Nusbaum, H. C., and Magnuson, J. S. (1997). Talker normalization: Phonetic constancy as a cognitive process. In K. Johnson and J. W. Mullennix (Eds.), Talker Variability in Speech Processing, pp. 109 - 132. San Diego: Academic Press.

6

Talker Normalization

Phonetic Constancy as a Cognitive Process

HOWARD NUSBAUM
JAMES MAGNUSON

6.1 LACK OF INVARIANCE AND THE PROBLEM OF PHONETIC CONSTANCY

Human listeners recognize and understand spoken language quite effectively regardless of the vocal characteristics of the talker, or how quickly the speech is produced, or what the talker has said previously. Even at the most basic level of recognizing spoken consonants and vowels, most humans have little difficulty maintaining phonetic constancy—stable recognition of the phonetic structure of utterances (Shankweiler, Strange, & Verbrugge, 1977) in spite of variation in the relationship between the acoustic patterns of speech and phonetic categories that results from these sources of variability (e.g., Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). Indeed, the perceptual ability of human listeners has still not been matched in engineering efforts to develop computer speech-recognition systems.

Furthermore, even after more than 30 years of scientific endeavor, there are no theories of speech perception that can adequately explain how humans recognize spoken consonants and vowels (see Nusbaum & Henly, in press). Although the theoretical problem posed by the lack of invariance in the relationship between linguistic categories and their acoustic manifestations in the speech signal has been attacked from a number of different perspectives, such as the use of articulatory knowledge (see Liberman, Cooper, Harris, & MacNeilage, 1962; Liberman & Mattingly, 1985; Stevens & Halle, 1967) or linguistic knowledge (G.A. Miller, 1962; Newell, 1975) or biologically plausible mechanisms such as feature detectors (Abbs & Sussman, 1971; McClelland & Elman, 1986), none of these approaches has credibly accounted for phonetic constancy. In theoretical terms, perhaps the most critical feature of the lack of invariance problem is that it makes speech recognition an inherentlynondeterministic process.

In order to understand the significance of this, we need to consider briefly the definition of a finite state automaton (Gross, 1972; Hopcroft & Ullman, 1969). A finite state automaton is an abstract computational mechanism that can represent (in terms of computational theory) abroad class of different "real" computational processes. A finite state automaton (FSA) consists of a set of states (that differ from each other), a vocabulary of symbols representing inputs, a mapping process that denotes how to change to a new state given an old state and an input symbol, a starting state, and a set of ending states. Finite state automata have been used to represent and analyze grammars (e.g.. Gross, 1972) and other formal computational problems (Hopcroft & Ullman, 1969). For our purposes, the states in an FSA representing speech perception can be regarded as denoting internal linguistic states, such as phonetic features or categories, and the input symbols can be thought of as acoustic properties present in an utterance. The possible orderings of sequences of permissible states in the processing carried out by the automaton can be thought of as the phonotactic constraints inherent in language. The transition from one state to another, which determines those orderings, is based on acoustic input with acoustic cues serving as the input symbols to the system. This is a relatively uncontroversial conceptualization of speech recognition (e.g., Klatt, 1979; Levinson, Rabiner, & Sondhi, 1983; Lowerre & Reddy, 1980) and is similar to the use of finite state automata in other areas of language processing, such as syntactic analysis (e.g., Chamiak, 1993; Woods, 1973).

A deterministic finite state automaton changes from one state to another such that the new state is *uniquely* determined by the information (i.e., next symbol) that is processed. In speech this means that if there were a one-to-one relationship between acoustic information and the phonetic classification of that information (i.e., each acoustic cue denotes one and only one phonetic category or feature), a wide variety of relatively simple deterministic computational mechanisms (e.g., some simple context-free grammars, Chomsky, 1957; feature detectors. Abbs & Sussman, 1971) could be invoked to explain the apparent ease with which we recognize speech. Unfortunately, as researchers know all too well by now, this relationship is much more complex. Instead, there is a many-to-many mapping between acoustic patterns and phonetic categories, which is referred to as the lack of invariance problem.

Any particular phonetic category (or feature) may be instantiated acoustically by a variety of different acoustic patterns. Conversely, any particular acoustic pattern may be interpreted as a variety of different phonetic categories (or features). Although a many-to-one mapping can still be processed by a deterministic finite state automaton, because each new category or state is still uniquely determined, albeit by different symbols or information (e.g., the set of different cues any one of which could denote a particular feature), the one-to-many mapping represents a nondeterministic computational problem. Given the current state of an FSA and the input acoustic information, there are often multiple possible states to which the system could change. There is nothing inherent in the input symbol or acoustic information, or in the system itself, that uniquely determines the classification of that information (i.e., the next state of the system). In other words, there is a computational ambiguity that is unresolvable given just the information that describes the system.

The classic empirical demonstrations of the lack of invariance problem come from early research on perception of synthetic speech (Liberman, Cooper, Harris, MacNeilage, & Studdert-Kennedy, 1967). Two different second formant (F2) transitions are heard as /d/ in the context of different vowels (Delattre, Liberman, & Cooper, 1955) demonstrating that very different acoustic patterns may be interpreted as signaling a single phonetic category. As already noted, this kind of many-to-one mapping between acoustic patterns and phonetic categories can be processed easily by a deterministic FSA mechanism, whereas the converse case of a one-to-many mapping has different computational-theoretic implications inherent in a nondeterministic mechanism. It is the demonstration that a single consonant release burst cue may be interpreted as either of two different phonetic categories /p/ or /k/ depending on the vowel context (Liberman, Delattre, & Cooper, 1952) which indicates that recognition of phonetic structure is inherently nondeterministic.

The problem for theories of speech perception is to explain how a listener can recover the phonetic structure of an utterance given the acoustic properties present in the speech signal. The real computational problem underlying this, which must be addressed by theories, is presented by those cases in which one acoustic cue can be interpreted as more than one phonetic feature or category. Because this is inherently a nondeterministic problem, it may require a different kind of computational solution than would be required by a deterministic problem.

6.2 COMPUTATIONAL CONSTRAINTS ON THEORIES OF SPEECH PERCEPTION

If the one-to-many mapping in speech specifically determines the class of computational mechanisms that can produce phonetic constancy, this should constrain the form of theories that are appropriate to explaining speech perception. Thus it is important to consider whether or not we need to be concerned about this kind of computational constraint. If this computational constraint is important (i.e., it cannot be dismissed or resolved by some trivial solution), it follows that it is important to consider how well extant theories of speech perception conform to this constraint.

Let us start by considering for a moment how we might distinguish between classes of computational mechanisms (Nusbaum & Schwab, 1986) and how this distinction relates to the issue of deterministic versus nondeterministic computational problems. Computational mechanisms can be thought of generally as consisting of three classes of elements: representations of information, transformations of the representations, and control structures. Control structures determine the sequencing of transformations that are applied to representations. When denned this way, computational mechanisms can be sorted into two types based on the nature of the control structure. In passive systems the sequence of transformations is carried out according to an open-loop control structure (MacKay, 1951, 1956). This means that given the same input at two different points in time, the same sequence of transformations will be carried out so that there is an invariant mapping from source to domain (in functional terms). For example, in motor control, a ballistic movement is considered to be controlled as an open-loop system. Also, feature detectors can be thought of as operating as passive mechanisms at least in the overly simplified form that is generally used in psychological models (Barlow, 1972). Passive systems constitute relatively simple and generally easily understood computational mechanisms. If speech perception could be carried out by a passive system, such as represented in a deterministic finite-state automaton, theories of speech perception would be relatively easy to specify and analyze.

By contrast, in an active system the sequence of transformations is adapdvely controlled by a closed-loop control structure (Nusbaum & Schwab, 1986). This means that the flow of computation is contingent on the outcome of certain comparisons or tests in a feedback loop. This kind of system is generally used when there is a need for an error-correcting mechanism that allows systematic adjustment of processing based on the outcome of the transformations and can be used to address nondeterministic computational problems. Nusbaum and Schwab (1986) described two different types of active systems. These types of systems can be thought of as hypothesize-test-adjust or approximate-test-adjust systems. In the former, higher level knowledge-based processes propose hypotheses that are tested against bottom-up transformations of input. In the latter, an approximate classification or target is proposed from the bottom-up, and derived implications of this approximation are compared with other analyses from either top-down or bottom-up processing. Both types of active systems have been proposed as important in various cognitive processes (e.g., Grossberg, 1986; Minsky, 1975; Neisser, 1967). In an active system, the sequence of transformations may differ given the same input in different contexts.

Because the relation between input and output is formally a mathematical function or computationally deterministic in passive systems, this class of systems cannot generally address nondeterministic computational problems, where one input may lead to either of two different states. (This is similar to the constraint that a mathematical function must produce one and only one value for each input.) Typically, an active system would be required to address these nondeterministic problems. Of course there are some exceptions, such as Hidden Markov Models (HMM) (Chamiak, 1993; Levinson et al., 1983, for reviews and discussion). HMMs are nondeterministic finite state automata that handle nondeterministic computational problems without a closed-loop control structure. Instead they select one of the alternative states based on statistics estimated from a set of "observations" made during a "training" process. An HMM resolves nondeterministic choices by estimating the distributional statistics for those choices and basing the decision on those statistics.

In general, HMMs have provided the most successful engineering solutions to the development of speech recognition systems because they explicitly recognize the inherent nondeterministic nature of the recognition problem. The most accurate and robust commercially available recognition systems are based on HMMs (e.g., Nusbaum & Pisoni, 1987). However, even though HMM-based systems are the most successful recognition systems, they do not perform as well as human listeners (e.g., Klatt, 1977; Nusbaum & Pisoni, 1987). These systems cannot recognize words in fluent, continuous speech as well as human listeners can, nor do they handle the effects of background noise or changes in speaking rate as well as human listeners do. In part, this may reflect the fact that the states and units of acoustic analysis are more rigid than are employed by human listeners (e.g., Nusbaum & Henly, 1992). This may also be partly due to the use of statistics to resolve the nondeterministic nature of the recognition problem; this kind of statistical approach may really only be a statistical approximation of the kind of mechanism used in human speech perception.

Rather than approximate a nondeterministic solution statistically, an alternative is to find a way of restructuring a nondeterministic problem (such as the lack of invariance problem) that eliminates the nondeterminism. This would make it possible to use a deterministic mechanism as the basis for phonetic constancy in speech perception. In computational theory of formal languages, there is a theorem that states that for any nondeterministic system there exists an equivalent deterministic system (Hopcroft & Ullman, 1969). Another way to say this is if there is a nondeterministic system such as the relationship between acoustic cues and phonetic categories, there exists a deterministic system that can account for this relationship. On the face of it, this suggests that although there are aspects of spoken language that might be characterized as requiring nondeterministic processing, it is possible to construct a deterministic mechanism to account for processing this information. Some deterministic automaton can be constructed that can provide a complete description of the nondeterministic problem represented by the mapping of acoustic cues onto phonetic structure. If such a deterministic description is possible, then this may describe the processing mechanism used in human speech perception and would allow the use of simpler computational devices such as passive mechanisms.

However, the proof of the theorem regarding the equivalence of a deterministic and nondeterministic system places certain constraints on the form of the deterministic system (Hopcroft & Ullman, 1969). The proof requires the construction of a deterministic system that contains states that are different from the nondeterministic system. Specifically, the new deterministic version of a nondeterministic system requires states that represent the *disjunction* of the set of states that would have been alternatives in the nondeterministic system. Thus these new states in the deterministic machine are actually compounds of the old states in the nondeterministic system. In other words, this does not deterministically resolve the ambiguity as to which of the states the machine should be in given an input. Rather it represents and recognizes the ambiguity explicitly as ambiguous states. So in the /pi/-/ka/-/pu/ example, when given the burst information, a nondeterministic machine could go to either the /p/ state or the /k/ state. A deterministic machine would have a single state called /p/-or-/k/. Clearly this does not resolve the lack of invariance problem in a satisfactory way because moving to the /p/-or-/k/ state would leave ambiguous the phonetic interpretation of the burst cue. Moreover, because phonetic segments are categorically perceived (Studdert-Kennedy, Liberman, Harris, & Cooper, 1970), this kind of phonetic perceptual ambiguity is never found in human listeners.

Another possible way of addressing the problem of a nondeterministic computational problem (or the lack of invariance) may be to change the definition of the states and the assumed form of the input. In a nondeterministic machine, there must be at least one state from which there are several alternative states that may be entered given exactly the same input information. By restructuring either the states or the input patterns, it might be possible to change a nondeterministic problem into a deterministic one, without retaining the ambiguity noted above. (However, I am unaware of any proof of this conjecture.)

For example, take the basic case in which one acoustic cue could lead to either of two states. If we consider the sequence of states that follow from each of those two possibilities given subsequent inputs, and the different sequences of inputs that would give rise to those sequences of states, it may be possible to convert the nondeterministic system into a deterministic system. Thus, rather than create compound states from the alternative states as described above, it is possible to create compound states that represent the alternative sequences of states that would be entered given a particular sequence of input symbols. This might require changing the definition of the states to be sequentially constituted and the definition of the next input to allow different lengths of sequences of acoustic cues depending on the current state. By the example from the /pi/-/ka/-/pu/ experiment, this would mean that to construct a deterministic system, we would need states /pi/, /ka/, and /pu/ (i.e., combining the consonant state with the subsequent vowel state to form a single sequentially defined state). To distinguish among these states, the input would then need to include information about the vowel in addition to the burst. In other words, we would be converting a system that is nondeterministic in phonetic terms to one that is deterministic in syllabic terms.

Of course, some speech researchers have proposed just this kind of approach by redefining the basic units of speech perception (see Pisoni, 1978, for a discussion) from phonemes to syllables (e.g., Massaro, 1972), or other context-sensitive units (e.g., Wickelgren, 1969) or to entire linguistic phrases (G. A. Miller, 1962). Unfortunately, this does not actually solve the problem because the coarticulatory processes that encode linguistic structure into sound do not respect any particular unit boundaries (e.g., for syllables, Ohman, 1966) so that the same problem of lack ofinvariance arises regardless of the size of the unit of analysis. Furthermore, empirical evidence suggests that listeners do not have a fixed size unit of analysis for recognizing speech (Nusbaum & Henly, 1992). The listener does not process a fixed amount of speech signal in order to recognize a particular unit. Nusbaum and Henly (1992) have argued that listeners dynamically adapt the analysis of the acoustic structure of speech as a result of available linguistic and informational constraints and the immediate linguistic-perceptual goals.

An alternative approach to restructuring the states and analysis of input is to change the definition of what is being recognized. For example, the ecological perspective on speech perception asserts that the states that are being recognized are phonetic gestures (because these are the distal objects of perception) and the input is considered to be gestural rather than acoustic (see Best, 1994; Fowler, 1989). Indeed, if we consider theories of speech perception generally, there has been a tendency to approach the problem of phonetic constancy by redefining the type of knowledge that is used in perception, often without considering the role or nature of the computational mechanism that is required. For example, articulatory theories (e.g., different forms of motor theory, Liberman et al., 1962; Liberman & Mattingly, 1985; and analysis-by-synthesis, Stevens & Halle, 1967) argue that knowledge of the process of speech production by which linguistic units are encoded into sound would resolve the lack of invariance problem. By contrast, Newell (1975) claimed that the acoustic signal underdetermines the linguistic structure that must be recovered, and so broader linguistic knowledge about lexical structure, syntax, and semantics must be used to constrain the recognition process (also see Klatt, 1977; G. A. Miller, 1962). From yet another perspective Stevens and Blumstein (1978, 1981) have argued, in essence, that the lack of invariance problem is a result of selecting the wrong acoustic properties for mapping onto phonetic features. Thus, their claim is that it is important to carefully define which acoustic properties are selected as the input tokens. In the Lexical Access From Spectra (LAFS) model, Klatt (1979) essentially combined this claim with a redefinition of which linguistic categories were actually being recognized.

However, none of these approaches has been entirely successful or convincing in explaining phonetic constancy. All depend on the assumption that the appropriate kind of knowledge or representation will be sufficient to restructure the nondeterministic relationship between the acoustic patterns of speech and the linguistic interpretation into a deterministic relationship, but empirical studies are not encouraging. Measures of speech production show as much lack ofinvariance in the motor system as there is in the relationship between acoustics and phonetics (e.g., MacNeilage, 1970). Similarly, there is as much lack of invariance between sound patterns and larger linguistic units (e.g., syllables, words, etc.) as there is with phonemes or phonetic features. And the perspective that better acoustic knowledge or the right acoustic analysis would provide an invariant mapping has failed as well. For the paradigm case of place of articulation perception, it turns out that listeners do not make use of the information that Stevens and Blumstein (1978, 1981) claimed was the invariant cue. Walley and Carrell (1983) demonstrated that listeners carry out phonetic classification by using the noninvariant portions of the signal rather than the invariant portions.

Perhaps theories of speech perception have largely failed to explain phonetic constancy given the problem of lack of invariance because they have taken the wrong tack on analyzing the problem. By focusing on a content analysis of the lack of invariance problem, these theories have tried to specify the type of information or knowledge that would permit accurate recovery of phonetic structure from acoustic patterns. As we have argued, this is an attempt to change the computational structure of the lack of invariance problem from a nondetenninistic problem to a deterministic problem. Perhaps the failure of these theories to yield convincing and completely explanatory accounts of phonetic constancy is a consequence of focusing on trying to find a kind of knowledge, information, or representation that resolves the lack of invariance problem. Instead, a more successful approach may depend on acknowledging and analyzing the computational considerations inherent in a nondeterministic system. The point of this section has been to argue that it is important to shift the focus of theories from a consideration of the problem of lack of invariance as a matter of determining the correct representation of the information in speech to a definition of the problem in computational terms. We claim speech perception is a nondeterministic computational problem. Furthermore, we claim that deterministic mechanisms and passive systems are incapable of accounting for phonetic constancy. Human speech perception requires an active control system in order to carry out processing. By focusing on an analysis of the specific nature of the active system used in speech perception it will be possible to develop theories that provide better explanations of phonetic constancy.

Active systems have been proposed as explanations of speech perception in the past (see Nusbaum & Schwab, 1986), including analysis-by-synthesis (Stevens & Halle, 1967) and Trace (McClelland & Elman, 1986). These theories have indeed acknowledged the importance of complex control systems in accounting for phonetic constancy. However, even in these theories, the focus has been on the nature of the information (e.g., articulatory in analysis-by-synthesis and acoustic-phonetic, phonological, and lexical in Trace); the active control system has subserved the role of delivering the specific information at the appropriate time or in the appropriate manner. Unfortunately, from our perspective, these earlier theories took relatively restricted views of the problem of lack of invariance (see Nusbaum & Henly, in press, for a discussion). Although all theories of speech perception have generally acknowledged that lack of invariance arises from variation in phonetic **context,** speaking rate, and the vocal characteristics of talkers, most theories have focused on the problem of variation in phonetic context alone. By focusing on the specific knowledge or representations needed to maintain phonetic constancy over variability in context, these theories developed highly specific approaches that do not generalize to problems of talker variability or variability in speaking rate (e.g., see Klatt, 1986, for a discussion of this problem in Trace; Nusbaum & Henly, in press). For these theories, there is no clear set of principles for dealing with nondetenninism in speech that would indicate how to generalize these theories to other sources of variability, such as talker variability.

Our goal is to specify a general set of computational principles that can address the nondeterministic problem posed to the listener by the lack of invariance between acoustic patterns and linguistic categories (see Nusbaum & Henly, in press). If these principles are sufficiently general, they may constitute the basic framework for a theory of speech perception that can account for phonetic constancy regardless of the source of acoustic-phonetic variability.

6.3 TALKER VARIABILITY AND TALKER NORMALIZATION

Two talkers can produce the same phonetic segment with different acoustic patterns and different segments with the same acoustic pattern (Peterson & Barney, 1952). As a result, there is the same many-to-many relationship between acoustic patterns and linguistic categories as a result of differences in the vocal characteristics of talkers. Nonetheless, human listeners are usually quite accurate in recognizing speech regardless of who produces it.

Engineers would love to build a speech recognition system that would accurately recognize speech regardless of who produced it, but this has yet to be done. Most speech recognition systems require some amount of training on the voice of the person who will be using the system in order to achieve accurate levels of performance (Nusbaum, DeGroot, & Lee, 1995; Nusbaum & Pisoni, 1987). Speech recognition systems would be much more useful if they did not require this kind of training on a talker's voice. Nonetheless, in spite of two decades of intense engineering effort directed at building speaker-independent speech recognition systems, this goal has been realized in only the most restricted sense: There are recognition systems that are relatively accurate for a very small vocabulary and there are systems that are relatively inaccurate (compared to humans) for larger vocabularies. In all cases, there are limitations to the set of talkers whose speech can be recognized. For example, one system that used statistical modeling techniques achieved a relatively high level of accuracy for speech produced by talkers from New Jersey, but performance was terrible when the same system was tested on speech produced by talkers from another dialect of American English (Wilpon. 1985). Thus from the engineering perspective, it is clear that speaker-independent speech recognition is an extremely difficult computational task, albeit one that we, as human listeners, solve all the time, such as whenever we answer the phone. The correct solution to this problem is probably not based solely on the perceptual experience listeners have with a wide range of talkers' vocal characteristics because unlike computer speech recognition systems, we can quickly generalize to an accent we have never heard before.

There is no deep mystery about why speaker-independent speech recognition is computationally challenging. Talkers differ in the structure of their vocal tracts and in the way they produce speech (e.g., Fant, 1973). This results in a nondeterministic relationship between acoustic properties and phonetic categories, which means that if given a particular acoustic pattern, there is uncertainty about how to classify it. In order to classify a pattern correctly (i.e., as the talker intended it), it is necessary to know something about the vocal characteristics of the talker who produced the speech. This is the crux of the purported solution to phonetic constancy given talker variability, and this is what distinguishes the problem of talker variability from the problem of variability in phonetic context.

When we consider the problem of talker variability and the theoretical approaches to speech recognition across talkers, we see very different kinds of theories than the more general theories described above such as motor theory or Trace (see Nusbaum & Henly, in press, for a discussion). First, whereas general theories of speech perception focus on the problem of consonant recognition, models of talker normalization address the problem of vowel perception. Thus, different classes of segments are generally targeted by these theories. This is probably because consonants are most greatly affected by changes in phonetic context (e.g., Liberman, Cooper, Shankweiler et al., 1967), whereas the effects of talker differences on vowel spaces are much better understood (e.g., Fant, 1973; Peterson & Bamey, 1952) than differences in the way talkers produce consonants. Second, theories of talker normalization fall into two categories depending on the kind of information used in normalizing talker differences (Ainsworth, 1975; Nearey, 1989). Some theories use extrinsic information (information from preceding context) to estimate or calibrate the talker's vowel space (e.g., Gerstman, 1968); other theories use intrinsic information, meaning that information within the acoustic pattern of the segment being recognized is used to achieve phonetic constancy (e.g., Shankweiler et al., 1977; Syrdal & Gopal, 1986).

Talker normalization is the purported process by which listeners compensate for differences among talkers in order to maintain phonetic constancy regardless of the vocal characteristics of the talker. Is this truly a different process from the process that characterizes the recognition of phonemes in spite of the variability in acoustic patterns produced by different phonetic contexts? From the term *normalization,* one might think so. For example, the term normalization has been used in computational vision to describe a set of passive and simple transformations that render an input pattern into a canonical form for pattern matching. In Roberts' (1965) early object recognition system, a set of prototypes or object templates were used as the basis for determining the identity of an input pattern. However, it was important to modify the input pattern by appropriate rotation, size expansion or compaction, and translation across spatial position, to optimally register the input pattern against the set of templates. If the size of the input and templates were different, or their major axes were not in registration, the contours or other visual features of the input and template set would mismatch for reasons unrelated to the basic level differences among the pattern properties of the objects to be recognized. Simple pattern differences due to orientation, distance, or location of an object should not affect the recognition of the object, so normalization processes were proposed, based on relatively simple criteria, that would eliminate these effects prior to pattern matching.

From this kind of work on computational vision and pattern matching, pattern normalization has been viewed as a distinct process from the process of recognition (e.g., see Uhr, 1973). First, normalization processes are typically viewed as preceding pattern recognition for the purpose of eliminating variation that is not intrinsic to the definition of a pattern. Second, normalization processes have been viewed as linear (simple) transformations of the input such as rotation, translation, and magnification. Finally, these processes have been viewed as passive filtering mechanisms.

In speech perception, some researchers have assumed that talker normalization operates in much the same way, although it is not clear why this should necessarily be the case. For example, Palmeri, Goldinger, and Pisoni (1993) have argued from recent data that talker information is retained within the episodic trace that is encoded during the perception of spoken words. They suggest that if normalization transforms an input pattern into some canonical form, thereby stripping out talker vocal characteristics, this kind of normalization cannot be carried out (see also Nygaard, Sommers, & Pisoni, 1994). (We can ignore for present purposes the fact that there exist parallel multiple representations of any stimulus pattern in the auditory system such that some may represent transforms of one kind and others may be relatively untransformed, rendering this logic questionable.) This assumption is largely based on the structure of many models of talker normalization: As in computational vision, these may take the form of a passive filtering process that transforms an input stimulus into some canonical or talker-independent form for subsequent matching to phonetic categories. The model proposed by Syrdal and Gopal (1986) is just this kind of system. Bark scaling by FO and F3 is used to modify the pattern of Fl and F2 of vowels in order to be compared to a set of prototype vowels. In this regard then, talker normalization is simply a passive filtering process that precedes the "real" computational work of recognizing phonetic structure.

Furthermore, the kind of knowledge and information that is used during talker normalization is thought to be different from the knowledge used to account for phoneme recognition. In order to carry out talker normalization, it is necessary to derive information about the talker's vocal characteristics. For example, in Gerstman's (1968) model, the point vowels are used to scale the **location** of the Fl-F2 space of all the other vowels produced by a given talker. Because the point vowels represent the extremes of a talker's vowel space, they can be used to characterize the talker's vocal tract extremes and therefore bound the recognition space. Similarly, Syrdal and Copal's model scales Fl and F2 using the **talker's** fundamental frequency and F3 because these are considered to be more characteristic of the talker's vocal characteristics rather than of vowel quality (e.g., Fant, 1973; Peterson & Bamey, 1952). Thus, talker normalization models use information about the talker's vocal characteristics rather than information about the specific message or phonetic context, as in models of phoneme perception such as Trace (McClelland & Elman, 1986) or Motor Theory (Libennan, Cooper, Hams, & MacNeilage, 1962) or analysis-by-synthesis (Stevens & Halle, 1967) or the Fuzzy Logical Model of Perception (Massaro, 1987; Massaro & Oden, 1980).

From all these considerations it appears that there is a belief among speech researchers that coping with talker variability is a different kind of process from coping with variability due to phonetic context. Thus in spite of the traditional acknowledgment that the lack of invariance problem in speech is manifest due to variability in context, speaking rate, and talker, this really means that there are different variability problems that require different theoretical solutions. Even in more recent models such as Trace, Elman and McClelland (1986) made the argument that speech perception requires a general approach to coping with "sources of lawful variability" in speech. However, Trace itself is highly specialized to address the specific problem of coping with variability due to phonetic context, and there is no set of general principles presented that would permit extension of this model to address talker variability or speaking rate variability (see Klatt, 1986, for a discussion). In spite of the general claims, and although seldom explicitly presented this way, the modal theoretical view of speech perception is that there are a set of specialized normalizing processes that act as passive filters (e.g., one for speaking rate, cf. J. L. Miller & Dexter, 1988; one for talker vocal characteristics, Syrdal & Gopal, 1986) that transform the input signal into some canonical form for comparison to a set of linguistic category prototypes. This final prototype-matching operation is the focus and concern of most theories of speech perception (e.g., Liberman et al., 1962; Massaro, 1987; McClelland & Elman, 1986).

Although this is the modal view, we claim that this balkanization of speech perception is part of the reason that adequate theories of speech perception have not emerged (see Nusbaum & Henly, in press). This kind of dissociation of spoken language understanding into separate perceptual problems may be a result of the fundamental approach that most theories have taken to speech perception. As we noted in the previous discussion of the general problem of lack of invariance, most theories have focused on an analysis of the kinds of knowledge or information representations needed by a perceiver to achieve phonetic constancy, even though this kind of content analysis or informational analysis cannot be expected to yield an effective account of phonetic constancy over variability in phonetic context without a consideration of the required computational control mechanisms. If a theory focuses only on the information that is relevant to determining a talker's vocal characteristics or relative speaking rate or identifying phonetic context, this will definitely make the effects of talker variability, context variability, and rate variability look like different kinds of perceptual problems. But a consideration of the computational structure of the problem leads to a different conclusion.

Because talker variability results in a one-to-many mapping between acoustic cues and phonetic categories, talker variability presents the same kind of nondeterministic computational problem that arises because of variation in phonetic context. This has two immediate implications. First, it is possible that a common computational architecture may mediate phonetic constancy resulting from either of these sources of variability. Indeed, it seems plausible to look for a general computational mechanism that could account for phonetic constancy as a general approach to coping with all forms of lawful variability (Elman & McClelland, 1986). Second, if talker variability results in a nondeterministic computational problem, as noted earlier, normalization cannot be accounted for by passive transformations, even if it may appear that way from some of the simple computational models and restricted analyses carried out (e.g., Gerstman, 1968; Syrdal & Gopal, 1986). These models take the simplest case possible and, because they are restricted to steady state vowels, may not be reflective of the entire scope of the talker normalization problem. Steady state vowels are seldom found in fluent speech where vowels are coarticulated into consonant contexts. However, even when constraining the problem of talker normalization to vowel space differences, these models are still not as accurate as human listeners (e.g., Syrdal & Gopal, 1986). By addressing only the problem of vocal tract scaling (Fant, 1973), these models cannot really address the problem of consonant perception. Vocal tract size differences will definitely affect the acoustic patterns of consonants but probably not in the simple way it does for steady state vowels. However, the speech of talkers differs in more than just the effects of differences in vocal tract size. Two talkers may use different cues, cue combinations, and coarticulation functions to express consonants (e.g., Dorman, Studdert-Kennedy, & Raphael, 1977). These kinds of effects are compensated for by the listener (e.g., Johnson, 1991; Nusbaum & Morin, 1992; but see Rand, 1971), and the vocal tract scaling models give no indication about how this is accomplished.

At this point, the concept of "lawful variability," introduced by Elman and McClelland (1986) becomes germane. They argued that speech is restructured in speech production to present a lack of invariance (cf. Liberman, Cooper, Shankweiler et al., 1967) not through capricious or random processes, but through structurally regular processes that impose this variability in systematic ways. To recognize speech, it is important to deconvolve these sources of lawful variability. At this level of description, this is also similar to claims made by Fowler and Smith (1986) regarding a "vector analysis" of speech. The moment-by-moment output acoustic signal is the instantaneous result of the convolution (or by their terminology, vector sum) of a set of sources of variability, including the phonetic context, the talker's vocal characteristics, and the speaking rate. In order to recover the phonetic structure, it is necessary to undo all these effects, although there is no explicit statement of how to accomplish that.

Both of these descriptions of speech perception share something in common, although neither provides a clear description of this commonality nor a clear theoretical solution. Both seem to acknowledge that there is a basic structural similarity to the recognition problem posed by variability in phonetic context and vocal characteristics but do not articulate its nature. Our claim is that this structural similarity is a consequence of the basic nondeterministic nature of the relationship between acoustic cues and linguistic categories that is imposed by these sources of variability. Furthermore, if it is the case that these sources of variability impose a nondeterministic computational structure on the problem of perception, then it must follow that we can reject all theories that have an inherently passive control structure. In other words, it is our contention that phonetic constancy must be achieved by an active computational system (e.g., Nusbaum & Henly, in press; Nusbaum & Morin, 1992; Nusbaum & Schwab, 1986).

6.4 EMPIRICAL EVIDENCE FOR ACTIVE PROCESSING IN TALKER NORMALIZATION

Active control systems employ a feedback loop structure to systematically modify computation in order to converge on a single, stable interpretation (MacKay, 1951, 1956). By comparison, passive control structures represent invariant computational mappings between inputs and outputs. In consideration of this distinction, there are two general patterns of behavioral performance that can be taken as empirical evidence for the operation of an active control system (see Nusbaum & Henly, in press; Nusbaum & Schwab, 1986, for a discussion). First, evidence of load sensitivity in processing should provide an argument for active processing. There are several ways to justify this claim. For example, automatized processing in perception occurs when there is an invariant mapping between targets and responses, whereas controlled—load-sensitive—processing occurs when there is uncertainty regarding the identity of targets and distractors over trials or when there is no simple single feature difference to distinguish targets and distractors (e.g., Shiffrin & Schneider, 1977; Treisman & Gelade, 1980). In other words, when there are multiple possible interpretations of a stimulus pattern, processing shows load sensitivity, which may be manifest as an increase in processing time, a decrease in recognition accuracy, or an interaction with an independent manipulation of cognitive load (Navon & Gopher, 1979) such as a digit preload task (e.g., Baddeley, 1986; Logan, 1979). Logically, if there are multiple interpretations for a particular pattern, these alternatives must be stored and processed simultaneously increasing possible memory load or processed in some serial fashion leading to an average increase in response time compared to conditions without this uncertainty.

Second, active processing is indicated by the appearance of processing flexibility as demonstrated by the effects of listener expectations, context effects, learning, or other forms of on-line strategic processing. Although an active process need not demonstrate this kind of flexibility, a passive process by virtue of its invariant computational mapping certainly cannot. This means, for example, that evidence for the effects of higher order linguistic knowledge on a lower level perceptual task, such as lexical influence on phonetic recognition (e.g., Ganong, 1980; Nusbaum & Henly, in press; Samuel, 1986), should implicate an active control system in processing.

There is definitely a great deal of evidence arguing that speech perception is load sensitive under conditions of talker variability. For example, the accuracy of word recognition in noise and word recall is reduced when there is talker variability (speech produced by several talkers) compared to a condition in which a single talker produced the speech (Creelman, 1957; Martin, Mullennix, Pisoni, & Summers, 1989; Mullennix, Pisoni, & Martin, 1989). Talker variability also slows recognition time for vowels, consonants, and spoken words in a number of different experiments using a range of different paradigms (Mullennix & Pisoni, 1990; Nusbaum & Morin, 1992; Summerfield & Haggard, 1975). This provides some basic evidence that perception of speech is sensitive to talker variability, but does not really indicate why this occurs.

One possibility is that the increase in recognition time could reflect the addition of a separate talker normalization process to the recognition process (e.g., Nusbaum & Morin, 1992; Summerfield & Haggard, 1975). This hypothesis of inserting a passive normalizing mechanism is consistent with the general perspective of this research area as we outlined earlier. However, Nusbaum and Morin (1992) demonstrated that an independent manipulation of load, using the digit preload method (Baddeley, 1986; Logan, 1979) produces an interaction with talker variability. When listeners hear speech produced by a single talker, recognition time is unaffected by increases in the size of this visually presented digit memory load. Talker variability by itself does slow recognition time (Nusbaum & Morin, 1992). However, when there is uncertainty about which talker could have produced the speech, increasing the memory load slows recognition time even more. There is no reason why adding a passive normalization stage would interact with the memory load. But if increasing talker variability increases the number of alternative interpretations that must be tested in order to recognize an utterance, we would expect this to interact with memory load (see Nusbaum & Morin, 1992).

These studies all indicate that talker variability increases the cognitive load of the listener and increases the difficulty of recognizing speech. On the one hand, as we argue above, this provides one kind of evidence supporting the idea that listeners recognize speech using an active computational mechanism to resolve the inherent nondeterminism presented by talker variability. On the other hand, there has been some research that implies that listeners may not engage in talker normalization during speech perception (Nygaard et al., 1994; Palmeri et al., 1993) based on two kinds of evidence: the first claim is that increasing the amount of talker variability (in terms of number of talkers) beyond simply having two talkers does not affect performance in a memory task (Palmeri et al., 1993). The second claim is that information about the talker is encoded into the episodic trace representing a spoken item in memory. This is considered critical because Nygaard et al. (1994) and Palmeri et al. (1993) assume that talker normalization must strip talker-specific information out of an utterance prior to—indeed as a precondition to—its recognition. Therefore, this information could not be present in an episodic trace of the item in a memory study.

Neither of these arguments presents a strong challenge to the claim that listeners engage in talker normalization, as Palmeri et al. acknowledge. The first issue is only significant if somehow the normalization process keeps track of how many talkers there are or somehow depends on variability defined over a very broad span of speech (i.e., many utterances). If talker normalization just uses information from the preceding talker to set a context of interpretation for a subsequent utterance or simply uses the presence of any talker variability to change the strategies used in recognition, the first part of this argument will not hold. Furthermore, the second argument depends on an overly simplistic view of normalization and of the human auditory system. If talker normalization is needed to address a nondeterministic mapping between acoustic properties and linguistic categories, it cannot operate as a passive filtering mechanism, as implied by Palmeri et al. and Nygaard et al. Instead, it must actively test hypotheses about the identity of an utterance using various sources of information about the talker's vocal characteristics. These could be derived from context or from the utterance itself (Ainsworth, 1975). But this information does not need to modify the auditory representation of an utterance as a precursor to recognition of that information. Furthermore, there is sufficient neurophysiological evidence about parallel representations in the human auditory system (Pickles, 1982) that there is no reason to assume that there is only one representation available to the listener. But of even greater theoretical concern is the fact that if, as suggested by this view, there is no talker normalization, how does the listener recognize speech given talker variability? Although there is a claim about storing auditory traces of spoken words, if recognition is conceived of as only based on a comparison of a stimulus utterance to these traces (as implied by Nygaard et al., 1994), this is exactly the approach that is used by many computer speech recognition systems and is quite possibly the reason that they fail to provide human levels of recognition performance.

Our view is that the evidence regarding the load sensitivity of the human listener when there is talker variability provides strong evidence that speech perception is carried out by an active process. Furthermore, evidence of the flexibility of human listeners in processing speech given talker variability provides additional support. For example, we have found that listeners shift attention to different acoustic cues when there is a single talker and when there is talker variability (Nusbaum & Morin, 1992). In one condition, subjects monitored a sequence of spoken vowels for a specified target vowel and all the vowels were produced by one talker. In a second condition, a mix of different talkers produced the vowels. Both of these conditions were given with four different sets of vowels that were produced by linear predictive coding (LPC) resynthesis of natural vowels used in our other experiments (Nusbaum & Morin, 1992). One set consisted of intact, four-formant voiced vowels. A second set consisted of the same vowels with voicing turned off to produce whispered counterparts. A third set was produced by filtering all information above F2 using hand-tuned digital niters. The final set combined whispering with filtering to eliminate FO and formant information above F2.

If listeners recognize vowels using a mechanism similar to the one described by Syrdal and Gopal (1986), fundamental frequency and F3 information (although see Johnson, 1989, 1990a) should be necessary for recognition under all circumstances, because in their view this information provides a talker-independent specification of vowel identity. This predicts that in both the single-talker and mixed-talker conditions, the intact voiced vowels should be recognized most accurately, with whispering or filtering reducing performance somewhat, and the combination reducing performance the most, because these modifications eliminate critical information for vowel recognition. Our results showed that in the single-talker condition, recognition performance was uniformly high across all four sets of stimuli. In the mixed-talker condition, however, accuracy dropped systematically as a function of the modifications of the stimuli, with the voiced, intact vowels recognized most accurately and the whispered, filtered vowels recognized least accurately (Nusbaum & Morin, 1992). If vowel recognition were carried out by a passive, talker-independent mechanism (e.g., Syrdal & Gopal, 1986), the same pattern of results should have been obtained in both the single-talker and mixed-talker conditions. The results we obtained suggest that listeners only direct attention to FO and F3 when there is talker variability (cf. Johnson, 1989, 1990b). This kind of strategic flexibility in recognition is strong evidence of an active mechanism. Furthermore, it suggests that the reason for the increase in cognitive load given talker variability may be that *the* listener must distribute attention over more cues in the signal than when there is a single talker.

More recently, we have found that listener expectations affect talker normalization processes as well. In an earlier study, we found that not all talker differences increase recognition time in a mixed-talker condition (Nusbaum & Morin, 1992; also see Johnson, 1990b). When the vowel spaces of talkers are sufficiently similar and their fundamental frequencies are similar, there may be no difference in recognizing targets when speech from these talkers is presented in separate blocks or in the same block of trials. Magnuson and Nusbaum (1993, 1994) carried out a study designed to investigate more specifically under what conditions talker variability increases recognition time. In this study, two sets of monosyllabic words were synthesized with two different mean FOs differing by 10 Hz. In one condition, a small passage was played to subjects in which two synthetic talkers, differing in FO by 10 Hz, have a short dialogue. In a second condition, another group of subjects heard a passage in which one synthetic talker used a 10-Hz pitch increment to accent certain words. Both groups then listened to exactly the same set of single-pitch and mixed-pitch recognition trials using the monosyllabic stimuli. The subjects who listened to the dialogue between two talkers showed longer recognition times when there was a mix of the two different F0s in a trial compared to trials that consisted of words produced at a single F0. By comparison, subjects who expected that the 10-Hz pitch difference was not a talker difference showed no difference in recognition times or accuracy between the single-pitch and mixed-pitch trials. This demonstrates two things: First, the effect of increased recognition time in trials with a mix of F0s cannot be attributed to a simple contrast effect (see Johnson, 1990a) because both groups received exactly the same stimuli. Instead, the increased recognition times in the mixed-pitch trials seem to reflect processing specific to the attribution of the pitch difference to a talker difference and not something about the pitches themselves. Second, and perhaps more important for the present argument, the listeners' expectations affected whether or not they showed any processing sensitivity to pitch variability. This kind of processing flexibility cannot be accounted for by a simple passive computational system and argues strongly for an active perceptual mechanism (Nusbaum & Schwab, 1986).

Thus there are two different kinds of evidence regarding the nature of the mechanism that mediates perception when there is talker variability. It is difficult to see how to link the observed load sensitivity under talker variability and the processing flexibility unless it is through a single computational architecture. When there is talker variability, recognition appears to be carried out by a computational system with an active control mechanism. One implication of this is that it rules out simple filtering models of talker normalization. It seems unlikely that there is some passive transformation of the signal that is carried out to render acoustic pattern information into some normalized or talker-independent form (e.g., Syrdal & Gopal, 1986). A second implication is that if talker normalization is accomplished by an active control system, perhaps this may provide a more general solution to the problem imposed by lawful variation. If we consider briefly how such an active system might operate, it could suggest whether such a generalization is possible. To do this it is important to consider some of operating principles that constrain this system.

6.5 TOWARD AN ACTIVE THEORY OF TALKER NORMALIZATION

First and foremost, our view is that talker normalization is carried out as a consequence of the normative process of speech perception. In other words, talker normalization is not carried out by a separate module or computational system, but is a consequence of the basic computational structure of the normal operations of speech perception. This stands in sharp contrast to most previous approaches to talker normalization, which emphasized the problem of computing talker vocal tract limits and scaling vowel spaces. It may be more productive to treat the processing of lawful variation as a single perceptual problem and focus on the commonalties rather than separate these problems based on the specific sources of information and knowledge needed to support normalization and recognition.

Second, the effects of talker variability on perceptual processing directly reflect the computational operations needed to achieve phonetic constancy. Increased recognition times and interactions of varying cognitive load with recognition reflect the increased processing demands on capacity that are incurred by talker variability. Talker variability increases the number of possible alternative interpretations of the signal, thereby increasing the processing demands on the listener. As a corollary of our first point, we predict that the same kinds of processing demands will be observed whenever there is any nondeterministic relationship between acoustic cues and linguistic categories during perceptual processing. Furthermore, even though there may be some relationship between the information used in talker identification and talker normalization, we claim that the perceptual effects of talker variability are not a consequence of talker identification processes competing with speech understanding.

Third, in order to achieve phonetic constancy, given a nondeterministic relationship between acoustic cues and perceptual categories, different sources of information and knowledge, beyond the immediate acoustic pattern to be recognized, must be brought to bear on the recognition problem. For example, if the Fl and F2 extracted from an utterance could have been intended as either of two different vowels given talker variability, information about the vocal tract that produced the vowels (e.g., from FO and F3) will be used to provide the context for interpretation. Whenever there is a one-to-many mapping between a particular acoustic pattern and linguistic categories, listeners will have to use information outside the specific pattern to resolve the uncertainty. This information could come from other parts of the signal, previous utterances, linguistic knowledge, or subsequent parts of the utterance.

In order to realize the kind of computational flexibility required for this approach, it is important to reconceptualize the basic process of speech perception. The standard view of speech perception is that phoneme recognition or auditory word recognition is a process of comparing auditory patterns extracted from an utterance with stored mental representations of pattern information associated with linguistic categories. Our view is that speech perception, as an active process, is basically a cognitive process as described by Neisser (1967) and is more akin to hypothesis testing than pattern matching (cf. Nusbaum & Schwab, 1986). Nusbaum and Henly (1992) have argued that linguistic categories need to be represented by structures that are much more flexible than have been previously proposed. They claimed that a particular linguistic category such as the phoneme *Pal* might be better represented by a theory of what a /b/ is. This view is an extension of Murphy and Medin's (1985) argument regarding more consciously processed, higher order categories. From this perspective, a theory is a set of statements that provide an explanation that accounts for membership in a category. Rather than view a theory as a set of explicit verbal statements, our view is that a theory representation of a linguistic category is an abstract, general specification regarding the identity and function of that linguistic category. Although this could be couched as a set of features, it is more reasonable to think of a theory as something that would generate a set of features given particular contextual constraints.

Recognizing a particular phoneme or word is a process of generating a set of candidate hypotheses regarding the classification of the pattern structure of an utterance. Conjectures about possible categories that could account for a section of utterance are proposed based on the prior context, listener expectations, and information in the signal. Given a set of alternative classifications for a stretch of signal information, the perceptual system may then carry out tests that are intended to diagnose the specific differences among the alternative classifications. Cognitive load increases as a function of the number of alternatives to be considered and the number of diagnostic tests that must be carried out.

By this view, phonetic constancy is the result of a process of testing hypotheses that have been tailored to distinguish between alternative linguistic interpretations of an utterance. An active control system mediates this process of hypothesis formation and testing. An abstract representation of linguistic categories in terms of theories provides the flexibility to apply diverse forms of evidence to this classification process, allowing the perceptual system to resolve the nondeterministic structure produced by talker variability. These components taken together form a complex inferential system that has much in common with conceptual classification (Murphy & Medin, 1985) and other cognitive processes.

6.6 SUMMARY AND CONCLUSION

Rather than view talker normalization as a separate process that is added onto the front end of speech perception when there is talker variability, we propose that talker normalization is a consequence of normal recognition operations carried out by an active control system. Although the sets of knowledge or the cues listeners attend to may differ for different forms of lawful variability, it is entirely possible that the same kind of processing is carried out for rate normalization and talker normalization and coping with the effects ofcoarticulation. An active computational architecture provides both a means of resolving the nondeterministic relationship between acoustic cues and linguistic categories and an account of the behavior data showing load sensitivity and strategic flexibility in recognition under conditions of talker variability.

Moreover, Nusbaum and Henly (in press) have argued that the problem of lack of invariance is not special to phoneme perception. Even above the level of acoustic-phonetic mapping, there is a lack of invariance at all levels of pattern-interpretation mappings. Within the realm of talker differences, for example, there are phonological differences in different talkers idiolects and dialects that must be normalized during word recognition. Specialized filtering mechanisms cannot begin to account for the way listeners resolve this aspect of the spoken language comprehension problem. However, the kind of active, hypothesis-testing approach we have outlined here makes clear predictions that under conditions of increased variability there will be increased cognitive load during comprehension. By taking a cognitive approach to the problem of talker normalization and speech perception, this framework offers the prospect of providing a general account for aspects of spoken language understanding that share the same kind of nondeterministic computational structure.

ACKNOWLEDGMENTS

We thank Terry Regier for helpful discussions regarding computational theory. Preparation of this manuscript was supported in part by a grant from the Division of Social Sciences at the University of Chicago.

REFERENCES

Abbs, J. H., & Sussman, H. M. (1971). Neurophysiological feature detectors and speech perception: A discussion of theoretical implications. *Journal of Speech and Hearing Research, 14,* 23-36.

Ainsworth, W. (1975). Intrinsic and extrinsic factors in vowel judgments. In G. Fant & M. Taiham (Eds.), *Auditory analysis and perception of speech* (pp. 103-113). London: Academic Press.

Baddeley, A. D. (1986). *Working memory.* Oxford: Oxford Science Publications.

Barlow, H. B. (1972). Single units and sensation: A neuron doctrine for perceptual psychology? *Perception, I,* 371-394.

Best, C. T. (1994). The emergence of native-language phonological influences in infants: A perceptual **assimilation model. In J. C.** Goodman & H. C. Nusbaum (Eds.), *The **development of speech** perception: The transition from speech sounds to spoken words* (pp. 167-224). Cambridge, **MA: MIT Press.**

Chamiak, E. (1993). *Statistical language learning.* Cambridge, MA: **MIT** Press.

Chomsky, N. (1957). *Syntactic structures.* The Hague: Mouton.

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

Delattre, P. C., Libennan, A. M., & Cooper, F. S. (1955). Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society a/America, 27,*769-773.

Donnan. M. F., Studdert-Kennedy, M., & Raphael, L. J. (1977). Stop consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues. ***Perception &** Psychophysics, 22.* 109-122.

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 processes* (pp. 360-380). Hillsdale, NJ: **Eribaum.**

Fant, G. (1973). *Speech sounds andfeatures.* Cambridge, MA: MIT Press.

Fowler, C. A. (1989). Real objects of speech perception: A commentary on Diehl and Kluender. *Ecological Psychology. I.* 145-160.

Fowler, C. A., & Smith, M. R. (1986). Speech perception as vector analysis: An approach to the problem of invariance and **segmentation.** In J. S. Perkell & D. H. Klatt (Eds.). *Invariance and variability in speech processes* (pp. 123-135). Hillsdale, NJ: Eribaum.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology. 6.* 110-125.

Gerstman, L. J. (1968). Classification of self-normalized vowels. *IEEE Transactions on Audio **Electroacoustics, AU-16,*** **78-80.**

Gross, M. (1972). *Mathematical models in linguistics.* Englewood Cliffs, NJ: Prentice-Hall.

Grossberg, S. (1986). The adaptive self-organization of serial order in behavior: speech, language, and motor control. In E. C. Schwab & H. C. Nusbaum (Eds.), *Pattern recognition by humans and machines: Vol. 1. Speech perception* (pp. 187-294). San Diego, CA: Academic Press.

Hopcroft, J. E., & Ullman, J. D. (1969). *Formal languages and their relation to automata.* Reading, MA: Addison-Wesley.

Johnson, K. (1989). Higher fonnant normalization results **from** integration of F2 and F3. *Perception & Psychophysics. 46,* 174-180.

Johnson, K. (1990a). Contrast and normalization in vowel perception. *Journal of Phonetics, 18,*229-254.

Johnson, K. (1990b). The role of perceived speaker identity in FO normalization of vowels. *Journal of the Acoustical Society of America, 88.* 642-654.

Johnson, K. (1991). Differential effects of speaker and vowel variability on fricative perceptions. *Language and Speech, 34,* 265-279.

Klatt, D. H. (1977). Review of die ARPA speech understanding project. *Journal of the Acoustical Society of America, 62,* 1345-1366.

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). Comment. In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and variability* ***in*** *speech processes* (pp. 381-382). Hillsdale, NJ: Eribaum.

Levinson, S. E., Rabiner, L. R., & Sondhi, M. M. (1983). An introduction to the application of the theory of probabilistic functions of a Markov process **to** automatic speech recognition. *Bell System Technical Journal, 62,* 1035-1074.

Liberman, A. M., Cooper, F. S., Harris, K. S., & MacNeilage, P. F. (1962). A *motor theory of speech perception. Proceedings of the speech communication seminar* (Vol. 2). Stockholm: Royal Institute of Technology.

Liberman, A. M., Cooper, F. S., Harris, K. S., MacNeilage, P. E, & Studdert-Kennedy, M. (1967). Some observations on a model for speech perception. In W. Wathen-Dunn (Ed.), *Modelsfor the perception ofspeech and visualform* (pp. 68-87). Cambridge, MA: MIT Press.

Liberman, A. M.. Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74,* 431-461.

Liberman, A. M., Delattre, P. C., & Cooper, F. S. (1952). The role of selected stimulus variables in the perception of unvoiced stop consonants. *American Journal of Psychology, 65,* 497-516.

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

Logan, G. D. (1979). On the use of a concurrent memory load to measure attention and automaticity. *Journal of Experimental Psychology: Human Perception and Performance, 5,* 189-207.

Lowerre, B., & Reddy, R. (1980). The Harpy speech understanding system. In W. A. Lea (Ed.), *Trends in speech recognition* (pp. 340-360). Englewood Cliffs, NJ: Prentice-Hall.

MacKay, D. M. (1951). Mindlike behavior in artefacts. *British Journalfor the Philosophy of Science, 2,* 105-121.

MacKay, D. M. (1956). The epistemological problem for automata. In C. E. Shannon & J. McCarthy (Eds.), *Automata studies.* Princeton, NJ: Princeton University Press.

MacNeilage, P. E (1970). Motor control of serial ordering of speech. *Psychological Review. 77.* 182-196.

Magnuson, J. S.. & Nusbaum, H. C. (1993). *Talker differences andperceptual* **normalization.** *Journal of the Acoustical Society of America. 93,* 2371.

Magnuson, **J. S.,** & Nusbaum, H. C. (1994). Some acoustic and nonacoustic conditions that produce talker normalization. *Proceedings of the Acoustical Society of Japan,* pp. 637-638.

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.

Massaro, D. W. (1972). Perceptual images, processing time, and perceptual units. *Psychological Review. 79,* 124-145.

Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry.* Hillsdale. NJ: Eribaum.

Massaro, D. W., & Oden, G. C. (1980). Speech perception: A framework for research and theory. In N. J. Lass (Ed.), *Speech and language: Advances in basic research and practice* (Vol. 3, pp. 129-165). New York: Academic Press.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18.* 1-86.

Miller, G. A. (1962). Decision units **in** the perception of speech. *IRE Transactions on Information Theory, IT-8,* 81-83.

Miller, J. L., & Dexter, E. R. (1988). Effects of speaking rate and lexical status on phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance, 14.* 369-378.

Minsky, M. (1975). A framework for representing knowledge. In P. H. Winston (Ed.), *The psychology of computer vision* (pp. 211-277). New York: McGraw-Hill.

Mullennix, J. W, & Pisoni, D. B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics, 47,* 379-380.

Mullennix, J. W., Pisoni. D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America, 85,* 365-378.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review. 92.* 289-316.

Navon, D.. & Gopher, D. (1979). On the economy of the human-processing system. *Psychological Review, 86,* 214-255.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustic Society of America, 85,* 2088-2113.

Neisser, U. (1967). *Cognitive psychology.* New York: Appleton-Century-Crofts.

Newell, A. (1975). A tutorial on speech understanding systems. In R. D. Reddy (Ed.), *Speech* ***recog-*** *nition* (pp. 4-54). New York: Academic Press.

Nusbaum, H. C., DeGroot, J., & Lee, L. (1995). Using speech recognition systems: Issues in cognitive engineering. In A. K. Syrdal, R. W. Bennett, & S. L. Greenspan (Eds.), *Applied speech technology* (pp. 127-194). Boca Raton, FL: CRC Press.

Nusbaum, H. C., & Henly, A. S. (1992). Listening to speech through an adaptive window of analysis. In B. Schouten (Ed.), ***The processing*** *of speech: From the auditory periphery to word recognition* (pp. 339-348). Berlin: Mouton-de Gruyter.

Nusbaum, H. C., & Henly, A. S. (in press). Understanding speech perception from the perspective of cognitive psychology. In J. Charles-Luce, P. A. Luce, & J. R. Sawusch (Eds.), *Theories in spoken language: Perception, production, and development.* Norwood, NJ: Ablex.

Nusbaum, H. C., & Morin, T. M. (1992). Paying attention to differences among talkers. In Y. Tohkura, Y. Sagisaka, & E. Vatikiotis-Bateson (Eds.), *Speech perception, production, and linguistic structure* (pp. 113-134). Tokyo: Ohmsha Publishing.

Nusbaum, H. **C., &** Pisoni, D. B. (1987). Automatic measurement of speech recognition performance: A comparison of six speaker-dependent recognition devices. *Computer Speech and Language, 2.* 87-108.

Nusbaum, H. C., & Schwab, E. C. (1986). The role of attention and active processing in speech perception. In E. C. Schwab & H. C. Nusbaum (Eds.), *Pattern recognition by humans and machines: Vol. 1. Speech perception* (pp. 113-157). San Diego, CA: Academic Press.

Nygaard, L. C., Sommers, M, S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science, 5,* 42-46.

Ohman, S. E. G. (1966). Coarticulation in VCV utterances: Spectrographic measurements. *Journal of the Acoustical Society of America, 39,* 151-168.

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken word. *Journal of Experimental Psychology: Learning, Memory, . and Cognition, 19,* 309-328.

Peterson, G., & Barney. H. (1952). Control methods used **in a** study of the vowels. *Journal of the Acoustical Society of America, 24,* 175-184.

Pickles, J. 0. (1982). *An introduction to the physiology of hearing.* London: Academic Press.

Pisoni, D. B. (1978). Speech perception. In W. K. Estes (Ed.), *Handbook of learning and cognitive processes: Vol. 6. Linguistic junctions in cognitive theory* (pp. 167-234). Hillsdale, NJ: Eribaum.

Rand, T. C. (1971). *Vocal tract size normalization in the perception of stop consonants.* Paper presented at the 81st meeting of the Acoustical Society of America, Washington, DC.

Roberts, L. G. (1965). Machine perception of three-dimensional solids. In J. T. Tippett (Ed.), *Optical and electro-optical information processing* (pp. 159-197). Cambridge, MA: MIT Press.

Samuel, A. G. (1986). The role of the lexicon in speech perception. In E. C. Schwab & H. C. Nusbaum (Eds.), *Pattern recognition by humans and machines: Vol. 1. Speech perception* (pp. 89-112). San Diego. CA: Academic Press.

Shankweiler, D., Strange, W. & Verbnigge, R. (1977). Speech and the problem of perceptual constancy. In R. Shaw & J. Bransford (Eds.), *Perceiving, acting, and knowing* (pp. 315-345). Hillsdale, NJ: Eribaum.

Shiffrin, R. M.. & Schneider, W. (1977). Controlled and automatic **human** information processing: n. Perceptual learning, automatic attending and a general theory. *Psychological Review, 84,* 127-190.

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.

Stevens, K. N., & Blumstein. S. E. (1981). The search for invariant acoustic correlates of phonetic features. In P. D. Eimas & J. Miller (Eds.). *Perspectives on the study of speech* (pp. 1-38). Hillsdale, NJ: Eribaum.

Stevens, K. N., & Halle, M. (1967). Remarks on analysis by synthesis and distinctive features. In W. Walthen-Dunn (Ed.), *Models for the perception of speech and visual form* (pp. 88-102). Cambridge, MA: MIT Press.

Studdert-Kennedy, M.. Libennan. A. M., Hams, K. S., & Cooper, F. S. (1970). Motor theory of speech perception: A reply to Lane's critical review. *Psychological Review, 77,*234-249.

Summerfield, Q., & Haggard, M. (1975). Vocal tract normalization as demonstrated by reaction times. In G. Fant & M. Tatham (Eds.), *Auditory analysis and perception of speech* (pp. 115-141). London: Academic Press.

Syrdal, A. K., & Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *Journal of the Acoustic Society of America, 79,* 1086-1100.

Treisman, A. M., & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology, 12,* **97-136.**

Uhr, L. (1973). *Pattern recognition, learning, and thought: Computer-programmed models of higher mental processes.* Englewood Cliffs, NJ: Prentice-Hall.

Walley, A. C., & Carrell, T. D. (1983). Onset spectra and fonnant transitions in the adult's and child's perception of place of articulation in stop consonants. *Journal of the Acoustical Society of America. 73,* 1011-1022.

Wickelgren, W. A. (1969). Auditory or articulaiory coding in verbal short-term memory. *Psychological Review. 76,232-235.*

Wilpon, J. G. (1985). A study on the ability to automatically recognize telephone-quality speech from large customer populations. *AT&T Technical Journal, 64,* 423-451.

Woods, W. A. (1973). An experimental parsing system for transition network grammars. In R. Rustin (Ed.), *Natural language processing* (pp. 111-145). New York: Algorithmics Press.