CHAPTER 1

Representations and representational specificity in speech perception and spoken word recognition

David B. Pisoni and Susannah V. Levi

1.1 Introduction

Approaches to the study of speech perception and spoken word recognition have undergone rapid change over the last few years due to theoretical and methodological developments in various subfields of cognitive science. In contrast to the traditional view that speakers only represent abstractions of linguistic structure from the speech signal, several exemplar-based approaches (e.g. Goldinger, 1998; Goldinger and Azuma, 2003; Johnson, 1997) to the study of speech perception and spoken word recognition have emerged from independent developments in categorization (Kruschke, 1992; Nosofsky, 1986) and frequency-based phonology (Bybee, 2001; Pierrehumbert, 2001). These alternatives offer fresh ideas and new insights relating to old problems and questions related to variability and invariance (Pardo and Remez, forthcoming). In this chapter, we will discuss how these new approaches—coupled with previous insights—provide a new framework for questions which deal with the nature of phonological and lexical knowledge and representation, processing of stimulus variability, and perceptual learning and adaptation (see papers in Pisoni and Remez, 2005).

The fundamental issue in speech perception and spoken language processing research is to describe the cognitive processes involved in a listener's recovery of her interlocutor's intended message. This complex problem has been typically broken down into several more specific research questions. First, what stages of perceptual analysis intervene between the presentation of the speech signal and recognition of the intended message? Second, what types of processing computations occur at each stage? Third, what are the primary perceptual processing units and what is the nature and content of representations of speech in memory? We provide an overview of some recent developments in the field that bear directly on the third question (for overviews of work pertaining to the first two questions, see McQueen, Chapter 3 this volume; Gaskell, Chapter 4 this volume). In this chapter, we will present evidence and provide arguments indicating that speakers encode and represent both individual instances (or exemplars) they have encountered and abstractions over those instances.

The chapter is structured as follows. Section 1.2 outlines the traditional view of speech perception and identifies some problems with assuming such a view in which only abstract representations exist. Section 1.3 discusses some new approaches to speech perception which retain detailed information in the representations.

In section 1.4 we discuss a view which rejects abstraction altogether, but then show that such a view has difficulty dealing with a range of linguistic phenomena. Section 1.5 provides a brief discussion of some new directions in linguistics that encode both detailed information and abstraction. Finally, in section 1.6 we discuss the coupling of speech perception and spoken word recognition.

1.2 The traditional view of speech perception

1.2.1 Overview

The traditional approach to speech perception has relied on the assumptions of generative linguistics, which adopts a formalist view and focuses 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 displayed by natural languages. Within the domain of speech perception, linguists have made several foundational assumptions about speech, assuming that speech is structured in systematic ways and that the linguistically significant information in the speech signal can be represented effectively and economically as a linear sequence of abstract, discrete units using an alphabet of conventional phonetic symbols (e.g. *speech* is represented with the segments /s/, /p/, /i/, /tʃ/). Segmental representations are designed to code only the linguistically significant differences in meaning between minimal pairs of words in the language (Twaddell, 1952), and segments therefore encode idealized abstractions of speech sounds. The strong view from generative linguistics is that a speaker's representation of the sounds in her language excludes redundant or accidental information that is present in the speech signal but not linguistically contrastive. Two examples of this traditional 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: 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: 510)

This traditional view that speech is encoded by speakers as a linear sequence of abstract symbols 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 underpinnings of this view date back to the early Paninian grammarians, who first noted that words have an internal structure and differ from each other in systematic ways reflecting the phonological (and morphological) contrasts of a particular language. Although not often made explicit, this view relies on several important theoretical assumptions that are worth mentioning because they bear directly on theoretical issues related to the nature and content of lexical representations.

First, the traditional view of 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 and Pierrehumbert, 1990). Second, in this view, 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 a "particulate system" (Abler, 1989; see section 1.4 below), some degree of uncertainty still remains about the precise elemental primitives of speech, even after many years of basic and applied research. For example, what is the size of the basic building blocks of speech? Are they features, phonemes, syllables, or gestures? Are they perceptual or articulatory in nature, or are they both?

Third, the traditional view of speech perception relies heavily on a set of psychological processes that function to "normalize" acoustically different speech signals and make them functionally equivalent in perception (Joos, 1948). In this view, it is generally assumed that perceptual normalization is needed in speech perception in order to reduce acoustic-phonetic variability in the speech signal, making physically different signals (e.g. from different speakers) perceptually equivalent by bringing them into conformity

with some common standard or referent (see Pisoni, 1997).

1.2.2 Problems with the traditional view of speech perception

Several aspects of the traditional view of speech as a linear string of discrete symbols are difficult to reconcile with the continuous nature of the acoustic waveform produced by a speaker. Importantly, the acoustic consequences of coarticulation, as well as other sources of contextually conditioned variability, result in the failure of the acoustic signal to meet two formal conditions: linearity and invariance. This failure in turn gives rise to a third related problem: the absence of segmentation of the physical, acoustic speech signal into discrete units (first discussed by Chomsky and Miller, 1963). This section provides an overview of these issues faced by the traditional view of speech, leading to the suggestion that speakers must represent detailed information about the speech signal in addition to the abstracted representations discussed above.

1.2.2.1 Non-linearity of the speech signal

One fundamental problem facing the traditional view is the lack of linearity. The linearity condition states that for each phoneme in the message there must be a corresponding stretch of sound in the utterance (Chomsky and 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 which "smear" acoustic features for adjacent phonemes. For example, perceptual cues regarding the place of articulation for onset stop consonants (e.g. /b/ vs. /d/ vs. /g/) are located in the formant transitions into the following segment which follows the release of the consonant. 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 et al., 1967).

1.2.2.2 Lack of acoustic-phonetic invariance

Another important property of the speech signal which is problematic for the traditional view is the fact that speech lacks acoustic-phonetic invariance (Chomsky and Miller, 1963). Acoustic-phonetic invariance entails that every phoneme must have a specific set of acoustic attributes in all contexts (Estes, 1994; Murphy, 2002; Smith and Medin, 1981). Because of coarticulatory effects in speech production, the acoustic properties of a particular speech sound vary as a function of the phonetic environment. For example, the formant transitions for syllable-initial stop consonants which provide cues to place of articulation vary considerably depending on properties of the following vowel (Liberman et al., 1954).

In addition to within-speaker variation, acoustic-phonetic invariance is also absent when we look across speakers of a language producing a particular segment in a particular context. For example, men, women, and children with different vocal tract lengths exhibit large differences in their absolute formant values in the production of vowels (Peterson and Barney, 1952). In each case, the absence of acoustic-phonetic invariance is inconsistant with the notion that speech is represented only as an idealized string of discrete segments.

1.2.2.3 Difficulties with speech segmentation

The non-linearity of the speech signal coupled with the context-conditioned variability leads to a third problem with the traditional view of speech perception: how do we segment the speech waveform into higher-order units of linguistic analysis such as syllables and words? The previous sections highlighted that the speech signal cannot be reliably segmented into discrete acoustically defined units that are independent of adjacent segments; in fluent speech, it is typically not possible to identify where one word ends and another begins using simple acoustic criteria. Precisely how the continuous speech signal is mapped onto discrete symbolic representations by the listener continues to be one of the most important and challenging problems for speech perception research, and critically suggests the existence of additional representations that encode the gradient, continuous aspects of the speech signal.

The description of the problem of speech segmentation 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: 210)

A major stumbling block for the traditional view is that it has routinely assumed a bottom-up approach to speech perception and spoken word recognition where phonemes are first recognized from the speech signal and then parsed into words (Lindgren, 1965; Gaskell, Chapter 4 this volume; McQueen, Chapter 3 this volume). An alternative view of speech perception that we will discuss in section 1.6 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. We believe that this latter view is critical for providing an account of speech perception which incorporates both detailed instance-based representations and abstractions over those instances.

In sum, the traditional view of speech perception which asserts that only abstract representations exist faces several problems in light of the fact that the speech signal is continuous. The next section will discuss new methods of approaching speech perception and spoken word recognition that take into account the continuous nature of speech and that represent in memory highly detailed information about the signal.

1.3 New approaches to speech perception and spoken word recognition

While traditional theories of word recognition and lexical access assumed that the mental lexicon consisted of a single canonical entry for each word (Marslen-Wilson, 1984; Morton, 1979; Oldfield, 1966), 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 (Elman, 2004; Goldinger, 1996; 1998; Goldinger and Azuma, 2003; Johnson, 1997). In contrast to the traditional views of the lexicon as containing linear strings of idealized sound segments, current episodic approaches to spoken word recognition and lexical access emphasize the coupling between the neural encoding of prior perceptual experiences and the representations of sound structure active in speech processing (see Goldinger, 1998 for a full exposition of this idea). In this section, we provide arguments that speech exhibits non-analytic properties (section 1.3.1), which favors an account in which individual episodes (or exemplars) are stored in memory (section 1.3.2). Section 1.3.2 further contains a synopsis of previous experiments revealing that particular components of the speech signal—unnecessary for identification of the linguistic target—are stored in memory and affect behavior across several language processing tasks (section 1.3.3).

1.3.1 Non-analytic cognition

Over the last twenty years, a large number of studies in cognitive psychology on categorization and memory have suggested that we encode and retain "instance-specific" information across a wide variety of cognitive domains (Brooks, 1978; Jacoby and Brooks, 1984; Schacter, 1990; 1992; Tulving and Schacter, 1990). According to a non-analytic approach to cognition, the stimulus variability which is present in these instances is viewed as "lawful" and informative in perceptual analysis (Elman and McClelland, 1986). Specific perceptual episodes are encoded in memory and active in the cognitive processes involved in recognition (Kolers, 1973; 1976). Given the emphasis on the details of individual percepts, the problem of variability raised in section 1.2.2.2 can be approached in fundamentally different ways by non-analytic accounts of perception and memory. Other examples of stimuli that encourage a non-analytic approach to perception are visual object recognition (Gautier and Tarr, 2002) and faces (Rhodes et al., 2004).

When the criteria used for postulating episodic or non-analytic representations (discussed in Brooks, 1978) are examined with respect to speech, it is apparent that a number of distinctive properties of speech make it amenable to this approach (Jacoby and Brooks, 1984). This section focuses on several properties that encourage a non-analytic processing strategy, including: high stimulus variability; complex stimulus–category relations; classification of inputs under incomplete information; and classification of structures with high analytic difficulty. These criteria—and their relationship to the speech signal—are summarized briefly below.

1.3.1.1 High stimulus variability

Stimuli with a high degree of acoustic-phonetic variability are compatible with non-analytic representations. Speech signals display a great deal of physical variability due to 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 transmission media (Klatt, 1986). These diverse sources of variability produce large changes in the acoustic-phonetic properties of speech. Variability must be taken seriously and approached directly because it is an integral property of natural speech, as well as all biological systems.

1.3.1.2 Complex stimulus–category relations

Speech also displays a complex relation between the stimulus and its category membership, another property of non-analytic systems. Despite the large amount of variability in the speech signal, 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, the set of allophones of a phoneme, and a non-acoustic abstract unit of a language. Thus, like other category domains studied by cognitive psychologists, speech sounds display complex stimulus–category relations which place strong constraints on the class of categorization models that can account for these operating principles.

1.3.1.3 Classifying stimuli with incomplete information

Classifying incomplete or degraded stimuli is also consistent with non-analytic analysis. Speech is a system that allows classification under highly degraded or incomplete information, such as silent-center vowels (Jenkins et al., 1999), speech processed through a cochlear implant simulator (Shannon et al., 1995), speech mixed with noise (Miller et al., 1951), and sinewave speech (Remez et al., 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).

1.3.1.4 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.

1.3.2 Evidence favoring episodic approaches to speech perception

The recent episodic approaches to the lexicon considered here (e.g. Goldinger, 1998; Johnson, 1997) assume that spoken words are represented in lexical memory as a collection of individual perceptual tokens rather than as abstract word types. Evidence supporting episodic exemplar-based approaches to representation in the mental lexicon has accumulated over the last few years.

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 (Estes, 1994; Nosofsky, 1986). A series of studies carried out in our lab has shown that "indexical" properties of a speech token (e.g. information

about a talker's voice and detailed information about speaking rate) are encoded into memory and become part of the long-term memory representation that a listener has about the words of her language (Pisoni, 1997). Rather than discarding talker-specific details of speech in favor of only highly abstract representations, these studies have shown that human listeners encode and retain very fine episodic details of the perceptual event (Pisoni, 1997). This evidence further supports the claim that assuming the existence of only abstract symbolic representations of speech cannot account for basic phenomena in speech and language processing.

1.3.2.1 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. In a series of studies, we specifically introduced variability in our stimulus materials using tokens from different talkers and different speaking rates to directly study the effects of these variables on perception (Pisoni, 1993). For example, Mullennix et al. (1989) observed that the intelligibility of isolated spoken words presented in noise was affected by the number of talkers 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 fifteen different talkers. Across three different signal-to-noise ratios, identification performance was always better when subjects were presented with stimuli produced by a single talker than for subjects presented with stimuli produced by multiple talkers. Thus, variability in the speaker's voice led to a decline in spoken word recognition performance. These findings replicated results originally reported by Peters (1955) and Creelman (1957), and suggest 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.

In a second set of experiments, Mullennix et al. (1989) measured repetition latencies to the same set of words presented under single- and multiple-talker test conditions. They found that subjects were slower and less accurate in repeating words presented in multiple-talker lists compared to single-talker lists. As all the test words used in the experiment were highly intelligible when presented in the quiet, these results are difficult to reconcile with a view in which spoken word recognition requires that the speech signal be "normalized," leading the perceiver to discard information regarding talker identity. Thus, the data from these studies raised a number of additional questions about how different perceptual dimensions of the speech signal are processed and encoded by the human listener.

One important issue raised by these results is whether linguistic information (e.g. identity of speech sounds) is processed separately from indexical (or extralinguistic) information such as the identity of the speaker. To address this issue, 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 words. Subjects were required to attend selectively to one stimulus dimension (e.g. either talker voice or phoneme identity) while simultaneously ignoring the other dimension. Across all conditions, Mullennix and Pisoni found that when subjects were required to attend selectively to one dimension, the other dimension interfered with their performance. If these perceptual dimensions were processed separately, as was originally assumed, interference from the non-attended dimension should not have been observed. However, the observed pattern of results suggested that words and voices were not processed separately; that is, the perception of one dimension (e.g. phoneme) affected classification of the other dimension (e.g. voice). Not only did we find mutual interference between the two dimensions, 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 than it was for them to ignore the voice dimension when they had to classify the phonemes.

To further study the effects of indexical properties on speech perception, 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 et al. (1989) and Goldinger et al. (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. Using a continuous recognition memory procedure, Palmeri et al. (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 another series of 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 found that the degree of perceptual similarity between voices affects the magnitude of repetition priming effects, suggesting that fine phonetic details are not lost and the perceptual system encodes detailed talker-specific information about spoken words in episodic memory representations (see Goldinger, 1997).

Another set of experiments was carried out to examine the effects of speaking rate on perception and memory. These studies, designed to parallel the earlier experiments on talker variability, also found that the perceptual details associated with differences in speaking rate are not lost as a result of perceptual analysis. In one experiment, Sommers et al. (1992) found that identification of words was affected by variation in speaking rate (i.e. fast, medium, and slow) compared to a condition in which the same words were produced at a single speaking rate. However, when differences in amplitude were varied randomly from trial to trial, identification performance was not affected by variability in overall signal level.

Effects of speaking rate variability have also been observed in experiments involving a serial recall task. Nygaard et al. (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, suggest that perceptual encoding and rehearsal processes are 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. Rather, detailed perceptual information of spoken words is represented in memory. In fact, in some cases increased stimulus variability in an experiment may actually help listeners encode items in long-term memory because variability helps keep individual items more distinct and discriminable, thereby reducing confusability and increasing the probability of correct recall (Goldinger et al., 1991; Nygaard et al., 1992). Listeners encode speech signals along many perceptual dimensions, and the memory system apparently preserves these details much more reliably than researchers believed in the past.

1.3.2.2 Talker-specific speech perception and spoken word recognition

Our findings on the effects of talker variability and speaking rate on perception encouraged us to examine perceptual learning in speech more carefully. Specifically, we investigated the rapid tuning or perceptual adaptation that occurs when a listener becomes familiar with the voice of a particular talker (Nygaard et al., 1994). This 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 (Bricker and Pruzansky, 1976; Fowler, 1990; Kakehi, 1992).

To determine how familiarity with a talker's voice affects the perception of spoken words, Nygaard et al. (1994) trained two groups of listeners to explicitly identify a set of ten unfamiliar voices over a nine-day period. 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 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. The results revealed that subjects who 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. 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, not inherent intelligibility between the two sets of words.

These findings demonstrate that exposure to a talker's voice facilitates subsequent perceptual processing of novel words produced by that talker. Thus, speech perception and spoken word recognition incorporate highly specific perceptual knowledge about a talker's voice.

More recently, Allen and Miller (2004) have also shown the effects of talker-specific knowledge in a task which examined listeners' sensitivity

to sub-phonemic acoustic differences. Listeners were trained on the voices of two talkers, one with long voice onset times (VOTs) and one with short VOTs. During the test phase, listeners were able to generalize talker-specific VOT differences to novel words, indicating that listeners' sensitivity to sub-phonemic acoustic-phonetic differences was retained and used in subsequent language processing tasks.

Similarly, Eisner and McQueen (2005) and Kraljic and Samuel (2005) also observed talker-specific sub-phonemic attunement for fricatives. Eisner and McQueen trained listeners with an ambiguous fricative in either an [f]- or [s]-biasing lexical context. During the testing phase, listeners categorized more stimuli on the f/s continuum depending on their previous training, but only when the same voice was used during both training and testing. Thus, listeners attended to talker-specific knowledge when categorizing ambiguous stimuli. In a similar experiment with ambiguous [s] and [ʃ] stimuli, Kraljic and Samuel showed that perceptual learning of talker-specific characteristics is retained up to at least 25 minutes.

What kind of perceptual knowledge do listeners acquire when they learn to identify a speaker's voice? One possibility is that the perceptual operations (Kolers, 1973) used to recognize voices become part of "procedural memory" and are activated when the same voice is encountered again in a subsequent intelligibility test. This kind of procedural knowledge might increase the efficiency of the perceptual analysis of 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 and Brooks, 1984).

Whatever the exact nature of this perceptual 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 individual items and the specific task that was employed to familiarize the listeners with the voices (Glanzer et al., 2004; Johnson et al., 1993; Mitchell and Johnson, 2000; Roediger, 1990; Schacter, 1992). These findings provide additional support for the view that the internal representation of spoken words encompasses a phonological description of the utterance as well as information about the source characteristics of the specific talker. The results of these studies suggest that normal speech perception is carried out in a "talker-contingent" manner; the indexical and linguistic properties of the speech signal are closely coupled in perceptual analysis.

Differences in the processing of detailed voice information and more abstract lexical information can be dissociated by familiarizing listeners with voices speaking in a foreign language. Inspired by previous work which showed that listeners were better able to identify voices speaking in a language familiar to the listeners (e.g. Goggin et al., 1991; Sullivan and Schlichting, 2000; Thompson, 1987), Winters et al. (2006) trained two groups of monolingual English listeners to identify the same ten voices speaking either in English or German. Following four days of training, listeners carried out a generalization task, in which they were asked to identify the same ten voices but in the untrained language (either German or English). Listeners from each group were able to generalize to the untrained language, indicating that the listeners' detailed knowledge of each speaker's voice characteristics is (at least) partially independent of the particular language being spoken. Further, voice information must be at least partially separate from lexical information, since listeners were able to generalize both to and from German, a language for which they had no lexical entries. However, the two groups differed in the degree to which they were able to generalize to the untrained language. Listeners trained in German were able to generalize to English with no loss in voice identification performance, whereas listeners trained in English exhibited a marked decline in their voice identification performance when presented with the same voices in German. This difference in generalization suggests that the listener's encoding of the indexical properties of the speech signal are not entirely dissociated from the listener's linguistic knowledge, and thus that knowledge of the training and testing languages can mediate performance in a voice-identification task.

1.3.3 Summary

The evidence presented in this section is consistent with a view of speech perception and spoken word recognition in which all information in the speech signal is processed and represented. This approach contrasts with the traditional view of speech perception in which a listener is assumed to analyze the speech signal

for its linguistic content, and discard extralinguistic information. This traditional view in which the speech signal is "normalized," and only abstract, symbolic representations are stored, is clearly not sufficient to account for the data presented above. In the next section, we will discuss the extreme position which states that abstract, symbolic representations are not necessary at all for language processing, a possibility which we ultimately reject in favor of a hybrid view in which listeners store the instances they encounter as well as abstractions over those instances.

1.4 The end of abstract representations?

A more radical approach to cognition in which there are no internal representations of the external world has been proposed recently by a group of artificial intelligence (AI) researchers working on behavior-based autonomous robotics and biological intelligence (Beer, 2000; Brooks, 1991a; 1991b; Clark, 1999). According to this perspective, called "embodied cognition," mind, body, and world are linked together as a "coupled" dynamical system (Beer, 2000; Clark, 1999); internal mental representations and information processing are not needed to link perception and action directly in real-world tasks, such as navigating 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; 1991b). Intelligent adaptive behavior reflects the operation of the whole system working in synchrony, without a central executive guiding behavior based on internal models of the world.

Although most research on embodied cognition has come from AI and is related to constructing autonomous robots and establishing links between perception and action in simple sensory-motor systems, the arguments against the necessity of abstract, symbolic representations and the mainstream symbol-processing views of cognition and intelligence have raised a number of issues that are directly relevant to current theoretical debates throughout cognitive science. With regard to representations in speech perception and spoken word recognition, these issues are concerned directly with questions about "representational specificity" and the necessity of lexical representations typically assumed to be active in spoken word recognition and comprehension. A strong non-representational view of spoken language has been proposed recently by Port and Leary (2005), who argued that discrete representations are not needed for real-time human speech perception.

Although the non-representational theorists have argued that it is not necessary to posit mediating states corresponding to internal representations of the external world, there are several reasons to believe that their global criticisms of the traditional symbol-processing approach to cognition may not generalize gracefully to more complex knowledge-based cognitive domains (Markman and 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 share 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.

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 non-representationalists would seriously argue that speech and spoken language is non-representational or non-symbolic in nature. 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 of a symbol-processing system. Indeed, this is one of the major "design features" of human language (Hockett, 1960). Evidence for symbolic representations comes from myriad sources of language data. Here, we briefly discuss two types of evidence that reveal the existence of discrete representations of sound structure in language.

The first general line of evidence we offer in favor of discrete representations of sound structure comes from linguistics. Indeed, one of the

fundamental assumptions within the generative linguistics tradition (e.g. Chomsky and Halle, 1968; Prince and Smolensky, 1993/2004) is that the continuous acoustic wave form is represented by speakers at various "grain" sizes, such as phonological features (subsegmental structure), phonemes (segmental structure), and syllabic and metrical structure (suprasegmental structure). These assumptions have proven quite useful in accounting for language-internal and cross-linguistic phonological patterns. For example, segments are composed of bundles of features, and these features are used to define natural classes of segments (fricatives, stops, etc.). It has been argued that sound change—both synchronic and diachronic—occurs at the level of natural classes. Additionally, although we discussed some criticisms of the traditional view in section 1.2.2 in which the only sound structure representations are discrete idealized symbols, there are certain phonological phenomena in human languages in which it appears that segments are discrete and psychologically real entities (or symbols) which may be individually manipulated in language use. One phonological phenomenon which reveals the psychological reality of the segmental level of representation is metathesis, in which adjacent segments are transposed to create a new sound structure sequence, as in the dialectal example *ask* → [æks]. This resequencing of the /s/ and /k/ critically requires that these sound structure elements are represented—at some point in the processing system—as abstract symbols in a string which can be reordered. The reader is referred to Elizabeth Hume's metathesis database (Hume, 2000) for a wide variety of metathesis examples across the world's languages.

An additional source of evidence suggesting that there is a level of discrete sound structure representation comes from studies of speech errors. For example, Nooteboom (1969) analyzed a corpus of speech errors in Dutch, and found that 89 percent of the errors involved a single segment (also see Jaeger, 1992; 2005 for similar data in children). An additional piece of evidence comes from the "repeated phoneme effect" (Dell, 1984; MacKay, 1970; Nooteboom, 1969), in which errors are more likely in sequences containing a repeated phoneme (e.g. the vowel in *time line*) than sequences without repeated phonemes (e.g. *heat pad*), indicating that the language processing system represents sublexical units. This result has been observed in both spontaneous speech errors and experimentally induced errors (Dell, 1984; 1986). Additionally, Stemberger (1990) reported that while repetition of identical segments increases the rate of speech errors, repetition of featurally similar segments does not. In addition to these speech production errors, it has been reported that a large number of misperceptions in fluent speech involve segments rather than syllables or words (Bond, 2005; Bond and Garnes, 1980; Bond and Robey, 1983).

It is worth noting that most speech error studies have analyzed speech transcribed into strings of phonemes, which leaves open the possibility that the segmental errors reported in these studies are an artifact of the methodology. In addition, there are several articulatory and acoustic studies which have provided evidence that certain speech errors typically thought to involve discrete insertions or substitutions actually result from gradient errors in production (e.g. overlapping gestures; see Pouplier, 2003). However, a recent articulatory study with an aphasic speaker indicates that discrete vowel segments can be inserted to "repair" problematic sound structure sequences. Buchwald (2005) reported on an aphasic English speaker (VBR) whose deficit leads her to insert a vowel in word-initial consonant clusters (e.g. *bleed* → [bəlid]). VBR's articulations were recorded with ultrasound imaging while she produced consonant cluster words (e.g. *bleed*) and words with schwa between the same two consonants (e.g. *believe*). The articulatory and acoustic data indicated that her productions of words with inserted schwa were identical to her productions of words with lexical schwa, and thus inconsistent with several gradient accounts of vowel insertion based on changes along the temporal dimension. This result indicates the existence of discrete, manipulable vowel units which may be inserted in the case of aphasic speech errors.

In our view, the current debate that emerges from the criticisms of traditional, symbolic representations is not about whether spoken language processing is strictly a symbol processing system. In the case of sound structure, the evidence is clear: we encode the instances we encounter, and form abstractions such as segmental representations. The principal theoretical issue revolves around a precise description of the exact nature of the phonetic, phonological, and lexical representations used in spoken language processing and the interaction among the abstractions and the encoded exemplars.

In our view, two major questions have emerged. First, how much perceptual detail of the original speech signal is encoded in order to support language processing. Second, how much detail can be later discarded as a consequence of phonological and lexical analysis? The evidence described

in the last two sections suggests that it is unlikely that there is only one basic unit of perception or only one common representational format active in speech perception and spoken word recognition. Rather, there is strong evidence for the existence of multiple units and representations—with different degrees of abstraction—that are used in parallel (see Pisoni and Luce, 1987).

The next section discusses some new directions in linguistic research that may be viewed as attempts to account simultaneously for the encoding of detailed information of perceptual experiences and for abstractions over those experiences.

1.5 Integrating abstractions and exemplars: new views from linguistics

In natural language contexts, one type of evidence that we encode the particular exemplars comes from certain phonological processes that affect words differently depending on their frequency of occurrence. Pierrehumbert (2001), citing Hooper (1976), noted a three-way distinction in the application of schwa lenition (or weakening) among words with word-medial obstruent-liquid clusters based on their relative frequency. For high frequency words, no schwa is present in the acoustic record (e.g. between the [v] and [r] of *every*); for words of low frequency, there is a schwa (e.g. mamm[ə]ry); and for mid-frequency words, there is a syllabic [r] (e.g. *memory*). This example is critical, as it contains a process targeting an abstract phonemic category (words with medial obstruent-liquid clusters), but applying differentially to particular members of that category depending on the number of times they have been encountered. Pierrehumbert proposes a framework in which individual exemplars of each word are stored and form part of the representation of a given lexical item. This framework permits a treatment of these frequency-based lenition effects if we assume that the representation of forms targeted by the lenition process changes at a rate commensurate with the absolute number of times we encounter that form.

Bybee (2005) has also recently suggested that fine phonetic details of specific instances of speech are retained in lexical representations. 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). The notion that acoustic-phonetic variability in speech needs to be captured and represented in some fashion in linguistic representations to reflect actual experience has been taken up by several other proposals in generative linguistics (see Steriade, 2001a; 2001b; papers in Hume and Johnson, 2001).

At this point, most of the proposals incorporating the strengths of exemplar-based accounts and accounts using abstract representations are in the speech production domain. Johnson (1997; 2005) has also proposed a model of speech perception that stores exemplars and therefore does not lose any token-specific details such as information about a talker's voice. While this proposal is consistent with the large body of results discussed in section 1.3.2, it is not at present integrated with a view in which sublexical information—abstracted over the stored exemplars—is represented separately by the language processing system. In short, while there are several exciting and promising new research directions, a full account of the wide body of data discussed in this chapter remains an active area of inquiry.

1.6 Representations and mechanisms in spoken word recognition

The discussion has so far focused on lexical and sublexical representations of speech without addressing the specific processing mechanisms involved in spoken word recognition. This section discusses several mechanisms of spoken word recognition proposed in the literature, and the types of representations associated with these mechanisms.

As discussed earlier, the traditional symbol-processing approach to spoken word recognition 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 words are recognized by recovering and identifying sequences of phonemes from the acoustic-phonetic information present in the speech waveform. If a listener could recognize and recover the phonemes from the speech waveform, she would be successful in perceiving the component words and understanding the talker's intended message (Allen, 2005). As foreshadowed in section 1.2.2, 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 bottom-up, "segmental view" of spoken word recognition was fundamentally transformed by Marslen-Wilson and his colleagues (Marslen-Wilson and Welsh, 1978), who 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 are carried out in language comprehension should be viewed as the functional locus of the interaction 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 the process of recognizing spoken words and understanding sentences. In Cohort Theory, segments and phonemes "emerge" from the process of lexical recognition and selection. 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 NAM has been the use of a simple similarity metric for estimating phonological distances of words using a one-phoneme substitution rule (Greenberg and Jenkins, 1964; Pisoni et al., 1985). This computational method has 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 spoken word recognition assume that speech signals are represented in memory using traditional abstract representational formats consisting of discrete features, allophones, or phonemes. Current models of spoken word recognition also routinely assume that individual words are represented discretely. 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 Goldinger, 1998; Kapatsinski, 2006 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.

An 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 is 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. While this approach successfully incorporates the details of perceptual experiences in the representations and mechanisms of spoken word recognition, it misses the critical generalizations available to the proposals that include abstract lexical and sublexical representations.

1.7 Conclusions

Evidence from a wide variety of studies suggests that highly detailed perceptual traces representing both the "medium" (detailed source information) and the "message" (linguistic content of the utterance) of the speech signal 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. 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. This information becomes 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 exists for both detailed episodic traces and abstract representations of sound structure, 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 stimuli over a wide range of environmental conditions. Listeners adapt very quickly and effortlessly to changes in speaker, dialect, and speaking rate, and are able to adjust rapidly to acoustic degradations 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 in listening conditions, is the major challenge for future research in speech perception and spoken word recognition.

Acknowledgments

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, Steve Winters and especially Adam Buchwald for invaluable discussions and comments on this chapter.

References

Abler, W. L. (1989) On the particulate principle of self-diversifying systems. *Journal of Social Biological Structure*, 12: 113.

Allen, J. B. (1994) How do humans process and recognizer speech? *IEEE Trans. Speech Audio*, 2: 567–77.

Allen, J. B. (2005) *Articulation and intelligibility*. Morgan & Claypool, San Rafael.

Allen, J. S., and Miller, J. L. (2004) Listener sensitivity to individual talker differences in voice-onset-time. *Journal of the Acoustical Society of America*, 115: 3171–83.

Beer, R. D. (2000) Dynamical approaches to cognitive science. *Trends in Cognitive Sciences*, 4: 91–9.

Blesser, B. (1972) Speech perception under conditions of spectral transformations, I: Phonetic characteristics. *Journal of Speech and Hearing Research*, 15: 5–41.

Bond, Z. S. (2005) Slips of the ear. In D. B. Pisoni and R. E. Remez (eds), *The Handbook of Speech Perception*, pp. 290–310. Blackwell, Oxford, UK.

Bond, Z. S., and Garnes, S. (1980) Misperceptions of fluent speech. In R. A. Cole (ed.), *Perception and Production of Fluent Speech*. pp. 115–32. Erlbaum, Hillsdale, NJ.

Bond, Z. S., and 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. 249–83. Academic Press, New York.

Bricker, P. D., and 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 and B. Lloyd (eds), *Cognition and Categorization*, 169–211. Erlbaum, Hillsdale, NJ.

Brooks, R. A. (1991a) New approaches to robotics. *Science*, 253(5025): 1227–32.

Brooks, R. A. (1991b) Intelligence without representation. *Artificial Intelligence*, 47: 139–59.

Buchwald, A. B. (2005) Sound structure representation, repair, and well-formedness: Grammar in spoken language production. Ph.D. dissertation, Johns Hopkins University.

Bybee, J. L. (1998) The emergent lexicon. *Chicago Linguistic Society*, 34: 421–35.

Bybee, J. L. (1999) Usage-based phonology. In M. Darnell, E. Moravcsik, F. Newmeyer, M. Noonan, and K. Wheatley (eds), *Functionalism and Formalism in Linguistics*, vol. II: *Case Studies*, pp. 211–42. 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, CA.

Chomsky, N. and Halle, M. (1968) *The Sound Pattern of English*. Harper & Row, New York.

Chomsky, N. and Miller, G. A. (1963) Introduction to the formal analysis of natural languages. In R. D. Luce, R. Bush, and E. Galanter (eds), *Handbook of Mathematical Psychology*, vol. 2, pp. 269–321. Wiley, 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: 345–51.

Creelman, C. D. (1957) Case of the unknown talker. *Journal of the Acoustical Society of America*, 29: 655.

Dell, G. (1984) Representation of serial order in speech: Evidence from the repeated phoneme effect in speech errors. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 10: 222–33.

Dell, G. (1986) A spreading activation theory of retrieval in sentence processing. *Psychological Review*, 93: 283–321.

Eisner, F., and McQueen, J. M. (2005) The specificity of perceptual learning in speech processing. *Perception and Psychophysics*, 67: 224–38.

Elman, J. L. (2004) An alternative view of the mental lexicon. *TRENDS in Cognitive Sciences*, 8: 301–6.

Elman, J. L., and McClelland, J. L. (1986) Exploiting lawful variability in the speech waveform. In J. S. Perkell and D. H. Klatt (eds), *Invariance and Variability in Speech Processing*, pp. 360–85. Erlbaum, Hillsdale, NJ.

Estes, W. K. (1994) *Classification and Cognition*. Oxford University Press, New York.

Fletcher, H. (1953) *Speech and Hearing in Communication*. Krieger, Huntington, NY.

Fowler, C. A. (1990) Listener–talker attunements in speech. *Haskins Laboratories Status Report on Speech Research*, 101:–2, 110–29.

Gautier, I., and Tarr, M. J. (2002) Unraveling mechanisms for expert object recognition: bridging brain activity and behavior. *Journal of Experimental Psychology: Human Perception and Performance*, 28: 431–46.

Glanzer, M., Hilford, A., and Kim, K. (2004) Six regularities of source recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30: 1176–95.

Gleason, H. A. (1961) *An Introduction to Descriptive Linguistics*. Holt, Rinehart, & Winston, New York.

Goggin, J. P., Thompson, C. P., Strube, G., and Simental, L. R. (1991) The role of language familiarity in voice identification. *Memory and Cognition*, 19: 448–58.

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: 1166–83.

Goldinger, S. D. (1997) Talker variability in speech processing. In K. Johnson and J. W. Mullennix (eds), *Talker Variability in Speech Processing*, pp. 33–66. Academic Press, San Diego.

Goldinger, S. D. (1998) Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, 105: 251–79.

Goldinger, S. D., and Azuma, T. (2003). Puzzle-solving science: the quixotic quest for units in speech perception. *Journal of Phonetics*, 31: 305–20.

Goldinger, S. D., Pisoni, D. B., and 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: 152–62.

Greenberg, J. H., and Jenkins, J. J. (1964) Studies in the psychological correlates of the sound system of American English. *Word*, 20: 157–77.

Halle, M. (1956) Review of *Manual of Phonology* by C. D. Hockett. *Journal of the Acoustical Society of America*, 28: 509–10.

Hockett, C. D. (1960) The origin of speech. *Scientific American*, 203: 88–96.

Hockett, C. F. (1955) *Manual of Phonology*. Indiana University, Bloomington.

Hooper, J. D. (1976) Word frequency in lexical diffusion and the source of morphophonological change. In W. Christie (ed.), *Current Progress in Historical Linguistics*, pp. 96–105. North-Holland, Amsterdam.

Hume, E. (2000) http://www.ling.ohio-state.edu/~ehume/metathesis/index.html)

Hume, E., and Johnson, K. (eds) (2001) *The Role of Speech Perception in Phonology*. Academic Press, San Diego.

Jacoby, L. L., and Brooks, L. R. (1984) Non-analytic cognition: memory, perception, and concept learning. In G. Bower (ed.), *The Psychology of Learning and Motivation*, pp. 1–47. Academic Press, New York.

Jaeger, J. (1992) Phonetic features in young children's slips of the tongue. *Language and Speech*, 35: 189–205.

Jaeger, J. J. (2005) *Kids' Slips: What Young Children's Slips of the Tongue Reveal about Language Development*. Erlbaum, Mahwah, NJ.

Jenkins, J. J., Strange, W., and 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 and J. W. Mullennix (eds), *Talker Variability in Speech Processing*, pp. 145–66. Academic Press, San Diego, CA.

Johnson, K. (2005) Resonance in an exemplar-based lexicon: the emergence of social identity and phonology. *UC Berkeley Phonology Lab Annual Report*, 95–128.

Johnson, M. K., Hashtroudi, S., and Lindsay, D. S. (1993) Source monitoring. *Psychological Bulletin*, 114: 3–28.

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, and Y. Sagisaka (eds), *Speech Perception, Production, and Linguistic Structure*, pp. 135–42. Ohmsha, Tokyo.

Kapatsinski, V. M. (2006) Towards a single-mechanism account of frequency effects. *Proceedings of LACUS 32: Networks*, pp. 325–35. Hanover, NH.

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

Klatt, D. H. (1986) The problem of variability in speech recognition and in models of speech perception. In J. S. Perkell and D. H. Klatt (eds), *Invariance and Variability in Speech Processing*, pp. 300–19. Erlbaum, Hillsdale, NJ.

Kolers, P. A. (1973) Remembering operations. *Memory and Cognition*, 1: 347–55.

Kolers, P. A. (1976) Pattern-analyzing memory. *Science*, 191: 1280–81.

Krajlic, T., and Samuel, A. G. (2005) Perceptual learning for speech: is there a return to normal? *Cognitive Psychology*, 51: 141–78.

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., and Studdert-Kennedy, M. (1967) Perception of the speech code. *Psychological Review*, 74: 431–61.

Liberman, A. M., Delattre, P. C., Cooper, F. S., and Gerstman, L. J. (1954) The role of consonant–vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs*, 68: 1–13.

Licklider, J. C. R. (1952) On the process of speech perception. *Journal of the Acoustical Society of America*, 24: 590–94.

Lindgren, N. (1965) Machine recognition of human language. *IEEE Spectrum*, Mar. and Apr.

Luce, P. A., and McLennan, C. T. (2005) Spoken word recognition: the challenge of variation. In D. B. Pisoni and R. E. Remez (eds), *The Handbook of Speech Perception*, pp. 591–609. Blackwell, Oxford.

Luce, P. A., and Pisoni, D. B. (1998) Recognizing spoken words: the Neighborhood Activation Model. *Ear and Hearing*, 19: 1–36.

MacKay, D. G. (1970) Spoonerisms: the structure of errors in the serial order of speech. *Neuropsychologia*, 8: 323–50.

Markman, A. B., and Dietrich, E. (2000). Extending the classical view of representation. *Trends in Cognitive Sciences* 4: 470–75.

Marslen-Wilson, W. D. (1984) Function and process in spoken word recognition: a tutorial review. In H. Bouma and D. G. Bouwhis (eds), *Attention and Performance X: Control of Language Processes*, pp. 125–50. Erlbaum, Hillsdale, NJ.

Marslen-Wilson, W. D., and 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., and Summers, W. V. (1989) Effects of talker variability on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 15: 676–84.

Miller, G. A., Heise, G. A., and Lichten, W. (1951) The intelligibility of speech as a function of the context of the test material. *Journal of Experimental Psychology*, 41: 329–35.

Mitchell, K. J., and Johnson, M. K. (2000) Source monitoring: attributing mental experiences. In E. Tulving and F. I. M. Craik (eds), *The Oxford Handbook of Memory*, pp. 179–85. Oxford University Press, New York.

Morton, J. (1979) Word recognition. In J. Morton and J. C. Marshall (eds), *Structures and Processes*, pp. 108–56. MIT Press, Cambridge, MA.

Mullennix, J. W., and Pisoni, D. B. (1990) Stimulus variability and processing dependencies in speech perception. *Perception and Psychophysics*, 47: 379–90.

Mullennix J. W., Pisoni, D. B., and Martin, C. S. (1989) Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America*, 85: 365–78.

Murphy, G. L. (2002) *The Big Book of Concepts*. MIT Press, Cambridge, MA.

Nooteboom, S. G. (1969) The tongue slips into patterns. In A. A. van Raad (ed.), *Leyden Studies in Linguistics and Phonetics*, pp. 114–32. Mouton, The Hague.

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., and 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. 209–12.

Nygaard, L. C., Sommers, M. S., and Pisoni, D. B. (1994) Speech perception as a talker-contingent process. *Psychological Science*, 5: 42–6.

Oldfield, R. C. (1966) Things, words and the brain. *Quarterly Journal of Experimental Psychology*, 18: 340–53.

Palmeri, T. J., Goldinger, S. D., and Pisoni, D. B. (1993) Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19: 309–28.

Pardo, J. S., and Remez, R. E. (2006) The perception of speech. In M. Traxler and 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, 629–37.

Peterson, G. E., and Barney, H. L. (1952) Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, 24: 175–84.

Pierrehumbert, J. B. (2001) Exemplar dynamics: word frequency, lenition and contrast. In J. Bybee and P. Hopper (eds), *Frequency and the Emergence of Linguistic Structure*, pp. 137–58. Benjamins, Amsterdam.

Pierrehumbert, J. B., and Pierrehumbert, R. T. (1990) On attributing grammars to dynamical systems. *Journal of Phonetics*, 18: 465–77.

Pisoni, D. B. (1993) Long-term memory in speech perception: some new findings on talker variability, speaking rate and perceptual learning. *Speech Communication*, 13: 109–25.

Pisoni, D. B. (1997) Some thoughts on "normalization" in speech perception. In K. Johnson and J. W. Mullennix (eds), *Talker Variability in Speech Processing*, pp. 9–32. Academic Press, San Diego.

Pisoni, D. B., and Luce, P. A. (1987) Acoustic-phonetic representations in word recognition. *Cognition*, 25: 21–52.

Pisoni, D. B., Nusbaum, H. C., Luce, P. A., and Slowiaczek, L. M. (1985) Speech perception, word recognition and the structure of the lexicon. *Speech Communication*, 4: 75–95.

Pisoni, D. B., and Remez, R. E. (eds) (2005) *The Handbook of Speech Perception*. Blackwell, Malden, MA.

Pollack, I. (1952) The information of elementary auditory displays. *Journal of the Acoustical Society of America*, 24: 745–9.

Pollack, I. (1953) The information of elementary auditory displays II. *Journal of the Acoustical Society of America*, 25: 765–9.

Port, R., and Leary, A. (2005) Against formal phonology. *Language*, 81: 927–64.

Pouplier, M. (2003. Units of phonological encoding: empirical evidence. Ph.D. dissertation, Yale University.

Prince, A., and Smolensky, P. (1993/2004) Optimality Theory: constraint interaction in generative grammar (technical report). Rutgers University, New Brunswick and University of Colorado, Boulder.

Raphael, L. J. (2005) Acoustic cues to the perception of segmental phonemes. In D. B. Pisoni and R. E. Remez (eds), *The Handbook of Speech Perception*, pp. 182–206. Blackwell, Oxford.

Reddy, R. D. (1975) *Speech Recognition.* Academic Press, New York.

Remez, R. E., Rubin, P. E., Pisoni, D. B., and Carrell, T. D. (1981) Speech perception without traditional speech cues. *Science*, 212(4497): 947–50.

Rhodes, G., Byatt, G., Michie, P. T., and Puce, A. (2004) Is the fusiform face area specialized for faces, individuation, or expert individuation? *Journal of Cognitive Neuroscience*, 16: 189–203.

Roediger, H. L. (1990) Implicit memory: retention without remembering. *American Psychologist*, 45: 1043–56.

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, 543–71.

Schacter, D. L. (1992) Understanding implicit memory: a cognitive neuroscience approach. *American Psychologist*, 47: 559–69.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (1995) Speech recognition with primarily temporal cues. *Science*, 270(5234): 303–4.

Smith, E. E., and Medin, D. (1981) *Categories and Concepts.* Harvard University Press, Cambridge, MA.

Sommers, M. S., Nygaard, L. C., and 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. 217–20.

Stemberger, J. P. (1990). Wordshape errors in language production. *Cognition*, 35: 123–57.

Steriade, D. (2001a) Directional asymmetries in place assimilation: a perceptual account. In E. Hume and K. Johnson (eds), *The Role of Speech Perception in Phonology*, pp. 219–50, Academic Press, San Diego.

Steriade, D. (2001b) The phonology of Perceptibility Effects: the P-map and its consequences for constraint organization. MS, UCLA.

Sullivan, K. P. H., and Schlichting, F. (2000) Speaker discrimination in a foreign language: first language environment, second language learners. *Forensic Linguistics*, 7, 95–111.

Thompson, C. P. (1987) A language effect in voice identification. *Applied Cognitive Psychology*, 1: 121–31.

Treisman, M. (1978a) A theory of the identification of complex stimuli with an application to word recognition. *Psychological Review*, 78, 420–25.

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., and Schacter, D. L. (1990). Priming and human memory systems. *Science*, 247, 301–6.

Twaddell, W. F. (1952) Phonemes and allophones in speech analysis. *Journal of the Acoustical Society of America*, 24: 607–11.

Winters, S. J., Levi, S. V., and Pisoni, D. B. (2006). The role of linguistic competence in cross-linguistic speaker identification. Talk presented at the 80th annual meeting of the Linguistics Society of America, Albuquerque, NM.