codes. *Cognitive Psychology*, 12, 1-31.
- Foss, D.J. & Gernsbacher, M.A. (1983). Cracking the dual code: Towards a unitary model of phoneme identification. *Journal of Verbal Learning and Verbal Behavior*, 22, 609-632.
- Foss, D.J., Harwood, D.A., & Blank, M.A. (1980). Deciphering decoding decisions: Data and devices. In RA Cole (ed.), *Perception and production of fluent speech*. pp. 165-199. Erlbaum, Hillsdale NJ.
- Foss, D.J. & Swinney, D.A. (1973). On the psychological reality of the phoneme: Perception, identification, and consciousness. *Journal of Verbal Learning and Verbal Behavior*, 12, 246-257.
- Fowler, C.A. & Balantucci, B. (2005) The relation of speech perception and speech production. In D.B. Pisoni & R.E. Remez (eds.), *The handbook of speech perception*, pp. 633-652. Blackwell Publishing Ltd: Oxford, UK.
- Fowler, C.A. (1990). Listener-talker attunements in speech. *Haskins Laboratories Status Report on Speech Research*, 101-102, 110-129.
- Fromkin, V. (1980). *Errors in linguistic performance*. Academic Press: NY.
- Fromkin, V. (1973). *Speech errors as linguistic evidence*. Mouton: The Hague.
- Garrett, M.F. (1980). Levels of processing in sentence production. In B. Butterworth (ed.), *Language production (Vol. 1)*. pp. 177-221. Academic Press: NY.

- Garrett, M.F. (1976). Syntactic processes in sentence production. In R.J. Wales & E. Walker (eds.), *New approaches to language mechanisms*. pp. 231-256. North-Holland: Amsterdam.
- Gaskell, M.G. & Marslen-Wilson, W.D. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes*, *12*, 613-656.
- Glanzer, M., Hilford, A., & Kim, K. (2004). Six regularities of source recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 1176-1195.
- Gleason, H.A. (1961). *An introduction to descriptive linguistics*. Holt, Rinehart, and Winston: New York.
- Goldinger, S.D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*, 251-279.
- Goldinger, S.D. (1997). Talker variability in speech processing. In K. Johnson & J.W. Mullennix (eds.), *Talker Variability in Speech Processing*. pp. 33-66. Academic Press: San Diego.
- Goldinger, S.D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1166-1183.
- Goldinger, S.D. & Azuma, T. (2003). Puzzle-solving science: the quixotic quest for units in speech perception. *Journal of Phonetics*, *31*, 305-320.
- Goldinger, S.D., Pisoni, D.B., & Logan, J.S. (1991). On the locus of talker variability effects in recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 152-162.
- Greenberg, J.H. & Jenkins, J.J. (1964). Studies in the psychological correlates of the sound system of American English. *Word*, *20*, 157-177.
- Grosjean, F. & Frauenfelder, U.H. (ed.) (1997). *Spoken Word Recognition Paradigms: Special Issue of "Language and Cognitive Processes"*, Psychology Press, Hove: England.
- Grossberg, S. (2003). Resonant neural dynamics of speech perception. *Journal of Phonetics*, *31*, 423-445.
- Halle, M. (1956). [Review of the book *Manual of Phonology* by C.D. Hockett]. *Journal of the Acoustical Society of America*, *28*, 509-510.
- Hockett, C.D. (1960). The origin of speech. *Scientific American*, *203*, 88-96.
- Hockett, C.F. (1955). *Manual of phonology*. Indiana University Publications in Anthropology and Linguistics (No. 11), Bloomington, IN.
- Jacoby, L.L. & Brooks, L.R. (1984). Nonanalytic cognition: Memory, perception, and concept learning. In G. Bower (ed.), *The psychology of learning and motivation*, pp. 1-47. Academic Press: NY.
- Jenkins, J.J., Strange, W., & Trent, S.A. (1999) Context-independent dynamic information for the perception of coarticulated vowels. *Journal of the Acoustical Society of America*, *106*, 438-448.
- Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson & J.W. Mullennix (eds.), *Talker Variability in Speech Processing*. pp. 145-166. Academic Press, San Diego, CA.
- Joos, M.A. (1948). Acoustic phonetics. *Language*, *24*, 1-136.
- Kakehi, K. (1992). Adaptability to differences between talkers in Japanese monosyllabic perception. In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka, eds. *Speech perception, production, and linguistic structure*, pp. 135-142. Ohmsha Publishing: Tokyo.
- Kapatsinski, V.M. (Forthcoming). Towards a single-mechanism account of frequency effects. *Proceedings of LACUS 32: Networks*, Hanover: NH.
- Kenstowicz, M. & Kisseberth, C. (1979). *Generative phonology*. Academic Press: New York/London.
- Klatt, D.H. (1986). The problem of variability in speech recognition and in models of speech perception. In J.S. Perkell & D.H. Klatt (eds.), *Invariance and Variability in Speech Processing*. pp. 300-319. Erlbaum, Hillsdale: NJ.
- Klatt, D.H. (1979). Speech perception: A model of acoustic-phonetic analysis and lexical access. *Journal of Phonetics*, *7*, 279-312.
- Kolers, P.A. (1976). Pattern-analyzing memory. *Science*, *191*, 1280-1281.

- Kolers, P.A. (1973). Remembering operations. *Memory and Cognition*, 1, 347-355.
- Kruschke, J.K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, 99, 22-44.
- Liberman, A.M., Cooper, F.S., Shankweiler, D.P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, 74, 431-361.
- Liberman, A.M., Delattre, P.C., & Cooper, F.S. (1952). The role of selected stimulus-variables in the perception of the unvoiced stop consonants. *The American Journal of Psychology*, 65, 497-516.
- Liberman, A.M., Delattre, P.C., Cooper, F.S., & Gerstman, L.J. (1954). The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs*, 68, 1-13.
- Liberman, I.Y., Shankweiler, D., Fischer, F.W., & Carter, B. (1974). Explicit syllable and phoneme segmentation in the young child. *Journal of Experimental Child Psychology*, 18, 201-212.
- Licklider, J.C.R. (1952). On the process of speech perception. *Journal of the Acoustical Society of America*, 24, 590-594.
- Lindgren, N. (1965). Machine Recognition of Human Language. *IEEE Spectrum*, March and April 1965.
- Luce, P.A. (1986). Neighborhoods of words in the mental lexicon. *Research on Speech Perception, Technical Report No. 6*. Department of Psychology, Speech Research Laboratory, Bloomington, IN.
- Luce, P.A. & McLennan, C.T. (2005). Spoken word recognition: The challenge of variation. In D.B. Pisoni & R.E. Remez (eds.), *The handbook of speech perception*, pp. 591-609. Blackwell Publishing Ltd: Oxford UK.
- Luce, P.A. & Pisoni, D.B. (1998). Recognizing spoken words: The Neighborhood Activation Model. *Ear and Hearing*, 19, 1-36.
- Markman, A.B. & Dietrich, E. (2000). Extending the classical view of representation. *Trends in Cognitive Sciences*, 4, 470-475.
- Marslen-Wilson, W.D. (1984). Function and process in spoken word recognition. A tutorial review. In H. Bouma & D.G. Bouwhuis (eds.), *Attention and Performance X: Control of Language Processes*. pp. 125-150. Erlbaum: Hillsdale, NJ.
- Marslen-Wilson, W.D. & Welsh, A. (1978) Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, 10, 29-63
- Martin, C.S., Mullennix, J.W., Pisoni, D.B., & Summers, W.V. (1989). Effects of talker variability on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 15, 676-684.
- Mason, H. (1946). Understandability of speech in noise as affected by region of origin of speaker and listener. *Speech Monographs*, 13, 54-58.
- McNellis, M.G. & Blumstein, S.E. (2001). Self-organizing dynamics of lexical access in normals and aphasics. *Journal of Cognitive Neuroscience*, 13, 151-170.
- Miller, G.A., Heise, G.A., & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test material. *Journal of Experimental Psychology*, 41, 329-335.
- Miller, G.A., Wiener, F.M., & Stevens, S.S. (1946). *Combat instrumentation. II. Transmission and reception of sounds under combat conditions*. Summary Technical Report of NDRC Division 17.3. NDRC (government): Washington, DC.
- Morton, J. (1979). Word recognition. In J. Morton & J.C. Marshall (eds.), *Structures and Processes*. pp. 108-156. MIT Press: Cambridge.
- Morton, J. & Long, J. (1976). Effect of word transitional probability on phoneme identification. *Journal of Verbal Learning and Verbal Behavior*, 15, 43-52.
- Mullennix, J.W., & Pisoni, D.B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception and Psychophysics*, 47, 379-390.

- Mullennix, J.W., Pisoni, D.B., & Martin, C.S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America*, 85, 365-378.
- Murphy, G.L. (2002). *The big book of concepts*. MIT Press: Cambridge, MA.
- Nosofsky, R.M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115, 39-57.
- Nygaard, L.C., Sommers, M.S., & Pisoni, D.B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, 5, 42-46.
- Nygaard, L.C., Sommers, M.S., & Pisoni, D.B. (1992). Effects of speaking rate and talker variability on the representation of spoken words in memory. *Proceedings 1992 International Conference on Spoken Language Processing*, Banff, Canada, Oct. 12-16, pp. 209-212.
- Oldfield, R.C. (1966). Things, words and the brain. *Quarterly Journal of Experimental Psychology*, 18, 340-353.
- Palmeri, T.J., Goldinger, S.D., & Pisoni, D.B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19, 309-328.
- Pardo, J.S. & Remez, R.E. (in press). The perception of speech. In M. Traxler & M.A. Gernsbacher (eds.), *The Handbook of Psycholinguistics*. Elsevier, New York.
- Peters, R.W. (1955). *The relative intelligibility of single-voice and multiple-voice messages under various conditions of noise* (Joint Project Report No. 56, pp. 1-9). US Naval School of Aviation Medicine: Pensacola, FL.
- Peterson, G. (1952). The information-bearing elements of speech. *Journal of the Acoustical Society of America*, 24, 629-637.
- Peterson, G.E. & Barney, H.L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, 24, 175-184.
- Pierrehumbert, J.B. (1999). What people know about sounds of language. *Studies in the Linguistic Sciences*, 29, 111-120.
- Pierrehumbert, J.B. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In J. Bybee & P. Hopper (eds.), *Frequency and the Emergence of Linguistic Structure*. pp. 137-158. John Benjamins: Amsterdam.
- Pierrehumbert, J.B. & Pierrehumbert, R.T. (1990). On attributing grammars to dynamical systems. *Journal of Phonetics*, 18, 465-477.
- Pisoni, D.B. (1997). Some Thoughts on "Normalization" in Speech Perception. In K. Johnson & J.W. Mullennix (Eds.), *Talker Variability in Speech Processing*. pp. 9-32. Academic Press, San Diego.
- Pisoni, D.B. (1993). Long-term memory in speech perception: Some new findings on talker variability, speaking rate and perceptual learning. *Speech Communication*, 13, 109-125.
- Pisoni, D.B. & Luce, P.A. (1987). Acoustic-phonetic representations in word recognition. *Cognition*, 25, 21-52.
- Pisoni, D.B., Nusbaum, H.C., Luce, P.A., & Slowiaczek, L.M. (1985). Speech perception, word recognition and the structure of the lexicon. *Speech Communication*, 4, 75-95.
- Pisoni, D.B. & Remez, R.E. (Eds.). (2005). *The Handbook of Speech Perception*. Blackwell Publishing: Malden, MA.
- Pollack, I. (1953). The information of elementary auditory displays II. *Journal of the Acoustical Society of America*, 25, 765-769.
- Pollack, I. (1952). The information of elementary auditory displays. *Journal of the Acoustical Society of America*, 24, 745-749.
- Port, R. & Leary, A. (2005). Against formal phonology. *Language*.
- Raphael, L.J. (2005). Acoustic cues to the perception of segmental phonemes. In D.B. Pisoni & R.E. Remez (eds.), *The handbook of speech perception*, pp. 182-206. Blackwell Publishing Ltd: Oxford, UK.

- Read, C. (1971). Preschool children's knowledge of English phonology. *Harvard Educational Review*, 41, 1-34.
- Reddy, R.D. (1975). *Speech recognition*. Academic Press: NY.
- Remez, R.E., Rubin, P.E., Pisoni, D.B., & Carrell, T.D. (1981). Speech perception without traditional speech cues. *Science*, 212(4497), 947-950.
- Roediger, H.L. (1990). Implicit memory: Retention without remembering. *American Psychologist*, 45, 1043-1056.
- Rozin, P. & Gleitman, L.R. (1977). The structure and acquisition of reading II: The reading process and the acquisition of the alphabetic principle. In A.S. Reber & D.L. Scarborough (eds.), *Toward a psychology of reading*. pp. 55-141. Erlbaum: Hillsdale, NJ.
- Samuel, A.G. (1981a). Phonemic restoration: Insights from a new methodology. *Journal of Experimental Psychology: General*, 110, 474-494.
- Samuel, A.G. (1981b). The role of bottom-up confirmation in the phonemic restoration illusion. *Journal of Experimental Psychology: Human Perception and Performance*, 7, 1124-1131.
- Sapir, E. (1933). La réalité psychologique des phonemes. *Journal de Psychologie Normale et Pathologique*, 30, 247-265.
- Schacter, D.L. (1992). Understanding implicit memory: A cognitive neuroscience approach. *American Psychologist*, 47, 559-569.
- Schacter, D.L. (1990). Perceptual representation systems and implicit memory: Toward a resolution of the multiple memory systems debate. *Annals of the New York Academy of Sciences*, 608, 543-571.
- Shannon, R.V., Zeng, F.G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, 270 (5234), 303-304.
- Shattuck-Hufnagel, S. & Klatt, D.H. (1979). The limited use of distinctive features and markedness in speech production: Evidence from speech error data. *Journal of Verbal Learning and Verbal Behavior*, 18, 41-45.
- Smith, E.E. & Medin, D. (1981). *Categories and Concepts*. Harvard University Press: Cambridge, MA.
- Sommers, M.S., Nygaard, L.C., & Pisoni, D.B. (1992). Stimulus variability and the perception of spoken words: Effects of variations in speaking rate and overall amplitude. *Proceedings 1992 International Conference on Spoken Language Processing*, Banff, Canada, Oct. 12-16, pp. 217-220.
- Sporns, O. (2003). Network analysis, complexity, and brain function. *Complexity*, 8, 56-60.
- Stemberger, J.P. (1982). *The lexicon in a model of language production*. Unpublished doctoral dissertation, University of California, San Diego.
- Steriade, D. (2001a). Directional asymmetries in place assimilation: A perceptual account. In E. Hume & K. Johnson (eds.), *The role of speech perception in phonology*, pp. 219-250, Academic Press, San Diego.
- Steriade, D. (2001b). *The phonology of Perceptibility Effects: the P-map and its consequences for constraint organization*. Unpublished manuscript, UCLA.
- Stevens, K.N. (1998). *Acoustic Phonetics*. MIT Press: Cambridge, MA.
- Stevens, K.N. (1972). The quantal nature of speech: Evidence from articulatory-acoustic data. In E.E. David & P.D. Denes (Eds.), *Human Communication: A Unified View*. pp. 51-66. McGraw-Hill: New York.
- Studdert-Kennedy, M. (1983). On learning to speak. *Human Neurobiology*, 2, 191-195.
- Studdert-Kennedy, M. (1974). The perception of speech. In T.A. Sebeok (ed.), *Current trends in linguistics (Vol. XII)*, pp. 2349-2385. Mouton: The Hague.
- Treiman, R. (1985). Onsets and rimes as units of spoken syllables. Evidence from children. *Journal of Experimental Child Psychology*, 39, 161-181.

- Treiman, R. (1983). The structure of spoken syllables: Evidence from novel word games. *Cognition*, *15*, 49-74.
- Treiman, R. (1980). *The phonemic analysis ability of preschool children*. Unpublished doctoral dissertation, University of Pennsylvania.
- Treisman, M. (1978a). A theory of the identification of complex stimuli with an application to word recognition. *Psychological Review*, *78*, 420-425.
- Treisman, M. (1978b). Space or lexicon? The word frequency effect and the error response frequency effect. *Journal of Verbal Learning and Verbal Behavior*, *17*, 37-59.
- Tulving, E. & Schacter, D.L. (1990). Priming and human memory systems. *Science*, *247*, 301-306.
- Twaddell, W.F. (1952). Phonemes and allophones in speech analysis. *The Journal of the Acoustical Society of America*, *24*, 607-611.
- Warren, R.M. (1970). Perceptual restoration of missing speech sounds. *Science*, *176*, 392-393.
- Whittlesea, B.W.A. (1987). Preservation of specific experiences in the representation of general knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 3-17.