Elman, J., & McClelland, J. (1986). Exploiting lawful variability in the speech wave. In J. S. Perkell & D. H. Klatt, (Eds.), *Invariance and variability in speech processing* (pp. 360–380). Hillsdale, NJ: Erlbaum.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics, 14,* 3–28.

Hintzman, D. L. (1986). Schema abstraction in a multiple-trace memory model. *Psychological Review, 93,* 411–428.

Joos, M. (1948). Acoustic phonetics. *Language,* V *24* (Suppl. 2), 1–136.

Liberman, A. M. (1957). Some results of research on speech perception. *Journal of the Acoustical Society of America, 29,* 117–123.

Liberman, A., & Mattingly, I. G. (1985). The motor theory revised. *Cognition, 21,* 1–36.

Potter, R., & Steinberg, J. (1950). Toward the specification of speech. *Journal of the Acoustical Society of America, 22,* 807–820.

Stevens, K. N., & Blumstein, S. E. (1978). Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America, 64,* 1358–1368.

# 2

# Some Thoughts on "Normalization" in Speech Perception

**DAVID B. PISONI**

## 2.1 INTRODUCTION

The theoretical problems confronting researchers working in the field of speech perception are, in principle, no different from the problems studied in other areas of perception and cognition. Basically, they involve issues of invariance and variability of the speech signal, the neural representation of speech in the auditory system, and the perceptual constancy maintained by human listeners in the face of diverse physical stimulation. These are well-known problems in speech perception that have occupied psychologists, linguists, and engineers for close to a half-century, since the beginning of modern speech research in the late 1940s (see Klatt, 1989, for a review).

When compared to other perceptual systems, there can be little doubt that speech perception is extremely robust and adaptive over a wide range of environmental conditions that introduce large physical changes and transformations in the acoustic signal. For example, normal hearing listeners can adapt easily to changes in speakers, dialects, speaking rate, and speaking style, as well as a wide variety of acoustic transformations, including the presence of noise, reverberation, and the use of different transducers without any noticeable decrease in performance. Even the most sophisticated state-of-the-art speech-recognition systems cannot compare to the speed and efficiency of the human listener. It has always

seemed that one of the principle tasks of speech science was to explain these remarkable perceptual and cognitive abilities and to model how the nervous system accomplishes this task so quickly and effortlessly in the face of continuous changes in the listener's environment.

For the last several years, I have been interested in the problems of stimulus variability in speech, in particular, the effects of stimulus variability from different talkers and different speaking rates on word-recognition performance. Recent findings have suggested to me that some of the long-standing theoretical assumptions that speech researchers have held about the existence of abstract units such as phonemes and words need to be reexamined and substantially revised. More specifically, the assumption of an idealized symbolic representation for spoken language has encouraged researchers to search for simple first-order physical invariants and to ignore the problem of stimulus variability in the listener's environment. Variability is simply treated as a troublesome source of "noise" in the acoustic signal. Following recent accounts of memory and concept learning by Jacoby and Brooks (1984), we will call this the traditional "abstractionist" or "analytic" approach to speech perception, an approach that places primary emphasis on the search for idealized categories that encode the linguistic content of speech into abstract symbolic units.

Findings from our laboratory on stimulus variability have suggested an alternative view of speech perception that is compatible with a large and growing body of literature in the field of cognitive psychology that deals with categorization. This view of speech perception focuses on the encoding of specific instances and assumes that very detailed stimulus information in the speech signal is processed by the listener and becomes part of the memory representation for spoken language. One of the assumptions of this "nonanalytic" approach is that stimulus variability is, in fact, a lawful and highly informative source of information for the perceptual process; it is not simply a source of noise that masks or degrades the idealized symbolic representation of speech in human long-term memory. According to this view, listeners encode particulars rather than generalities. Our research has shown that source information about the talker's voice and detailed information about speaking rate is, in fact, encoded into memory and forms part of the neural representation of speech. Rather than discarding the "indexical" attributes of speech in favor of a highly abstract symbolic code like a string of segments or phonemes, the human perceptual and memory systems appear to encode and retain very fine details of the perceptual event. Our results have a number of implications for future research on spoken-language processing as we continue to gain new insights into the "nonanalytic" aspects of speech and to reconsider the long-standing problems of invariance and variability in speech perception.

In this chapter, I consider the problem of perceptual normalization, specifically, talker normalization, in light of some recent findings and new theoretical developments in the field of perceptual learning and categorization. In past theoretical accounts of speech perception, a strict dissociation has been made between



the linguistic properties of speech that carry the speaker's intended message and the indexical features of the signal that provide information about the talker's voice (Studdert-Kennedy, 1974). The dissociation between the form and content of the speech signal has a long history in the fields of linguistics and phonetics, which has been carried over to theoretical accounts of speech perception despite the obvious fact that both sources of information are carried simultaneously and in parallel by the same acoustic signal. An excellent example of the functional parallelism of the indexical and linguistic properties of speech can be seen in Figure 1, which shows the encoding of speech in the auditory periphery (Hirahara & Kato, 1992). The absolute frequencies of the formants provide cues to speaker identification (A), whereas the relative differences among the formants specify information for vowel identification (B). Both sets of attributes are carried simultaneously by the same acoustic signal.

Our recent findings suggest that the indexical attributes of a talker's voice are perceived and encoded in memory by the perceptual system along with the linguistic message, and that information about the talker's voice is not lost or discarded as a consequence of perceptual analysis. A variety of behavioral studies involving perceptual identification, speeded classification, and recognition memory tasks have demonstrated that very detailed information about the talker's voice and the specific stimulus tokens is encoded and retained in memory and subsequently affects spoken-word-recognition performance. Other studies have shown that listeners acquire information about unfamiliar voices that improves the intelligibility of novel words and sentences from these same talkers.

These results imply that the indexical and linguistic attributes of speech are not neatly partitioned into two independent channels of information by the nervous system. Indeed, the mutual dependencies observed in perceptual analysis between these two sources of information suggest very close interactions between the form and content of the linguistic message and the listener's linguistic knowledge. Thus, what a listener learns about a talker's voice—the acoustic correlates of gender, dialect, speaking rate, and so forth—are encoded and subsequently used to facilitate a phonetic interpretation of the linguistic content of the message.

These recent findings raise fundamental questions about the nature of the traditional symbolic representations that have been routinely assumed in previous accounts of speech perception. Current views about the neural representation of speech and the basic perceptual units will need to be revised and substantially enriched to accommodate the additional fine details of stimulus encoding that listeners carry out in speech perception and spoken-language processing tasks in the laboratory and real world. I believe these new perceptual findings are important because they suggest an alternative approach to traditional accounts of speech perception and spoken word recognition, which have relied heavily on abstract, idealized symbolic representations of the linguistic message with little or no concern about the contribution of the talker's voice and the indexical features of speech to the perceptual process and the neural representation of speech in memory. The findings from these recent studies also raise several interesting questions

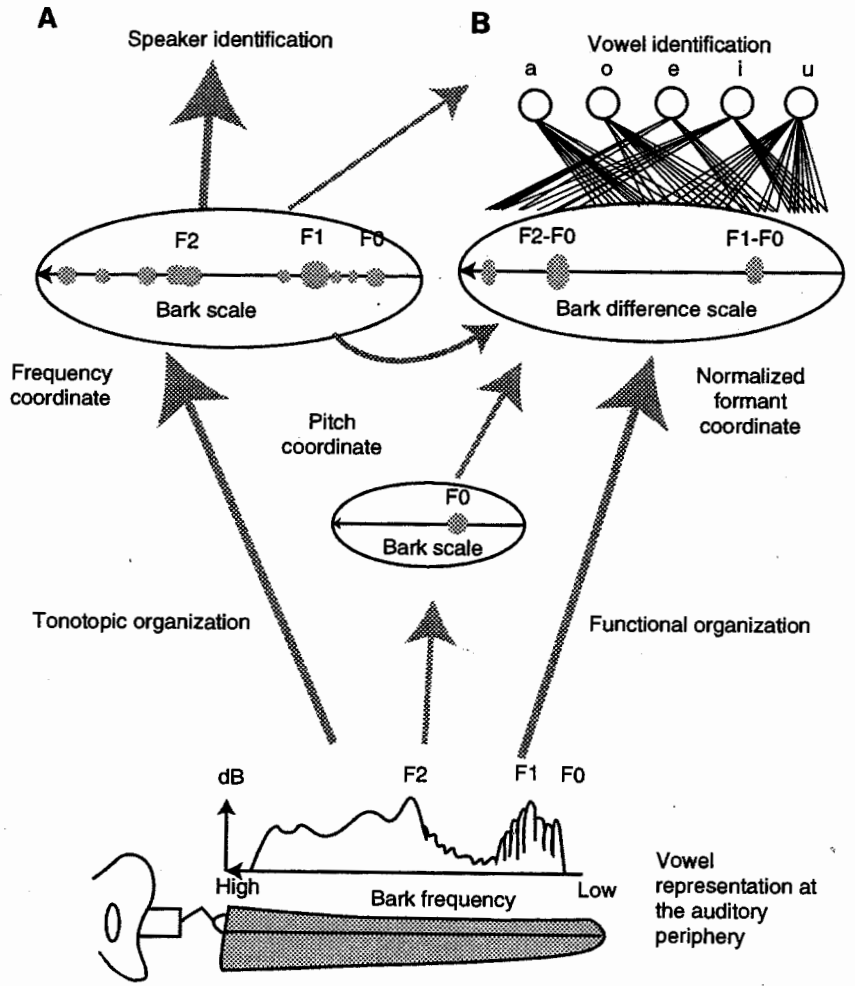


**FIGURE 1** Vowel projections at the auditory periphery show that information for speaker identification (A) and perception of vowel quality (B) is carried simultaneously and in parallel by the same signal. The tonotopic organization of frequencies using a bark scale provides cues to speaker identification, whereas the relations among the formant (F) patterns in terms of difference from F0 in barks provide cues to vowel identification. (From Hirahara & Kato, 1992.)

about the theoretical assumptions that have been made over the years concerning the process of normalization in speech perception.

This chapter is divided into three sections. First, I begin by reviewing several definitions of the term *normalization* and consider the implications that follow from adopting the traditional meanings of this term. Then I briefly summarize several recent findings on the contribution of the talker's voice to speech perception and spoken word recognition. Finally, I describe an alternative approach to perceptual normalization that can accommodate these new findings within a theoretical framework that emphasizes the role of nonanalytic cognition and the contribution of specific instances to perception, learning, and cognition, especially memory and perceptual learning in speech perception and spoken-word recognition.



My goal in putting these ideas together at this time is to argue that it is possible to approach some of the old traditional problems in speech perception and spoken-word recognition, such as invariance, variability, and the need for perceptual normalization, in novel ways that can lead to new knowledge and insights into the fundamental process of speech perception. I believe this was not possible because of the traditional prevailing metatheory that was adopted by everyone working in the field and because of the experimental methodologies routinely employed in the past to study speech perception. However, this is a new era in the study of brain, behavior, and computation. Recent developments in cognitive and neural science, ecological psychology, and neural modeling, along with the widespread availability of high-speed computing technology, have provided new tools and powerful methods to study speech perception under much more demanding and robust conditions; thus many different sources of variability can be manipulated and studied.

Two key features of the traditional approach to speech perception are quite striking when viewed from the nonanalytic perspective, which emphasizes the role of specific instances to perception and cognition, that I am suggesting here. First is the dissociation, mentioned earlier, between the linguistic and indexical properties of the speech signal. Second is the assumption that stimulus variability is a source of noise and is not informative to the listener. Both of these assumptions have had profound effects on the study of speech perception and have strongly affected the kind of research problems that speech scientists have investigated over the years. These two assumptions have also affected the specific experimental methodologies used to study speech perception. Because stimulus variability was thought to mask or obliterate the underlying idealized symbolic message, factors known to create variability in the speech signal were deliberately reduced or eliminated. These factors were viewed as nuisance variables that needed to be controlled in experiments. For example, the traditional approach to acoustic-phonetic research typically used a small number of talkers, usually only one or two, reading carefully constructed experimental materials in citation format under extremely good recording conditions in the laboratory (Byrd, 1992). Also, each experiment typically addressed only a single specific research issue using a very small sample of highly controlled test materials, such as isolated nonsense syllables, words, or short sentences. In contrast, the current approach to acoustic-phonetic research relies very heavily on the use of large speech databases, such as the TIMIT corpus, which was collected with many different talkers producing speech under a wide variety of different conditions. Moreover, the test materials were specifically constructed to sample a larger and more diverse phonetic space,

permitting computation of a variety of lexical statistics and the development of different types of models that can be used to characterize the structure and organization of lexical patterns in a given language (see Huttenlocher & Zue, 1984; Pisoni, Nusbaum, Luce, & Slowiaczek, 1985; Shipman & Zue, 1982; Zue, 1985).

The consequence of this traditional research strategy in acoustic-phonetics was that little if any effort was made to study the contribution of different sources of variability directly or to try to understand how variability affected speech perception or spoken-word-recognition performance in human listeners. Based on our recent findings, I believe that we have made some progress in understanding the role of stimulus variability in speech and how it affects performance in a variety of tasks with different populations of listeners. Also, a substantial body of new knowledge has been obtained about the interdependencies between the indexical and linguistic attributes of the speech signal. Listeners are sensitive to fine phonetic details in the signal, and they encode and represent acoustic changes in their listening environment that are potentially useful.

## 2.2 WHAT IS MEANT BY NORMALIZATION?

My observations and conclusions about stimulus variability and the encoding of fine phonetic details in speech are not entirely novel and should not be too surprising to people working in the mainstream of speech research. About 15 years ago, Dennis Klatt (1979) made a similar proposal about the need to preserve potentially useful acoustic-phonetic information in the context of his LAFS (Lexical Access From Spectra) model of speech recognition. His proposal was based on the idea that an optimally efficient speech-recognition system should be one that can recover gracefully from incorrect labeling or identification without catastrophic effects on performance. Klatt argued that models of speech perception that assume some form of intermediate level of discrete symbolic representation like phonemes or segments discard potentially useful acoustic-phonetic information that might be necessary if backtracking were required. Without retaining fine stimulus details, it would be impossible, according to Klatt, to recover from an errorful interpretation based on ambiguous or incorrect information in the signal.

Without remembering very much about Klatt's earlier proposal concerning the importance of retaining fine phonetic details, I initially found our results on talker variability troublesome for traditional accounts of speech perception that assumed that the speech signal is quickly encoded into a sequence of discrete abstract symbolic units like phonemes or phonetic segments. Of course, at that time, there were not many other viable alternatives to the traditional view of speech or competing accounts of speech perception. Everyone just assumed that phonemes or segments or some kind of symbolic units were perceived during the process of speech perception and constituted the basic perceptual units employed in accessing words from the mental lexicon. These theoretical views about the nature of speech have been around for a long time and have been prominent in



discussions of speech perception and speech recognition. One of the best examples of the traditional idealized view of speech is expressed in Charles Hockett's well-known description of speech as a sequence of Easter eggs:

> 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)

Although this description of speech may seem a little simplistic, it does illustrate nicely the prevailing theoretical views at the time and the very strong emphasis on abstract symbolic coding of speech into discrete idealized linguistic units. Hockett was in good company, as shown by the approach of two prominent theorists in the field of speech communication, J. C. Licklider and Morris Halle, who both endorse the same set of fundamental assumptions about the discrete symbolic nature of speech:

> 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)

As I went back recently and thought about some of these old views in greater detail, I realized that it might be a useful intellectual exercise to review and reconsider what researchers mean by the term *normalization* in speech perception and what some of the implicit assumptions are that follow from adopting this particular approach to dealing with stimulus variability and perceptual constancy in speech. Before proceeding to this task, however, it should be pointed out here that although the most important and probably the most distinctive property of speech is its inherent physical variability (see Table I), these properties are not what are typically discussed in traditional accounts of speech perception or spoken-word recognition. Instead of encouraging research on the study of variability, most theorists have tended to emphasize the discrete symbolic nature of speech and its linguistic function in conveying meaning. Moreover, there has always been a great deal of interest in figuring out ways to eliminate variability from experiments using speech stimuli rather than studying the problem of variability directly.

**TABLE I** Sources of Variability in Speech Acoustics[a]

| Type of variability | Sources of variability |
|---|---|
| Ambient conditions | Background noise, room reverberation, microphone/telephone characteristics |
| Within-speaker variability | Breathy/creaky voice quality, shifting formants, and fundamental frequencies, changing speaking rate, variable degrees of articulatory undershoot, imperfect repetition across tokens of same gesture. |
| Cross-speaker variability | Differences of dialect, vocal tract length and shapes, detailed articulatory habits |
| Segment realization variability | Coarticulatory changes, articulatory modification due to stress or duration changes, optional deletions/simplifications in fluent speech. |
| Word environment variability in continuous speech | Cross-word-boundary coarticulation, phonetic and phonological recoding of words in sentences, changes in word duration due to syntax, pragmatics |

[a]After Klatt, 1986.

Viewed in this way, the study of variability in speech has historically taken a backseat to the formalist approach to speech derived from linguistic theory, especially the recent views of language promoted by transformational grammar with its emphasis on the linguistic competence of an idealized speaker–hearer as summarized in Chomsky's well-known passages reproduced below:

> Linguistic theory is concerned primarily with an ideal speaker-listener, in a completely homogeneous speech-community, who knows its language perfectly and is unaffected by such grammatically irrelevant conditions as memory limitations, distractions, shifts of attention and interest, and errors (random or characteristic) in applying his knowledge of the language in actual performance.
>
> We thus make a fundamental distinction between competence (the speaker-hearer's knowledge of his language) and performance (the actual use of language in concrete situations). Only under the idealization set forth in the preceding paragraph is performance a direct reflection of competence. (1965, p. 2–3)

To help in this exercise and to find a couple of good working definitions of what "normalization" means, I turned first to *Webster's New Twentieth Century Dictionary* (1983). I also consulted English and English's *Dictionary of Psychological Terms* (1958) and Reber's *Dictionary of Psychology* (1985) for definitions of several other related terms used in perception and cognition. The dictionary definitions of the words "normal," "normalization," and "normalize" are given below:

> **normal:** (adj) Conforming with or constituting an accepted standard, model, or pattern; especially, corresponding to the median or average of a large group in type, appearance, achievement, function, development, etc.; natural; standard; regular. (*Webster's*)



> **normal:** (noun) The usual state, amount, degree, etc.; especially, the median or average. (*Webster's*)
>
> **normalization:** (noun) Reduction to a normal or standard state. (*Webster's*)
>
> **normalize:** (vt) To bring into conformity with a standard, pattern, model, etc. (*Webster's*)

The commonality among all three terms shown above is the proposal that some object, concept, or idea is made to conform to a "standard" or a "model" in order to produce or create equivalent forms from diverse inputs. Although not stated explicitly here, there is the further assumption of some loss or reduction of information after normalization is completed.

The definition of the closely related word "standardize," also from *Webster's* is given below:

> **standardize:** (vt) To cause to conform to a given standard; to make standard or uniform; to cause to be without variations or irregularities. (*Webster's*)

Here we see an additional property regarding the elimination of variability added to the previous definitions of "normalization" that require conformity with a standard, pattern, or model. Thus, at least according to the dictionary definitions considered here, the process of normalization consists of at least two separate components. The first deals with bringing something into conformity with a preexisting standard so as to produce equivalent forms, whereas the second has to do with the reduction or elimination of variability thereby creating a standard or normal state.

I believe these definitions of the word *normalization* are fairly good approximations of how this term has been used technically in speech perception over the years to deal with the problems of perceptual invariance and perceptual constancy to nonequivalent forms. Indeed, the intended meanings are consistent with the definitions of the terms *invariance* and *equivalence* given below by Reber:

> **invariance:** (noun) Generally, characteristic of that which does not change. The term is most often used with the qualifier relative. That is, few things in this world are truly invariant but some display greater invariance, greater consistency from circumstance to circumstance than others. In general, in the study of perception and learning, those aspects of the stimulus world that display the higher invariances, relative to other aspects, are learned most quickly and easily.
>
> **equivalence:** (noun) In general, any relationship between two "things" such that one may be substituted for the other in a particular setting and not alter significantly the situation. The term is often modified so that the particular form of equivalence is specified; e.g., stimulus equivalence refers to two or more stimuli that are sufficiently similar that they evoke the same or nearly the same response, response equivalence refers to similar responses made to similar stimuli, etc.

I have gone through the effort of consulting several dictionaries to make explicit several aspects of the meaning of the term normalization that most speech researchers assume more or less implicitly in their experimental and theoretical work. In considering these definitions, I believe it is reasonable to conclude that the process of "normalization" produces three consequences in perception. First,

there is the generation of equivalent forms from diverse inputs. Normalization is assumed to convert physically different tokens into some common representational format and to store these "standardized" representations in some kind of memory. Second, the process of normalization entails a loss of information and, as a consequence, a reduction in stimulus variability. Finally, although not explicitly stated, I believe that normalization also entails the additional assumption that stimulus variability is an undesirable source of noise in the speech signal, which produces perturbations on abstract underlying idealized forms that are assumed to be the true objects of perceptual analysis.

The process of perceptual normalization is, of course, a key component of traditional abstractionist theories of perception, memory, and learning (Shankweiler, Strange, & Verbrugge, 1977). According to this approach, the stimulus environment is highly impoverished and, as a consequence, the perceiver must rely on top-down knowledge to recognize the intended perceptual object and recover the underlying abstract symbolic form. In the case of speech perception, these objects are the talker's intended linguistic message. Other accounts assume these objects are the talker's intended gestures (Fowler & Rosenblum, 1991; Liberman & Mattingly, 1985). In addition to these two assumptions, there is also the proposal that information processing is selective and results in substantial stimulus reduction because the nervous system cannot maintain all aspects of stimulation in the perceiver's environment. A good summary of the traditional information-processing view of perception is given by Haber (1969) below:

> The second assumption, regarding limited information-handling capacities, is also an important one. The problem of limited channel capacity has been clear in the study of perception for most of the history of experimental psychology; witness concepts such as selective attention, and immediate memory span. The nervous system is apparently just not large enough to maintain all aspects of stimulation permanently. What this suggests for information-processing analyses is that we should look for instances in which recoding of information takes place—recoding generally in such a way that some of the content is maintained more explicitly at the expense of the other aspects which are dropped out. The points in time where the recoding occurs should be particularly important ones in the study of information processing, and it is not surprising that most information-processing models refer to these points almost exclusively. (p. 4)

## 2.3 STIMULUS VARIABILITY IN SPEECH PERCEPTION

My colleagues and I have completed a number of experiments on the effects of different sources of variability in speech perception and spoken-word recognition (Pisoni, 1990). Instead of reducing or eliminating variability in the stimulus materials, as most speech researchers have routinely done, we deliberately introduced variability from different talkers to study the effects of these variables on perception (Pisoni, 1992). Our research on this problem began with the 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 same words were produced by 15 different talkers, including male and female voices. Subjects were asked to identify the words using an open-set test format. No feedback was provided. The results were very clear. Across three different signal-to-noise ratios (SNRs), identification performance was always better for words that were produced by a single talker than for words produced by multiple talkers. Trial-to-trial variability in the speaker's voice affected recognition performance. These findings replicated results reported many years ago by Peters (1955) and Creelman (1957) and suggested that the perceptual system must engage in some form of adjustment or "recalibration" each time a novel voice is encountered during the set of trials using multiple voices.

In a second experiment, my colleagues and I measured naming latencies to the same words presented in both blocked (single-talker) and mixed (multiple-talker) test conditions (Mullennix et al., 1989). We found that subjects were not only *slower* to name words presented in multiple-talker lists but they produced *more errors* when their performance was compared to the same words from single-talker lists. Both sets of findings were surprising at the time because all the test words used in the experiment were highly intelligible when presented in isolation under quiet listening conditions. The intelligibility and naming data from these studies raised a number of additional questions about how the various perceptual dimensions of the speech signal are processed and encoded by the human listener. At the time, we naturally assumed that the acoustic attributes used to perceive voice quality were independent of the more abstract linguistic properties of the signal. However, no one had ever tested this assumption directly.

To assess whether attributes of a talker's voice were perceived independently of the phonetic form of the words, we used a speeded classification task (Mullennix & Pisoni, 1990). Subjects were required to attend selectively to one stimulus dimension (e.g., voice) while simultaneously ignoring another stimulus dimension (e.g., phoneme). Across all conditions, we 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, and subjects could not selectively ignore irrelevant variation on the nonattended dimension. If both perceptual dimensions were processed separately, as we originally assumed, we should have found little, if any, interference from the nonattended dimension. Not only did we find mutual interference, suggesting that the two dimensions, voice and phoneme, 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 to ignore the voice dimension when they had to classify the phonemes.

The results from these perceptual experiments were surprising given our 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 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 also encoded into long-term memory. 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 another set of experiments, Goldinger (1992) found evidence that implicit memory for attributes of a talker's voice persists for a relatively long time after perceptual analysis has been completed. He also showed that the degree of perceptual similarity between voices affects the magnitude of the repetition effect in several implicit memory tasks. For example, he found that subjects identified spoken words more accurately when the words were repeated using the same voice in which they had originally been presented than when the words were repeated in a different voice. These findings suggest that the perceptual system encodes very detailed talker-specific information about spoken words in episodic memory representations.

Another series of experiments examined the effects of speaking rate on perception and memory. These studies, designed to parallel the earlier experiments on talker variability, have also shown that the perceptual details associated with differences in speaking rate are not lost as a result of perceptual analysis. In one experiment, Sommers, Nygaard, and Pisoni (1994) found that words produced at different speaking rates (i.e., fast, medium, and slow) were identified more poorly than the same words produced at only one 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 level.

Other experiments on serial recall have also examined the encoding and representation of speaking rate in memory. Nygaard, Sommers, and Pisoni (1995) found that subjects recall 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. Differences in speaking rate, like those observed for talker variability in our earlier experiments, suggest that perceptual encoding and rehearsal processes, which are typically thought to operate on only abstract symbolic representations, are also influenced by instance-specific sources of variability. If these sources of variability were somehow "filtered out" or "normalized" by the perceptual system at relatively early stages of analysis, differences in recall performance would not be expected in memory tasks like the ones used in these experiments.



Taken together with the earlier results on talker variability, the findings on speaking rate suggest that details of the early perceptual analysis of spoken words are not lost. Instead, they become 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 (Goldinger et al., 1991). Listeners encode speech signals in multiple ways along many perceptual dimensions, and the nervous system apparently preserves these perceptual details much more reliably than researchers have believed in the past.

Considering the overall pattern of results, our findings on the effects of talker variability in perception and memory tasks provide support for the proposal that detailed perceptual information about a talker's voice is preserved, and that detailed attributes of the listener's stimulus environment are encoded implicitly into long-term memory. At present, it is not clear whether there is a single "composite" representation in memory or whether these different attributes are encoded in parallel in separate representations (Eich, 1982; Hintzman, 1986). Also unclear is whether spoken words are encoded and represented in memory as a sequence of abstract symbolic phoneme-like units along with much more detailed episodic information about specific instances and the processing operations used in perceptual analysis. These are important questions for future research on the representation of speech in memory.

## 2.4 PERCEPTUAL LEARNING OF VOICES

Findings on talker variability have also encouraged my colleagues and me to examine the 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 received little attention in the speech perception literature despite the obvious relevance to problems of speaker normalization, acoustic-phonetic invariance, and the potential application to automatic speech recognition and speaker identification (Fowler, 1992; Kakehi, 1992). Our search of the research literature on talker adaptation revealed only a small number of behavioral studies with human listeners on this topic, and all of them appeared in obscure technical reports from the mid 1950s.

To determine how familiarity with a talker's voice affects the perception of spoken words, we had two groups of listeners learn to explicitly identify a set of unfamiliar voices over a 9-day period using common names (e.g., Bill, Joe, Sue, Mary). After the subjects learned to recognize the voices, we presented them with a set of novel words mixed in noise at several SNRs; one group heard the words produced by previously heard talkers, whereas the other group heard the same words produced by new talkers to whom they had not been exposed during the perceptual learning task. In this phase of the experiment, which was designed to measure speech intelligibility, subjects were required to identify the words rather

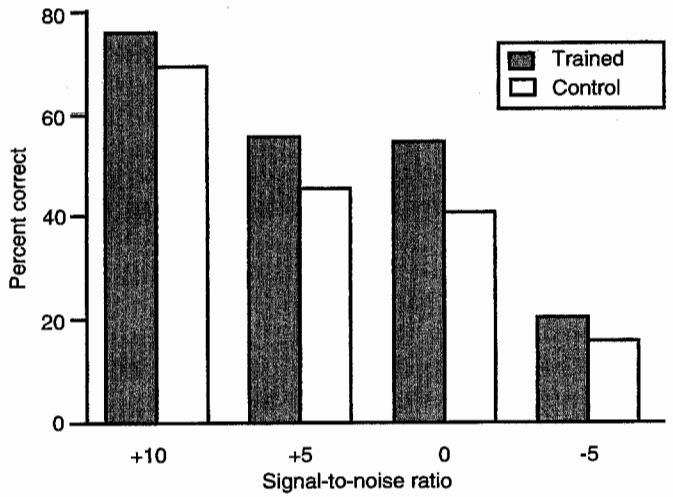


**FIGURE 2** Percent correct word recognition (intelligibility) as a function of signal-to-noise ratio for trained and control subjects on the transfer task administered after voice-recognition training was completed. (From Nygaard, Sommers, & Pisoni, 1994.)

than simply recognize the voices, as they had done in the first phase of the experiment.

The results of the intelligibility experiment are shown in Figure 2 for the two groups of subjects. We found that identification performance for the trained group was reliably better than the control group at each of the SNRs tested. The subjects who had heard novel words produced by familiar voices were able to recognize words more accurately than subjects who received the same novel words produced by unfamiliar voices. Two other groups of subjects were also tested in the intelligibility experiment as controls; however, these subjects did not receive any training in recognizing the voices and were therefore not exposed to any of the stimuli prior to listening to the same set of words in noise. One control group received the set of words presented to the trained experimental group; the other control group received the words that were presented to the trained control subjects. The performance of these two control groups was not only the same, but was also equivalent to the intelligibility scores obtained by the trained control group. Thus, only the subjects in the experimental group who were explicitly trained on the voices showed an advantage in recognizing novel words produced by familiar talkers.

The findings from this perceptual learning experiment demonstrate that exposure to a talker's voice facilitates subsequent perceptual processing of novel words produced by the same talker. Thus, speech perception and spoken-word recognition draw on highly specific perceptual knowledge about a talker's voice that is obtained in an entirely different experimental task—explicit voice recognition as compared to a speech intelligibility test. Listeners show evidence of tracking changes in their listening environment, specifically, information about a talker's vocal tract transfer function and how it changes over time, and encoding this information in memory for later use when they have to process the linguistic attributes of the signal in order to identify the words in an intelligibility test.



Two additional studies were designed to assess the nature and extent of this kind of perceptual learning (Nygaard & Pisoni, 1995). Subjects in both experiments were trained to recognize a set of ten talkers from sentence-length utterances. In the first experiment, after training was completed, intelligibility was assessed using isolated words produced by familiar and unfamiliar talkers. The aim was to determine if the information learned about a talker's voice from sentences generalizes to the perception of spoken words. The assumption was that training with sentence-length utterances would focus listeners' attention on a different set of acoustic properties than training with isolated words. It was hypothesized that because sentences contain extensive prosodic and rhythmic information in addition to the specific acoustic-phonetic implementation strategies unique to individual talkers, perceptual learning of voices from sentences would require attentional and encoding demands specific to those particular test materials.

In the second experiment, after training on sentences was completed, listeners were given an intelligibility test consisting of sentence-length utterances produced by familiar and unfamiliar talkers. Two issues were addressed here. First, does specific training on sentence-length utterances generalize to similar test materials? Second, are sentence-length utterances that have higher-level semantic and syntactic constraints susceptible to the effects of familiarity with a talker's voice?

All subjects showed continuous improvement over the 3 days of training. Both groups of subjects identified talkers consistently above chance even on the first day of training, and performance rose to nearly 85% correct by the last day of training.

Surprisingly, in the first experiment we found only a small unreliable difference in the word intelligibility test for subjects who heard familiar voices during training compared to the control subjects. These results suggest that perceptual learning of talkers' voices from sentence-length utterances does not generalize to the perception of isolated words.

The results obtained in the second experiment, which assessed sentence intelligibility at three SNRs, were quite different. Figure 3 shows the total number of "key words" that were correctly identified in each sentence. Large differences were observed for the subjects who had heard novel sentences produced by *familiar* talkers. These differences became even larger as the listening conditions became poorer, demonstrating transfer of knowledge about the talker's voice from sentence-length utterances.

The results of these experiments suggest that perceptual learning of voices is both talker- and task-specific. Perceptual learning transfers in a task-specific manner, suggesting that attention must be directed to learning the specific voice attrib-

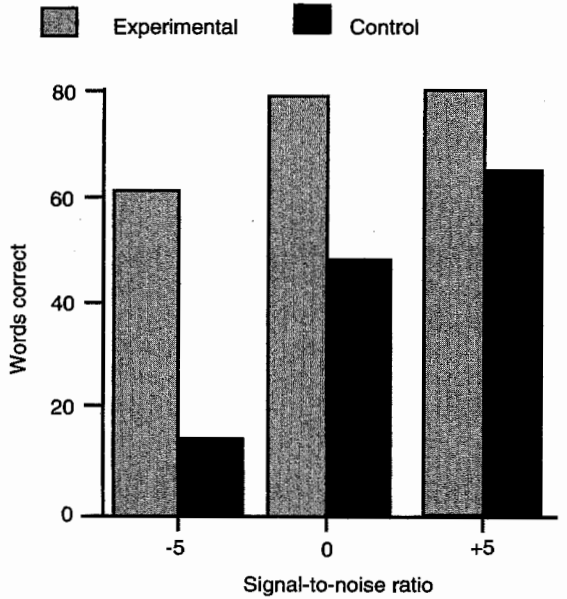


**FIGURE 3** Percent correct recognition of key words in sentences as a function of signal-to-noise ratio for experimental and control groups. (From Nygaard & Pisoni, 1995.)

utes that will be relevant at test. These findings also show that talker-specific effects occur with sentence-length materials that contain higher-level semantic and syntactic constraints. Thus, talker-specific effects operate in a variety of listening situations from isolated words to sentence-length utterances, but there are both talker- and task-specific constraints on the transfer of this kind of knowledge.

Familiarity with a talker's voice involves a form of perceptual tuning and adjustment that results in modification of speech- and language-processing mechanisms. Listeners appear to retain talker-specific information about individual articulatory idiosyncrasies both at the level of acoustic-phonetic implementation in isolated words and at a more global level found in sentence-length utterances. Once again, we see close interactions and dependencies between the indexical properties of speech associated with the talker's voice and the linguistic analysis of the speech signal used in recognizing words in sentences.

What kind of perceptual knowledge does a listener acquire when he or she listens to a speaker's voice and is required to carry out an explicit name-recognition task, as our subjects did in these experiments? One possibility is that the analysis procedures or perceptual operations (Kolers, 1973) used to recognize the voices are retained in some type of "procedural" memory system and that these same processing 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 each time a new word was encountered. Another possibility is that the specific instances—perceptual episodes or exemplars of each talker's voice—are stored in a composite memory system and then later retrieved during the process of word recognition when new tokens from a familiar talker are presented (Estes, 1994; Jacoby & Brooks, 1984).

Whatever the exact nature of this 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 and sentences produced by the same talker. These findings demonstrate a form of implicit memory for a talker's voice that is distinct from the retention of the individual lexical items and sentences used and the specific perceptual learning task that was employed to familiarize the listeners with the voices (Roediger, 1990). These results provide additional support for the view that the neural representation of spoken words and sentences encompasses both a symbolic description of the utterance in terms of a phonetic representation *and* additional information about the structural description of the source characteristics of the specific talker. Thus, speech perception appears to be carried out in a "talker-contingent" manner; indexical and linguistic properties of the speech signal are apparently closely interrelated and are not dissociated in perceptual analysis (Nygaard & Pisoni, 1995; Nygaard et al., 1994). One of the important discoveries of studies done within the framework of nonanalytic cognition is that many aspects of perception, learning, and categorization may not be available to conscious recollection but nevertheless will affect perceptual analysis and memory processes in a variety of ways. Thus, sweeping conclusions like those of Halle (1985) and Brown (1990) about the retention of voice-specific information in speech will obviously have to be revised in light of the present set of findings on perceptual learning of voices:

> When we learn a new word we practically never remember most of the salient acoustic properties that must have been present in the signal that struck our ears; for example, we do not remember the voice quality of the person who taught us the word or the rate at which the word was pronounced. Not only voice quality, speed of utterance, and other properties directly linked to the unique circumstances surrounding every utterance are discarded in the course of learning a new word. (Halle, 1985, p. 101)

> Clearly most of the time anyone is listening to English being spoken, he is listening for the meaning of the message—not to how the message is being pronounced. Indeed if you listen to how the words are spoken it is very unlikely that you can simultaneously understand what it is that is being said. On the whole people do not listen critically to the way the message is pronounced. The odd glottal stop or unusual pronunciation of a word may strike the listener, but most of the time he is busy abstracting the meaning of the message, and preparing his own mental comments on it. This is why most people are quite unaware of how English is actually spoken. (Brown, 1990, p. 3)

## 2.5 EXEMPLAR-BASED APPROACH TO SPEECH PERCEPTION

The approach to speech perception proposed here relies heavily on earlier work by Jacoby and Brooks on nonanalytic concept formation. Their studies on

categorization and memory have provided evidence for the encoding and retention of episodic information and the details of perceptual analysis (Brooks, 1978; Jacoby & Brooks, 1984; Tulving & Schacter, 1990). According to this approach, stimulus variability is informative for perceptual analysis. Memory involves encoding both specific instances and the specific processing operations used during recognition (Kolers, 1973; 1976). The major emphasis of this view of cognition is on particulars, rather than abstract generalizations or symbolic coding of the stimulus input into idealized categories. I believe that the problems of variability and invariance found in speech perception can be approached in a fundamentally different way by nonanalytic or instance-based accounts of perception and memory. Moreover, this view of cognition provides insights into several long-standing theoretical issues in speech and offers a potentially useful way of dealing with findings that show that listeners encode very fine details of their stimulus environment.

The findings from studies on nonanalytic cognition are directly relevant to theoretical questions about the nature of perception and memory for speech and to assumptions about abstractionist representations based on formal linguistic analyses. When the criteria used for postulating episodic or nonanalytic representations are examined carefully, it becomes clear that speech signals display a number of distinctive properties that make them excellent candidates for this approach (Brooks, 1978; Jacoby & Brooks, 1984). These criteria are summarized in the sections below.

### 2.5.1 High Stimulus Variability

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, changes in speaking rate and dialect, differences in social contexts, syntactic, semantic, and pragmatic effects and emotional state, as well as a variety of ambient environment effects, 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.

### 2.5.2 Complex Category Relations

The 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 meaning contrast, phonetic similarity, and distributional characteristics. In traditional taxonomic linguistics, for example, the concept of a phoneme is used in a number of different ways, as shown by the definitions from Gleason (1961) given in Table II.



**TABLE II** Definitional Characteristics of the Phoneme[a]

- The phoneme is the minimum feature of the expression system of a spoken language by which one thing that may be said is distinguished from any other thing that might have been said.
- A phoneme is a class of sounds. There is no English phoneme that is the same in all environments, though in many phonemes the variation can easily be overlooked, particularly by a native speaker.
- A phoneme is a class of sounds that: (1) are phonetically similar and (2) show certain characteristic patterns of distribution in the language or dialect under consideration.
- A phoneme is one element in the sound system of a language having a characteristic set of interrelationships with each of the other elements in that system.
- The phoneme cannot, therefore, be acoustically defined. The phoneme is instead a feature of language structure. That is, it is an abstraction from the psychological and acoustical patterns that enables a linguist to describe the observed repetitions of things that seem to function within the system as identical in spite of obvious differences. The phonemes of a language are a set of abstractions.

[a]From Gleason, 1961

Thus, the perceptual categories used in speech display complex relations that place a number of strong constraints on the class of models that can account for these operating principles. These categories cannot be defined clearly and easily by simple rules and must be identified via a set of relations involving sound contrast and lack of contrast in meaning within a specific phonological system.

### 2.5.3 Incomplete Information

Spoken language is a highly redundant symbolic system that has evolved to maximize transmission of linguistic information. In the case of speech perception, research has demonstrated the existence of multiple speech cues for almost every phonetic contrast. Although these speech cues are, for the most part, highly context-dependent, they also provide reliable information that can facilitate comprehension of the intended message when the signal is presented under degraded conditions. This feature of speech perception permits very high rates of information transmission even under poor listening conditions.

### 2.5.4 High Analytic Difficulty

Speech is inherently multidimensional in nature. As a consequence, many quasi-independent articulatory attributes can be mapped onto the phonological categories of a specific language. 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 formalize a set of explicit rules that can successfully map speech cues onto discrete phoneme categories. 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.

### 2.5.5 Perceptual Spaces for Perception and Production

Among category systems, speech appears to be unique because of the close interrelations between speech production and speech perception. Speech exists simultaneously in several very different domains: the acoustic domain, the articulatory domain, and the perceptual domain. Although the relations among these domains are complex, they are not arbitrary and reflect properties of a unitary articulatory event. And, even within the domain of production and articulation, speech is encoded simultaneously in the optical display of the talker's face and the acoustic signal generated by the vocal tract. The sound contrasts used in a given language function within a common linguistic system that is shared by both production and perception. Thus, the phonetic contrasts generated in speech production by the vocal tract are precisely the same acoustic differences that are distinctive in perceptual analysis (Stevens, 1972). As a result, any theoretical account of speech perception must also take into consideration aspects of speech production and acoustics, as well as the multimodal relations between the auditory and visual correlates of speech.

In learning the sound system of a language, children not only develop abilities to discriminate and identify sounds, but they also learn to control the motor mechanisms used in articulation to generate precisely the same phonetic contrasts in speech production to which they have become attuned in perception. One reason that the developing perceptual system might preserve very fine phonetic details, 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). This perceptuomotor skill would provide children with an enormous benefit in acquiring the phonology of the local dialect from speakers they are exposed to early in life.

In short, when properties of speech signals are examined more closely, and when the criteria for nonanalytic cognition are applied and evaluated, it becomes plausible to assume that very detailed information about specific instances in speech perception might be stored in memory. In contrast to a traditional symbolic rule-based representational approach, listeners may store a very large number of specific instances or perceptual episodes and then use them in an analogical rather than analytic way to perceive and categorize novel stimuli (Brooks, 1978; Whittlesea, 1987). Recent findings from studies on talker variability in speech perception provide support for this conclusion.

The traditional view of "normalization" assumes that perceptual constancy is achieved via a process or set of procedures involving a reduction in variability and a "loss" of detailed stimulus information. But this view is wrong. Human listeners preserve fine phonetic details and appear to encode other aspects of



their listening environment, including indexical information about the talker's voice, gender, dialect, speaking rate, affect, and speaking style, among other nonlinguistic attributes in their surroundings. We believe these are important new insights into speech perception and spoken-language processing that will open up new areas of research on the study of sources of variability in speech perception.

## 2.6 SUMMARY AND CONCLUSIONS

This chapter began with a review and discussion of the term *normalization* as it has been used in the field of speech perception. An examination of several dictionary definitions revealed a number of distinctive properties of normalization that were generally consistent with the way this term is used technically in the field of speech perception to deal with perceptual constancy and invariance. Three consequences of normalization were addressed in greater detail. First, I discussed the assumption that normalization produces equivalent forms. Second, I examined the proposal that normalization entails a loss or reduction of information. And third, I considered the property of normalization that involves elimination of variability among specific instances. All three attributes of "normalization" are assumed in traditional abstractionist accounts of speech perception and spoken-word recognition, and all three of these attributes have played important roles in theorizing about the nature of the perceptual and neural mechanisms used in speech perception.

In the second section, I summarized findings from a series of recent experiments on the contribution of stimulus variability to speech perception and spoken-word recognition. Other results on learning novel voices and the contribution of voice information to word recognition and sentence intelligibility were also described. These studies with normal-hearing listeners demonstrate that fine details of the listener's perceptual environment are encoded and stored in memory and used at a later time during the perceptual analysis of novel stimuli. The results on the perceptual learning of voices demonstrate that the linguistic and indexical properties of speech are not maintained independently as separate channels of information by the nervous system. Interactions between these two properties of the speech signal occur early on in perceptual analysis and produce mutual dependencies that affect the efficiency of subsequent perceptual operations.

In the third section, the above-referenced observations and recent findings were considered together within the framework of recent developments in perceptual learning, memory, and categorization, specifically, the nonanalytic or instance-based approach to cognition that emphasizes the episodic encoding of specific details of the stimulus environment. The studies on talker and rate variability provide important information about speech perception and spoken-word recognition and have raised a set of new questions for future research. These

findings encourage researchers to look at several of the long-standing problems in speech perception in different ways than they have been able to do in the past.

The present findings and reanalysis of the term normalization suggests several directions for research on speech perception. Exemplar-based or episodic models of categorization provide new solutions to the problems of invariance, variability, and perceptual normalization, which have been difficult to resolve with traditional models of speech perception that were motivated by formal linguistic analyses of spoken language. These problems can now be approached in quite different ways when viewed within the general framework of nonanalytic or instance-based models of cognition, which provide novel ways of dealing with stimulus variability—one of the most difficult problems in the field of speech perception and spoken-language processing.

## ACKNOWLEDGMENTS

This research was supported in part by National Institutes of Health (NIH) Research Grant DC-00111 and in part by National Institute on Deafness and Other Communication Disorders (NIDCD) Research Training Grant DC00012 to Indiana University in Bloomington, Indiana. I thank Steve Chin for his editorial help and Darla Sallee for her assistance in preparing the manuscript.

## REFERENCES

Brooks, L. (1978). Non-analytic concept formation and memory for instances. In E. Rosch & B. Lloyd (Eds.), *Cognition and categorization* (pp. 169–211). Hillsdale, NJ: Erlbaum.

Brown, G. (1990). *Listening to spoken English* (2nd ed.). New York: Longmans.

Bryd, D. (1992). Sex, dialects, and reduction. *ICSLP 92 Proceedings,* pp. 827–830.

Chomsky, N. (1965). *Aspects of the theory of syntax.* Cambridge, MA: MIT Press.

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

Eich, J. E. (1982). A composite holographic associative memory model. *Psychological Review, 89,* 627–661.

English, H. B., & English, A. C. (1958). *A comprehensive dictionary of psychological and psychoanalytical terms: A guide to usage.* New York: Longmans, Green.

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

Fowler, C. A. (1990). Listener-talker attunements in speech. *Haskins Laboratories Status Report on Speech Research, SR-101/102,* 110–129.

Fowler, C. A., & Rosenblum, L. D. (1991). The perception of phonetic gestures. In I. G. Mattingly & M. Studdert-Kennedy (Eds.), *Modularity and the motor theory of speech perception* (pp. 33–59). Hillsdale, NJ: Erlbaum.

Gleason, H. A. (1961). *An introduction to descriptive linguistics.* New York: Holt, Rinehart & Winston.

Goldinger, S. D. (1992). *Words and voices: Implicit and explicit memory for spoken words. Research on Speech Perception* (Tech. Rep. No. 7). Bloomington: Indiana University.

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.



Haber, R. N. (1969). *Information-processing approaches to visual perception.* New York: Holt, Rinehart & Winston.

Halle, M. (1956). *For Roman Jakobson: Essays on the occasion of his sixtieth birthday, 11 October 1956.* The Hague: Mouton.

Halle, M. (1985). Speculations about the representation of words in memory. In V. A. Fromkin (Ed.), *Phonetic linguistics* (pp. 101–104). New York: Academic Press.

Hintzman, D. L. (1986). Schema abstraction in a multiple-trace memory model. *Psychological Review, 93,* 411–423.

Hirahara, T., & Kato, H. (1992). The effect of FO on vowel identification. In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (Eds.), *Speech perception, production and linguistic structure* (pp. 89–112). Tokyo: Ohmsha Publishing.

Hockett, C. F. (1955). *A manual of phonology.* Baltimore: Waverly Press.

Huttenlocher, D. P., & Zue, V. W. (1984). A model of lexical access based on partial phonetic information. *Proceedings ICASSP-84,* pp. 1–4.

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). New York: Academic Press.

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). Tokyo: Ohmsha Publishing.

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 & D. H. Klatt (Eds.), *Invariance and variability in speech processes* (pp. 300–319). Hillsdale, NJ: Erlbaum.

Klatt, D. H. (1989). Review of selected models of speech perception. In W. Marslen-Wilson (Ed.), *Lexical representation and process* (pp. 169–226). Cambridge, MA: MIT Press.

Kolers, P. A. (1973). Remembering operations. *Memory & Cognition, 1,* 347–355.

Kolers, P. A. (1976). Pattern analyzing memory. *Science, 191,* 1280–1281.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21,* 1–36.

Licklider, J. C. R. (1952). On the process of speech perception. *Journal of the Acoustical Society of America, 24,* 590–594.

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.

Mullennix, J. W., & Pisoni, D. B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception & 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.

Nygaard, L. C., & Pisoni, D. B. (1995). Talker- and task-specific perceptual learning in speech perception. *Proceedings of the XIIIth International Congress of Phonetic Sciences, Stockholm, Sweden,* pp. 194–197.

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. (1995). Effects of stimulus variability on perception and representation of spoken words in memory. *Perception & Psychophysics, 57,* 989–1001.

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,* 1–20.

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). Pensacola, FL: U.S. Naval School of Aviation Medicine.

Pisoni, D. B. (1990). Effects of talker variability on speech perception: Implications for current research and theory. *Proceedings of the 1990 International Conference on Spoken Language Processing, Kobe, Japan,* pp. 1399–1407.

Pisoni, D. B. (1992). Some comments on invariance, variability and perceptual normalization in speech perception. *Proceedings of the 1992 International Conference on Spoken Language Processing, Banff, Canada,* pp. 587–590.

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.

Reber, A. S. (1985). *The Penguin dictionary of psychology.* Harmondsworth, UK: Penguin Books.

Roediger, H. L. (1990). Implicit memory: Retention without remembering. *American Psychologist, 45,* 1043–1056.

Shankweiler, D., Strange, W., & Verbrugge, R. (1977). Speech and the problem of perceptual constancy. In R. Shaw & J. Bransford (Eds.), *Perceiving, acting, and knowing: Toward an ecological psychology* (pp. 315–345). Hillsdale, NJ: Erlbaum.

Shipman, D. W., & Zue, V. W. (1982). Properties of large lexicons: Implications for advanced isolated word recognition systems. *Proceedings ICASSP-82,* pp. 546–549.

Sommers, M. S., Nygaard, L. C., & Pisoni, D. B. (1994). Stimulus variability and spoken word recognition: I. Effects of variability in speaking rate and overall amplitude. *Journal of the Acoustical Society of America, 96,* 1314–1324.

Stevens, K. N. (1972). The quantal nature of speech: Evidence from articulatory acoustic data. In E. E. David, Jr., & P. B. Denes (Eds.), *Human communication: A unified view* (pp. 51–66). New York: McGraw-Hill.

Studdert-Kennedy, M. (1974). The perception of speech. In T. A. Sebeok (Ed.), *Current trends in linguistics* (pp. 2349–2385). The Hague: Mouton.

Studdert-Kennedy, M. (1983). On learning to speak. *Human Neurobiology, 2,* 191–195.

Tulving, E., & Schacter, D. L. (1990). Priming and human memory systems. *Science, 247,* 301–306.

*Webster's new twentieth century dictionary.* (1983). New York: Simon & Schuster.

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.

Zue, V. W. (1985). The use of speech knowledge in automatic speech recognition. *Proceedings of the IEEE, 73*(11), 1602–1615.

# 3

# Words and Voices

## *Perception and Production in an Episodic Lexicon*

**STEPHEN D. GOLDINGER**

## 3.1 INTRODUCTION

In the first decade of the twentieth century, a German scientist named Richard Semon described a theory of memory that anticipated many aspects of contemporary theories (Schacter, Eich, & Tulving, 1978; Semon, 1909/1923). In modern parlance, we would call Semon's view an "episodic," "exemplar," or "multiple-trace" theory, because it assumed that every stimulus, such as a spoken word, leaves a unique trace in memory. Upon presentation of a new stimulus, all traces are activated in proportion to their similarity to the stimulus. The most activated traces come to consciousness and the stimulus is thus "recognized." The assumption of unique traces allowed Semon's theory to explain the permanence of specific memories; the challenge was to also create *abstraction* from a collection of idiosyncratic traces. The resolution apparently derived from Galton, who discovered that blending many faces into a photographic composite creates the image of a "generic" face. Galton applied this phenomenon as a memory metaphor: "Whenever a single cause throws different groups of brain elements simultaneously into excitement, the result must be a blended memory" (1883, p. 229; cited in Hintzman, 1986). Semon used this idea, assuming that abstraction occurs at *retrieval* as countless partially redundant traces respond to an input.

This book is printed on acid-free paper. ♾



Academic Press, Inc.
525 B Street, Suite 1900, San Diego, California 92101-4495, USA
http://www.apnet.com

Academic Press Limited
24-28 Oval Road, London NW1 7DX, UK
http://www.hbuk.co.uk/ap/

Library of Congress Cataloging-in-Publication Data

Talker variability in speech processing / edited by Keith Johnson, John W. Mullennix. -- Ed. 1.
p. cm.
Includes index.
ISBN 0-12-386560-3 (alk. paper)
1. Psycholinguistics. 2. Speech perception. 3. Automatic speech recognition. 4. Language and languages--Variation. I. Johnson, Keith, date. II. Mullennix, John W.
P37.T24 1996
401'.9--dc20
96-31802
CIP

PRINTED IN THE UNITED STATES OF AMERICA
96 97 98 99 00 01 BB 9 8 7 6 5 4 3 2 1

# Contents