

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 22 (1998)
Indiana University

**Speech Perception and Spoken Word Recognition:
Research and Theory¹**

Miranda Cleary and David B. Pisoni²

*Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405*

¹ This work was supported by NIH NIDCD Research Grant DC00111 and Training Grant DC00012 to Indiana University.

² Also DeVault Otologic Research Laboratory, Department of Otolaryngology-Head and Neck Surgery, Indiana University School of Medicine, Indianapolis, IN.

Speech Perception and Spoken Word Recognition: Research and Theory

Introduction

The study of speech perception investigates how we are able to identify in the human voice the meaningful patterns that define spoken language. Research in this area has traditionally focused on what parameters in the acoustic signal influence the sensation of particular speech sounds, specific consonants or vowels, for example. This work has typically been couched in terms of how humans perceive minimal linguistic contrasts—how we distinguish “pat” from “bat” or “bit” from “bet” for example. In short, speech perception has traditionally been the study of phoneme perception.³

The study of speech perception also needs, however, to account for how we identify spoken words and comprehend sentences in *connected fluent* speech. Defining “speech perception” very narrowly in terms of phoneme perception or nonsense syllable identification was a useful and reasonable simplification in the early days of the field, but the drawbacks of couching the problem purely in these terms have become increasingly apparent in recent years.

The study of speech perception has therefore undergone some important theoretical shifts over the last fifty years. Many of the fundamental assumptions of the area’s first heyday in the late 1940’s have been slowly jettisoned one by one, or have undergone significant modification. This chapter reviews key findings and ideas contributing to the current state of the field. With an aim towards stimulating further research, we will indicate areas we feel have already been well (if not overly) “mined” for their secrets, and what we currently see as the most potentially useful and unexplored avenues of future investigation on speech sound perception and spoken language processing.

Speech Perception in the Context of the Other Senses

Speech perception stands to audition much as face perception stands to vision, in that it is a differentially developed response to a particular type of physical signal within a more general sensory modality. Since speech perception is the study of how we categorize the range of vocalizations characteristic of our own species, the key issues cannot be addressed through the use of animal models; the relevant data come, almost exclusively, from external observations of human behavior. The relevance of how other species perceive human speech signals is doubtful (see Trout, 1998), although to the extent that such research helps provide a psychophysical account of the early processing activity that takes place at the auditory periphery, it can have some limited value (see Kluender, 1994 for a review).

Speech is only one type of auditory input of interest to humans. Perceiving speech is an ability arguably lying somewhat intermediate between the fundamentally pragmatic skill of perceiving meaningful environmental sounds (where the sound-meaning correspondence is relatively static across members of the species), and the somewhat more mysterious perceptual capacity we have to store, process, and appreciate musical input (of rather more ambiguous semantic content). Certain acoustic

³ A “phoneme” is an abstract linguistic entity referring to the smallest unit of sound that can distinguish one meaningful word from another in a particular language. What constitutes a phoneme differs from language to language. Phoneme symbols are convenient shorthand for particular combinations of acoustic feature attributes—the attributes judged to be most critical to identification being simplified to an abstract binary labeling of either present (+) or absent (-). Specific combinations of these features are categorized as instances of a particular phoneme. The term “phone” is sometimes used to refer to one possible acoustic instantiation of a phoneme.

characteristics, however, tend to differentiate spoken language from other auditory signals. It therefore makes sense to review these before proceeding to a discussion of the key theoretical issues.

The Acoustic Speech Signal

In speech production, the lungs can be crudely described as a sophisticated pair of bellows with the moving larynx, tongue, lips, and soft palate serving to constrain the air-stream in certain characteristic ways. The articulators impose continually shifting filtering functions atop temporally quasi-periodic or aperiodic sound sources, causing the acoustic energy to be attenuated at certain frequencies and amplified at others.

Fluent connected speech can be grossly characterized by its average amplitude. A simple speech waveform of a female speaker saying, “You are not walking in my street,” plotting intensity as a function of time, can be seen in Figure 1 (b). The highlighted section of (b) is magnified in (a). The intensity level for conversational speech usually averages around 60-65 dB SPL in a quiet environment. A whisper drops to about 30 dB SPL and loud speech close to shouting approaches 75-80 dB SPL.

It is possible to analyze complex signals such as the speech waveform as the summation of a large number of simple sinusoidal frequency components. A two-dimensional plot showing the relative amplitude of these components over some small window of time is called a power spectra. A single power spectra from Point X in Figure 1 (b) is shown in Figure 1 (f). A third dimension representing time can be added to this 2-D frequency vs. amplitude representation by computing many power spectra consecutively over time. A spectrogram provides a visual three-dimensional representation of energy (indicated by darkness of shading) at a range of frequencies (Y-axis) as a function of time (X-axis). Essentially, it is a temporally ordered “stack” of numerous power spectra viewed from a “vantage point” above their peaks.

Wide-band spectrograms (as in Figure 1 (c)) are used when temporal fidelity is important for analysis. Rapid frequency changes in the signal can therefore be observed, but a wide band spectrogram gives up in the frequency domain what it gains in the temporal domain. Actual intensities at different frequencies are averaged by the analysis, yielding visually enhanced groups of harmonics “smudged-together” in a band that is termed a “formant”—essentially an energy concentration centered around a particular frequency. A narrow-band spectrogram, on the other hand, provides more detailed information about the relative strength of individual horizontally-running harmonics rising and falling in time (Figure 1 (d)).

 Insert Figure 1 about here

Laryngeal Source Characteristics

The complex acoustic patterns in speech arise from the anatomy and physiology of the vocal tract and the physics of sound (Denes & Pinson, 1993; Flanagan, 1972; Rossing, 1990). The most basic characteristic of a human voice is its fundamental frequency, symbolized as “ f_0 ”, which closely corresponds to the perceived pitch of a speaker’s voice (perception of a “deep” versus “shrill” voice, for example). Determined by the rate at which the muscular vocal folds within the larynx flap open and shut, f_0 is the lowest harmonic component of the complex periodic sound made when air presses through these folds. In the wide-band spectrogram shown in Figure 1 (c), each of the semi-regularly spaced vertical striations corresponds to one opening and closing cycle of the vocal folds. Narrow-band spectrograms

illustrate f_0 via the spacing of the harmonics. Widely spaced harmonics indicate high pitch, whereas closely spaced harmonics indicate lower pitch. The rate of vocal fold vibration can be influenced by

many factors, but the mass and flexibility of the folds largely constrains this value so as to produce the pitch ranges associated with males, females, and children. (The average f_0 s for men, women and children are approximately 125 Hz, 200+ Hz, and 300+ Hz, respectively.) Individual speakers do, however, have some degree of muscular control of the tension and position of the folds. During fluent speech, the vocal cords are vibrating about seventy percent of the time (Fry, 1979).

Changes in f_0 play a major role in perception of intonation, phonemic tone, and word stress. Declination is the term used to refer to the gradual (non-monotonic) decrease in f_0 from the beginning to the end of a declarative utterance. The “pitch-track” in Figure 1 (e) shows the f_0 of this female voice, for example, dropping from a value of about 320 Hz to 250 Hz. Variations on this general pattern, such as inserting a prominent pitch rise, are used for different purposes; to indicate an interrogative, for instance. Intonation patterns consisting of f_0 peaks and valleys, typically carry semantic import, and appear to be “scaled” by speakers when applied to utterances or syntactic phrases of various durations.

Relative f_0 levels (or “tones”) are used contrastively by a number of so-called tone languages (such as Mandarin Chinese and many of the Bantu languages spoken in Africa), such that words may be distinguished in meaning only by the pitch or pitch contour of the vowel. The perception of word or syllable stress in a language such as English results from a combination of adjusted pitch (higher), amplitude (louder), and duration (longer). Stress, like tone, is defined quantitatively in relative terms, depending on its surrounding speech context.

The timing of the onset of vocal cord vibration can also serve to distinguish one spoken word from another. The delay between a noise-burst caused by sharply releasing a set of closed articulators (such as the lips) to the first vocal fold vibration following, such as occurs in producing a syllable like “pah,” is used in many languages to contrast consonant sounds that are otherwise alike in how they are produced in front of vowels. This difference in “voice onset time” distinguishes, for example, early “voiced” (short VOT) sounds like [b] and late-voiced (long VOT) “voiceless” sounds like [p]. There are also “pre-voiced” sounds in the world’s languages for which vocal cord activity starts at, or before, the release of closure.

Sources and Filters Above the Vocal Cords

As can be seen in Figure 2 (a), the human vocal tract is a cavity of rather irregular dimensions. While vocal tract modeling has become increasingly sophisticated in recent years, highly simplified tube models of the human vocal tract can account for some of its general resonance characteristics and interactions. Early work by Chiba and Kajiyama (1941), Fant (1960), and Stevens (Stevens & House; 1955) for example, modeled the human vocal tract during vowel production as a pair of simple tube resonators, one representing the front oral cavity and one representing the back pharynx cavity. For nasalized sounds, like the English [m] and [n], involving the lowering of the soft palate, the added large nasal cavity requires the model be modified to a “pronged” set of three main cavities. Equations for calculating hollow tube resonances and explanations of how various articulatory constrictions are modeled can be found in most introductory acoustic phonetics texts (e.g., Johnson, 1997; Kent, 1997; Ladefoged, 1996).

Insert Figure 2 about here

Coarticulation. In fluent speech, as Figure 2 illustrates, (“You lose your yellow roses early”), no tidy one-to-one mapping necessarily exists between the percepts we arrive at and the acoustic signal. Even word boundaries are often not well defined physically, (as one may realize from hearing a foreign language spoken conversationally). The non-linear nature of this relationship is illustrated in that while a certain sound X may be heard as preceding sound Y, the acoustic characteristics of X do not necessarily leave off where the acoustics of sound Y begin. For example in Figure 2, the “z” sound in “lose” clearly shows the influence of the “y” sound in “your” that immediately follows. This “spreading” influence of certain articulations, known technically as “coarticulation,” can last many milliseconds past the perceived temporal center of a particular speech sound. Since the combinatorial possibilities of one sound following another are enormous, this means that a particular percept exists in a one-to-many relation with its acoustic expressions in the physical signal.

Discrete perceptual units like consonants and vowels are linguistic abstractions that do not exist in the raw acoustic signal. Some theorists have argued that phonemes emerge as perceptual categories as a result of the type of training involved in learning to read and write (see Liberman, Shankweiler, Fischer & Carter, 1974). More commonplace, however, is the argument that there are, in fact, “low level” basic perceptual categories that have an independent existence regardless of meta-linguistic training (Liberman, Cooper, Shankweiler & Studdert-Kennedy, 1967; Nearey, 1990). Most models of speech perception and spoken word recognition assume the extraction of some intermediate representational unit such as the phoneme. The process by which such units are “segmented” out of the continuous signal must therefore then be specified. Several other alternative units of perceptual analysis have been proposed, including, the syllable, context-sensitive units such as the diphone⁴ or triphone⁵ or even entire word forms. The issue of what unit of analysis is “most fundamental” is often hotly debated, with opposing camps each citing a different body of experimental evidence. With speech perception being a temporally unfolding process however, likely involving hierarchically related mental representations of the input being constructed “on-line,” the “unit” upon which a particular experimental response is based probably depends on the precise nature of the information processing task assigned to the human perceiver.

Even so, for many years, the predominant focus of study has been on phoneme-level units of perception. Much effort has been expended over the last fifty years in trying to find invariant correlates of phonemic categories in the acoustic signal. This search has not yielded the necessary and sufficient predictors for the percept of individual phonemes that was originally its goal. It has, however, told us a good deal about the acoustic tendencies of larger classes of sounds (e.g., vowels vs. consonants), and about the acoustic information that can contribute to (not determine) the perception of individual phonemes under various circumstances. We review some of these findings below.

Articulatory and Acoustic Correlates of Vowel Perception. In wide-band spectrograms of slowly, carefully articulated speech, vowels are the most visually prominent feature, usually marked clearly by a sequence of vertical voicing striations and characterized by horizontally-running formants (see Figure 1 (c)). Formants are not harmonics of f_0 . Rather, formants are defined by the enhanced intensity of certain harmonics and the damping or attenuation of other harmonics due to the natural resonance characteristics of the vocal tract.⁶ Sloping formants indicate changes in resonant frequencies due to articulator movement. At any one point in time, the vowel signal typically has several sets of such enhanced harmonics, and each set of enhanced harmonics defines a single formant.

⁴ Defined by from the approximate center of one phone [see footnote 3] to the midpoint of the following phone.

⁵ Defined from the midpoint of the preceding phone, through the target phone, to the midpoint of the following phone.

⁶ The lowest frequency formant is typically referred to as “F1”, the next highest formant as “F2” and so on.

The relative positioning of the two or three lowest frequency formants over time helps to perceptually distinguish the different vowel sounds. (In Figure 1 (c), compare the “ee” sound in “street” to the “oo” in “You” for example.) Low frequency first formants tend to be caused by the raising or arching of the tongue up towards the roof of the mouth. The front/back location of this tongue raising in the oral cavity roughly correlates with the second formant value, with lower frequency F2s associated with backness. Relative duration is also used to distinguish different vowels. Often, the brief vowels found in rapid speech or unstressed syllables, are said to “lose their spectral quality” in the sense that their formant values cannot change rapidly enough to reach the frequency values typical of longer instances, and thus are less characteristic, or “undershot” (Fant, 1960). While this might be seen as problematic for discriminating the different vowels, it has been shown that some information about vowel identity is present even in the rapid formant frequency changes known as “transitions” that initiate and conclude a vowel sound flanked by consonants (Strange, Jenkins, & Johnson, 1983).

Every speaker uses some subset of possible formant frequencies to produce recognizable vowels. The bounds of this acoustic space are defined by the physiology of the speaker’s vocal tract and translate into a “vowel space” whose edges are defined by the most extreme values of the first two or three formants in the speaker’s vowel set (Peterson & Barney, 1952; Gerstman, 1968). Perceptually, the vowels produced in these extreme regions correspond to the so-called “point vowels”—the [u] as in “boot”, the [i] as in “beet” and the [A] as in “bah!” Languages differ on how many regions within this space they use to represent different vowels, and in where these regions lie in the space. A rough characterization of a typical vowel space for a male speaker of American English is given in Figure 3 (b) (For a fuller characterization of American English vowels, see Hillenbrand, Getty, Clark, and Wheeler, 1995; Peterson and Barney, 1952.)

 Insert Figure 3 about here

Articulatory and Acoustic Correlates of Consonant Perception. A good deal of what distinguishes the speech signal goes on at frequencies higher than those typical of vowel F1s and F2s. The human ear’s receptive frequency range of approximately 20 to 20,000 Hz boasts its best discrimination in the interval of about 1500 to 6000 Hz. Most of the acoustic cues that differentiate the consonants lie in this range, and although their intensity is weak in absolute terms compared to that of vowels, the auditory system of the human listener—with its increased sensitivity to intensities in this range—partially compensates for this fact. It should also be noted, however, that the frequency resolution ability of the human ear (see Moore, this volume) is reduced in these higher frequencies.

Consonants in English can be roughly sorted along three major articulatory dimensions: (1) manner of articulation (very generally, how the articulators move), (2) place of articulation (where the most constriction occurs in the vocal tract), and (3) voicing (whether and when the vocal cords are vibrating during articulation).

Fricatives involve the making of a narrow but incomplete closure of the articulators and are distinguished phonetically by intervals of aperiodic noise at particular frequencies. The strident fricatives [σ], [Σ], [ζ], [Z] ⁷ have energy concentrated at the higher frequencies (higher for the alveolars than the palatals) and are of relatively long duration, while the non-strident [ϕ], [T], [Ϙ], [Δ] ⁸ have shorter

⁷ [s] as in “sip,” [Σ] as in “ship,” [z] as in “zip,” and [Z] as in “azure”

⁸ [f] as in “fin,” [T] as in “thin,” [v] as in “van,” and [Δ] as in “then”

durations, and are typically characterized by weaker energy spread over a wider frequency range.

Stops (also called plosives) involve making a complete vocal tract closure which is then abruptly released, allowing air to escape (as in [b], [d], and [g], for example). Thus, abrupt stretches of silence often indicate the presence of stop consonants in spectrograms. The release of a stop closure is obligatorily marked by very brief high frequency noise strongest at a frequency indicative of the place of articulation. The burst may be followed by additional “aspiration” noise generated at the larynx.

Somewhat more acoustically complex are the nasal stops [n], [m], and [ŋ].⁹ These are produced with the bulk of the air traveling from the lungs up into the nasal cavity via a side-branch route channeled off by lowering the velum, a flap of tissue at the extreme back of the mouth. Because the nasal cavity is relatively soft and large, the bandwidth of the nasal formants is wider than that of orally-produced (non-nasal) formants. The shape of the tract during nasal production also results in antiformants, frequency ranges where the lack of intensity is due to active interference and canceling-out of some frequencies by others. Nasals show voicing striations and weak higher formant structure, with low and relatively intense F1s.

Finally, the American English liquids [λ] and [ɹ] involve only partial constriction, permitting strong formant structure, and are characterized by a low first formant.¹⁰ These two sounds are distinguished from each other by a strong downwards drop in F3 for [ɹ], a drop which [λ] lacks. The glides [j] and [w]¹¹ share a resemblance to the liquids being obligatorily voiced and having relatively rapid spectral transitions. [w] tends to also show the influence of lip protrusion, resulting in a lowered F2.

Each of these “manner” classes (fricative vs. stop vs. nasal, etc.), contains sounds distinguished from each other only by place of stricture in the vocal tract. Perception of place contrasts has been a long-standing area of research and debate. For example, the bilabial (closure at the lips), alveolar (closure at the alveolar ridge), and velar (closure at the velum) stops of English share most of their first formant transition characteristics. They differ however in their second and third formant transitions, which contributes towards place discrimination. Early work by Delattre, Liberman & Cooper (1955) suggested that consonants possessed invariant “loci” or frequencies from which a formant transition would emerge (after a silent interval of closure in the case of stops and nasals) and interpolate toward the characteristic frequencies of the following sound, e.g., of a vowel. Unfortunately, the notion of a “locus equation” did not work well for velar consonants across all vowels. A similar problem was encountered trying to use a static “template” defined by an average short-term power spectra calculated across the release burst (if one was present) and the formant transitions directly following the stop release (in a CV syllable), or vowel offset (in a VC syllable) (Stevens & Blumstein, 1981). Once variation in surrounding context was taken into account, the utility of a single locus or characteristic burst frequency per place of articulation appeared to be diminished. In subsequent years, the search for invariant acoustic information regarding place of articulation turned to somewhat more complex measures of the rate and direction of spectral change over time (e.g., Kewley-Port, 1983; Kewley-Port, Pisoni, & Studdert-Kennedy, 1983; Kewley-Port & Luce, 1984). More recently, Sussman and colleagues have tried to revive a modified version of the locus equation by detailing strong linear relationships in CV syllables between the frequency at which the F2 transition begins and F2 values measured from the temporal center of particular vowels (see Sussman, & Shore, 1996; Sussman, McCaffrey, & Matthews, 1991; for cross-linguistic analysis, Sussman, Hoemeke & Ahmed, 1993). This correlation appears to be an informative though not infallible predictor

⁹ [ŋ] as in “sing,” “wrong” and “rang”

¹⁰ [λ] as in “light,” [ɹ] as in “right”

¹¹ [j] as in “yawn,” [w] as in “want”

of consonant place (but see Fowler, 1994; Brancazio & Fowler, 1998 for counter-arguments and discussion).

The type of findings surveyed above provide useful heuristics for predicting the phoneme category associated with a particular acoustic signal, but they do not constitute invariant acoustic “markers” unique to particular phonemes. Phonemes are really only perceived in the linguistic context in which they occur. They are linguistic abstractions, not physical entities.

Nevertheless, as mentioned in the introduction, speech perception was equated for many years with what we would term, “phoneme perception.” As an outgrowth of this approach, in the history of the field, there have been a handful of largely phoneme-based perceptual phenomena that have been quite extensively studied. While for completeness’ sake we will review a selection of these below, it can be argued that these particular perceptual effects have, in a sense, been exhausted in their usefulness, and are now distracting many researchers from some of the more pressing fundamental problems of fluent speech perception and spoken language processing. We will try to show what we consider to be the potentially adverse consequences of building a comprehensive theory of speech perception on a selective smorgasbord of relatively low-level perceptual effects without a unified framework that binds these separate levels together.

Findings From the Traditional Domain of Inquiry: Phoneme Perception

Categorical Perception

Categorical perception refers to the phenomenon of perceiving in terms of only a small set of discrete categories, items from some larger stimulus set whose members vary continuously along some manipulated acoustic parameter. In an influential paper, Liberman, Harris, Hoffman, and Griffith (1957) reported that listeners “categorically” perceived changes in the direction, frequency of origin, and duration of the F2 transitions of synthesized consonant-vowel syllables. A syllable best labeled as “bah” was reported for steeply rising F2 transitions, “dah” for short gradual F2 transitions, and “gah” for longer steeply falling F2 transitions. Sharp “cutoff points” for the manipulated values appeared to exist, above and below which, unequivocal labeling was favored. From individual performance on a identification/labeling task, parameter values for which the discrimination of acoustically similar stimuli would be nearly perfect, as well as intervals within which discrimination would approach chance, could be predicted (see Figure 4 (a)). While actual discrimination was somewhat better “within-category” than the identification task predicted, it was still markedly worse than would be expected from psychophysical data on human discrimination of different pure tone frequencies in the same range.

Insert Figure 4 about here

Parameters beside F2 transitions, such as voice-onset time, have also been manipulated to produce “categorical” responses (Liberman, 1970). Discrimination of vowels, however, does not show the same general pattern (Fry, Abramson, Eimas, & Liberman, 1962; Pisoni, 1973), although reducing the duration of the vowel (Pisoni, 1973) or placing vowels in dynamic context (Stevens, 1968; Sachs, 1969) generates discrimination functions that better resemble the data from the consonant studies.

A claim from this early work on categorical perception was that different perceptual mechanisms were used to process speech versus non-speech auditory stimuli. Comparisons were drawn between the fine discriminatory abilities humans have for tones or, in vision, for color, and the apparent phonetic

insensitivity within intervals that define native speech sounds (e.g., Eimas, 1963). During the 1970's and 80's, however, this interpretation of categorical perception came under serious scrutiny. New evidence suggested that within-category discrimination was not as poor as the strong version of categorical perception claimed, and that categorical perception effects appeared to be enhanced by the particular response format of the task (Fujisaki & Kawashima, 1968; Pisoni & Lazarus, 1974). The relevance of categorical perception to naturalistic speech perception was also debated (Crowder, 1989).

Categorical perception of voice-onset time differences has been explained as being partially due to peripheral auditory processing limitations on the temporal resolution of acoustic detail. Non-speech stimuli with roughly similar temporal design can also be discriminated categorically (Miller, Wier, Pastore, Kelly & Dooling, 1976; Pisoni, 1977). Non-human species such as the chinchilla (Kuhl & Miller, 1975), macaque (Kuhl & Padden, 1983), and Japanese quail (Kluender, Diehl, & Kileen, 1987) can also be trained to categorically discriminate the type of stimuli used in categorical perception studies, although they have not been shown to be able to identify these signals in the same manner that human listeners do.

Early work in categorical perception fueled an interest in how and when tendencies toward categorical discrimination of speech sounds develop in humans. Do they emerge as a result of experience with an ambient language, or do these tendencies require little in the way of post-natal experience to be expressed? Pioneering work by Eimas, Siqueland, Jusczyk, and Vigorito (1971) used a habituation-based high-amplitude sucking procedure to show that infants of four-months and even one-month of age responded to VOT differences of 20 ms straddling a crucial cutoff point of around 30 ms VOT, as if the stimuli were different. Twenty millisecond differences to one or the other side of this cutoff point were not treated as distinguishable. Whether or not infants are explicitly *identifying* phonetic categories in such a task, remains at issue, but unlike the non-human species mentioned above, human infants do not require hundreds of training trials to show discrimination (see Eimas, 1985).

Infant discrimination of many (but not all) contrasts used in adult language has been demonstrated. (For reviews see Aslin, Jusczyk & Pisoni, 1998; Jusczyk, 1997; Werker, 1992.) Sensitivity to sound contrasts not used in the child's particular linguistic environment appears to be attenuated or lost within the first year of life. American infants, for example, lose the distinction between pre-voiced and voiced categories (Aslin, Pisoni, Hennessy & Perey, 1981).

Categorical perception has also been explored in the context of perceiving or learning a foreign or a second language. It has long been known that listeners often show an apparent inability to perceptually discriminate speech tokens that resemble items of a single phonemic category in their native language but which would be of different phonemic categories in another language (e.g., Lisker & Abramson, 1967; Goto, 1971). The ability, for example, of adult monolingual Japanese speakers to perceive and produce the distinction in English between [♦] and [l], a contrast not utilized in Japanese, has attracted considerable attention over the years (Miyawaki, Strange, Verbrugge, Liberman, Jenkins & Fujimura, 1975). Although initial reports indicated that very little "re-acquisition" of this particular contrast was possible, training incorporating a wide variety of phonetic contexts and a selection of different speakers has demonstrated reliable improvements in this discrimination by adult Japanese monolinguals (Logan, Lively, & Pisoni, 1991; Lively, Logan & Pisoni, 1993; Bradlow, Pisoni, Akahane-Yamada, & Tohkura, 1997). Reacquisition success appears to also depend crucially on the particular contrast involved and its relation to an individual's native sound inventory (Best, McRoberts, & Sithole, 1988; Best, 1994).

Duplex Perception

Another laboratory phenomenon frequently cited as evidence for specialized mechanisms in speech perception is duplex perception (for a review, see Repp, 1982). Early investigators, including Broadbent (1955), reported that listeners perceived a single “fused” syllable when a synthetic first formant was presented to one ear and a synthetic second formant (or second and third formant) to the other. A variant of this (see Figure 4 (a)) presents only the frequency transition portion of the higher formant(s) to one ear and a complex “base” of the complete F1 and the steady-state portion of the higher formant(s) to the other (Rand, 1974). This manipulation yields reports of not one, but two percepts—a “fused” speech syllable in the ear that was played the syllable with the missing high frequency transition(s), and a non-speech “chirp” in the ear to which the lone transition was played.

It has been argued that a speech-specific route of perception leads to the fused syllable percept, while a separate “general auditory” process results in the non-speech percept. Fowler and Rosenblum (1990) reported, however, that duplex percepts could be elicited by using components analogously isolated from complex non-speech stimuli—the sound of a slamming door, for example. While Fowler & Rosenblum suggest that the duplex percept may still arise from a special system designed to attribute causal sources to familiar sounds (Gaver, 1993; Ballas, 1993), these results clearly call into question the idea that a “speech module” is necessarily responsible for the duplex perception results. Duplex-type effects using musical chords have also been reported (Pastore, Schmeckler, Rosenblum, & Szczesiul, 1983; Hall & Pastore, 1992). While these last authors argue for an explanation via a general perceptual mechanism, the issue of whether highly familiar sound complexes might be processed differently than less common ones remains unresolved.

Duplex perception is often discussed in the context of perceptual organization and auditory grouping mechanisms (e.g., Remez, Rubin, Berns, Pardo, & Lang, 1994). Perception of the duplex syllable component implies automatic integration of acoustic information from more than one source into a single speech percept. The perceptual mechanism seems to demonstrate a preference for integrating two streams of information as having a common speech-related source if at all possible. Stored information reflecting a statistical correlation between the past occurrences of similar frequency sweeps and similar base frequencies may favor this type of processing for familiar complex sounds. Next we discuss a different type of integration in speech perception, this time involving input across different modalities.

Visual Contributions to Phoneme Perception and the McGurk Effect

It has been known since the early days of telephone communications research that certain sounds are more susceptible to being misheard than others. Often the resulting confusions ([v] for [b], for example), tend to share acoustic attributes along dimensions such as voicing and nasality, but differ along dimensions such as place of articulation (Fletcher, 1929; Miller & Nicely, 1955). Some of this ambiguity is removed in natural settings by the fact that a speaker’s face is often available to the perceiver. Access to a speaker’s face has been shown to significantly improve intelligibility under less than ideal conditions, such as low signal-to-noise ratios (Sumbly & Pollack, 1954). It is not only under auditorily-degraded conditions that visual information has a functional impact, however. A now-classic paper published by McGurk and MacDonald (1976) reported that auditory recordings of “ba” dubbed to filmed articulations of “ga” often led to reports of “da”—a “fused” utterance with an intermediate place of articulation (at the alveolar ridge) that was never actually presented (see Figure 4 (b)). An auditory “ga” dubbed to a visual “ba” led to reports of a “combination” of both sounds (e.g., “gabga,” “baga,” etc.). Analogous effects have been found with a number of different CV combinations across a range of listener ages. The fusion effect is not lost by an awareness of the manipulation or by repeated exposure. The effect can also be insensitive to gender mismatches between face and voice (Green, Kuhl, Meltzoff & Stevens, 1991), and to asynchrony of the dubbing process (centered around the release burst) up to a delay of about 180 ms between the audio followed by the visual input (e.g., Munhall, Gribble, Sacco & Ward, 1996). The

asymmetry of the effects (fusion vs. combination) has been explained in terms of the visual prominence of a labial closure (as in “ba”) vs. the relative lack thereof in a velar articulation (as in “gah”). Also contributing to the phenomena may be the fact that high frequency auditory cues to place of articulation are relatively fragile and may be easily “overridden” by the unambiguous visual place cues that are present for labial closures (but not for closures further back in the vocal tract).

A number of recent studies have looked at the McGurk effect cross-linguistically (e.g., see Sekiyama & Tohkura, 1993; Sekiyama, 1997; Hardison, 1996). The fact that different languages utilize certain places of articulation more than others may have interesting implications for bimodal speech perception. For example, Summerfield (1987) notes that there are languages that frequently use place distinctions that have few visual correlates. Research is ongoing on the question of whether perception of speech in certain languages may rely more heavily on visual information than in other languages.

The McGurk effect is an illusion based on an unnatural co-occurrence of inputs, but it helps to demonstrate that speech perception is susceptible to the influence of visual information even when the auditory signal is not degraded. This finding has led to interest in just how much information the visual channel can provide during speech perception. “Viseme” has been the term coined to refer to an abstract linguistic unit of visual analysis analogous to the phoneme (Fisher, 1968). The set of all speech sounds that involve a visually apparent closure of the lips, for example, might constitute a single viseme. Confusability statistics have been generated for different visemes and the idea of “viseme-defined” lexicons has been explored (e.g., Auer & Bernstein, 1997). In general, the equivalence classes for visemes are somewhat larger than those for phonemes. One viseme category such as that defined by a labial closure may encompass several phonemes— [p], [b], and [m], in this case. This is a consequence of the fact that many of the contrasts used in spoken language (voicing, and nasalization, for example) have no visual analogue. Lip-reading is still, however, effectively used by many of the hearing-impaired to aid comprehension of spoken language.¹²

While the visual channel clearly can provide valuable information about speech, listeners often manage quite well enough without it. In telephone conversation, for example, not only does one not have the option of looking at the talker’s face, but the typical telephone receiver only reproduces frequencies between 350-3500 Hz. Nevertheless, this filtered speech is usually quite intelligible. What characteristics of speech help to make this possible?

From Phoneme Perception to Perceiving Fluent Speech

Perceiving Phonemes in Context

Several results from phoneme identification studies demonstrate that phonemically contrastive information is widely and variably distributed temporally across the incoming speech signal. Most speech sound contrasts are supported redundantly by several different aspects of the acoustic signal. Some of these aspects manifest themselves in a “trading relation,” in that the stronger presence of one characteristic can (or must) potentially compensate for the weaker presence or non-expression of another. For example, there are several aspects of the speech signal that can contribute to the perception of voicing in stops, including the duration of voice onset time and the first formant onset frequency following release. If one shortens the VOT of a particular synthetic stimulus, in order to generate the same percept as previously obtained, the onset frequency of the F1 transition must be increased—as would naturally

¹² “Speech reading” refers to the combining of lip-read information with liberal use of contextual information to achieve comprehension of spoken language. Speech reading bears a similar relation to lip-reading of visemes as auditory speech perception does to phoneme perception.

occur in a faster rate of speech. In terms of natural speech, this means that speakers have the flexibility of producing acoustic cues of variable values so long as necessary compensations are made—listeners will still perceive the intended contrast.

Coarticulatory context effects (Repp, 1982) are a body of data detailing how a single spectrographically salient acoustic cue can be perceived differently depending on other sounds in its immediate context. A synthetic release burst centered at 1600 Hz for example, is perceived as [p] before steady state vowels with low F1s such as [i], but as a [k] before the vowel [A] (Liberman, Delattre & Cooper, 1952). Frication noise with spectral intensities ambiguous between an alveolar and a palatal will be perceived as an [s] (an alveolar sound) if preceding a rounded vowel such as in “Sue,” but more often as a [ʒ] before the vowel [A] as in “shot” (Mann & Repp, 1980). Such “context-conditioned” effects are quite numerous, and are often cited as problematic for the idea that there are invariant cues for phonemes in speech. Syllables are often suggested as a viable alternative, but these units are also affected by context and speaking mode, although less so than phonemes. These results simply underline the points made earlier in this chapter regarding co-articulation; information relevant to identifying particular phoneme-level contrasts is often distributed across much wider temporal intervals of the signal than might be intuitively expected (based on orthography or spelling, for example).

In naturalistic circumstances, we are rarely required to identify phonemes in meaningless nonsense syllables. We distinguish between phonemes in the context of spoken words. A number of studies have examined how the criterion for interpreting particular acoustic information as giving evidence for one phonemic category versus another is affected by the lexical status of the context in which the acoustic energy occurs. An experienced native listener not only knows how various phonemes are phonetically implemented in his/her language, but also comes equipped with neural representations of the vocabulary of that language. The listener’s lexical knowledge can subtly influence the weighting of acoustic cues in phoneme categorization. A common illustration of this is the finding that the VOT criterion for what constitutes the voiced vs. voiceless contrast for stimuli with ambiguous VOTs is affected by the lexical status of the sounds formed by interpreting the acoustic information as one phoneme versus the other. The “cut-off” VOT value for labeling a stimulus as “beef” versus “peef,” for example, is different than the criterion value used to distinguish “beace” vs. “peace” (Ganong, 1980). A similar result is obtained if the frequencies of the two possible word interpretations are manipulated, with subjects choosing to classify ambiguous stimuli as the higher frequency item (e.g., Connine, Titone, & Wang, 1993).

Lexical knowledge can even cause a “filling-in effect” to occur when part of a familiar or contextually predictable word is artificially obliterated with a patch of noise. Listeners often report hearing the missing segment in addition to the noise. This phenomenon is called the “phoneme restoration effect” (Bagley, 1900/1901; see also discussion in Cole & Rudnick, 1983; Warren, 1970). Mispronunciations within familiar words often go unreported for apparently similar reasons. It has also been found that while utterances filtered to remove high frequency information are usually difficult to understand “in the clear,” they are, almost paradoxically, more intelligible when played through noise—presumably because “restoration mechanisms” are recruited to process this degraded form of the input (Miller, 1956).

Some theorists argue that a discussion of lexical effects on perceiving phonemic contrasts does not properly belong in a discussion of speech perception, stating that lexical influences are merely “biases” that operate “post-perceptually” only in circumstances where the auditory signal is degraded. While we do not believe this to be a tenable position, clearly a major issue in speech perception is how and when different types of “top down” knowledge and information in long-term memory come in contact with incoming “bottom-up” data (Luce & Pisoni, 1998; Massaro & Oden, 1995; McQueen, 1991; Pitt 1995; Pitt & Samuel, 1993). If speech cannot be recognized without the participation of top-down

information and knowledge, then defining “speech perception” independently of top-down information flow is arguably not speech perception at all. The section on spoken word recognition discusses this problem in more detail.

The Intelligibility of Speech Under Transformation

Trading relations, coarticulatory context effects, and phoneme restoration all demonstrate some of the ways in which the speech signal tends to be informationally redundant and consequently quite resistant to degradation. Experiments looking at the intelligibility of speech under gross distortion have shown that listeners can still recognize many words in a non-anomalous sentence even if, for example, the signal is interrupted by silence at rates faster than 10 times a second (Miller & Licklider, 1950), or is transformed to pitch-preserving time-compressed speech (Foulke, 1971). Other studies have shown that speech waveforms that have been “clipped and squared off” at their intensity peaks, or speech that has had each of its formant centers replaced with only a simple frequency-varying sinusoidal tone with all other intensities removed (Remez, Rubin, Pisoni, & Carrell, 1981) can also be linguistically intelligible. The acoustic patterns in even low-quality synthetic speech and speech-like productions by avian mimics that have vocal tracts far dissimilar to humans can similarly be identified.

The findings from these diverse studies seriously weaken the notion of speech perception as just a simple pattern-recognition process of identifying canonical strings of phonemes. Another rarely cited extreme manipulation conducted by Blesser (1969; 1972) took the spectral information in the speech signal between 200-3000 Hz and simply rotated it around an axis at 1600 Hz. This spectral transformation seriously distorts place cues but manages to maintain perceived pitch and intonation (Blesser, 1972). Blesser found that after several training sessions, listeners were managing to partially comprehend sentences in this highly unnatural degraded format even though recognition performance on isolated words remained low. He concluded that “learning to comprehend sentences or holding a conversation through the spectral transformation does not require the pre-requisite ability of perceiving isolated phonemes or syllables. Speech perception under normal conversing is not simply the result of processing a sequence of speech segments” (Blesser, 1969, p. 5). Similar findings have been reported more recently by Shannon, Zeng, Kamath, et al., (1995).

Normalization Issues and Indexical Information

The fact that the speech signal is so redundantly specified to the experienced speech perceiver suggests that it might be possible to perform some type of stimulus reduction such that only non-redundant information would play a role in the recognition process. The sheer number and variety of distortion types that can be successfully dealt with however, make this a daunting proposition. Nevertheless, implicit in nearly all work in speech perception is the assumption that hearers “normalize” the acoustic speech signal in order to recognize an abstract idealized symbolic linguistic message (Pisoni, 1997b). That is to say, variability not directly associated with the abstract linguistic message, such as information about speech rate or talker identity, is somehow culled from the signal, providing a “standardized” representation of an idealized canonical linguistic form. This “normalized” signal is the true object of speech perception, according to this abstractionist/symbolic view.

A variety of evidence has been offered to show that normalization must, by necessity, be taking place. If absolute frequency values were being used to identify linguistic units, the perceptual equivalence of utterances spoken by individuals with different-sized vocal tracts would be impossible. If absolute duration were used, the identical phrase spoken at different rates could not be recognized as such. (See Figure 5 to compare speech spectrograms of the utterance “You ought to keep one at home,” spoken in a selection of dissimilar voices.) Moreover, it has been suggested that once a “normalization adjustment”

takes place, this “calibration” continues to apply to all input perceived as coming from the same source. Ladefoged and Broadbent (1957) for example, used synthetic speech to demonstrate that manipulation of the range of the formant frequencies of a carrier phrase could affect the perception of vowel identity in a target word occurring later in the sentence. The issue then becomes a matter of specifying what exactly takes place during a listener’s exposure to the early part of an utterance that leads to processing later occurring target items in a particular way. One popular proposal is that the bounds of a speaker’s vowel space are used to scale or “self-scale” an utterance. Computer algorithms implementing this strategy have met with some limited success (Gerstman, 1968; Strom, 1997).

 Insert Figure 5 about here

Information used to make “linguistic” discriminations has traditionally been conceived of as having a separate and independent existence apart from non-linguistic, “indexical” information also contained in the signal. The term “indexical” refers to aspects of the signal that can provide information regarding the speaker’s identity and condition. (The speaker is male, the speaker is sad, the speaker is John, etc.) Traditionally, this type of information was thought to be “channeled-off” to some other processing mechanism during a normalization stage of speech perception, and not thought to have much of a short or long-term impact on perceiving phonemic contrasts or recognizing spoken words. In recent years however, evidence has begun to accumulate suggesting that indexical information is not discarded during or after normalization. There is evidence, however, that this distinction may not be as clear-cut as once thought. Although linguistics has treated indexical attributes of speech as variability that is irrelevant to understanding the abstract structure of language, theories of speech perception grounded in general theories of cognition and human information processing now recognize that indexical information interacts with attention and memory processes during speech perception.

Short-term as well as long-term influences of indexical properties on speech perception have been demonstrated. Palmeri, Goldinger, and Pisoni (1993), Goldinger (1996), Nygaard, Sommers, and Pisoni (1994), and Nygaard and Pisoni (1998), among others, have shown that talker-specific information is stored in long-term memory rather than discarded, and can affect recognition of spoken words studied, as well as the identification of non-studied words spoken by a familiar versus unfamiliar talker. This information appears to be stored in an “implicit” form of memory, which can impact performance on a variety of speech-related tasks, even if no explicit memory for indexical information can be reported by participants in the study.

A number of other experiments have shown decrements in word and phoneme identification when the stimuli involved are spoken by different talkers from trial to trial (see Mullennix, 1997, for a review). Lippman (1997), among others, has suggested that approximately two to four seconds of continuous speech from a single talker may be enough for a human listener to adjust sufficiently to perform at near optimal levels. This roughly matches results reported by Kakehi (1992) who found (using a stimulus set of one hundred talkers each recorded speaking 100 different Japanese CV monosyllables in isolation) that after approximately four to five syllables, the disruption in identification caused by changing the talker was overcome, with performance returning to single talker levels.

In general, familiarity with a speaker’s voice tends to make the speech perception process faster and less prone to error. However, recalling the results discussed earlier in the context of acquisition of

non-native sound categories, it also appears that experience with a variety of different talkers' tokens of a particular speech unit facilitates the generalized instance of perceiving an unfamiliar talker's version of the same token. Also worth noting is Mullennix and Pisoni's (1990) finding that listeners cannot simply choose to completely ignore one type of information (phoneme identifying characteristics vs. voice identifying characteristics), even when explicitly instructed to do so.

In short, there is now a large body of evidence suggesting that speech perception involves extracting both indexical information and a "symbolic" linguistic message from the speech signal, and that perception is necessarily influenced by representations of both kinds of information in memory. These sources of information are likely to be extracted at different rates from the signal, but these extraction processes are not independent of each other as once thought.

Also relevant to this issue is data on how people perceive, remember, and respond to synthetic speech—synthetic speech generated with the ideas of a "canonical representation" and invariant acoustic cues underlying its design (see Pisoni, 1997a). The intelligibility of synthetic speech produced by rule tends to be less than that of human speech, most markedly in non-optimal listening conditions (Pisoni, Nusbaum, & Greene, 1985; Logan, Greene, & Pisoni, 1989). There is also evidence that comprehension of text-to-speech produced passages can be poorer than that for naturally read versions of the same passage (see reviews in Ralston, Pisoni, & Mullennix, 1995; Duffy & Pisoni, 1992), and that recall for lists of items presented in synthetic speech is not as good as for lists presented in natural speech (Luce, Feustel, & Pisoni, 1983). These findings have led to interest in developing synthetic voices with more naturalistic patterns of indexical variability (Lambert, Cummings, Rutledge, & Clements, 1994).

The data discussed above suggest that results using typical "laboratory speech" will be difficult to generalize to speech perception under more realistic conditions. "Lab-speech" has traditionally involved items spoken (or synthesized) in isolation, consisting typically of nonsense syllables, and sometimes monosyllabic and spondaic single words. It is typically spoken slowly, using very careful articulation, and full vowels. "Lab speech," more often than not, is recorded in the clear, and is frequently read rather than spontaneously spoken. Often stimuli are created using only a single talker or speech synthesis system. These simplifications were practical and necessary in the past, however, we strongly feel the time has come to move on to a somewhat less conservative corpus of data on which to build a theory of speech perception (see also Stevens, 1996). In recommending that the variety of speech styles studied be expanded, we are advocating that this be done in a principled fashion, with naturalistic speech use kept clearly in mind. The next section briefly outlines why this issue is of importance to speech perception and spoken language processing.

The Impact of Speaking Styles on Perception

One of the most interesting current issues in speech perception concerns the nature of the communication process and how this is reflected in the production provided to a hearer's perceptual system. The basic point to be made, as reviewed by Lindblom (1996), is that the speaker and hearer typically share knowledge about what is "old" and "new" information. Consequently, not every utterance of the same intended canonical string of phonemes contains exactly the same amount of information. The predictability of the elements contained in the utterance will influence both the articulatory precision of the talker and subsequent perception by the listener. This account of speech communication says that, "...the (ideal) speaker makes a running estimate of the listener's need for explicit signal information on a moment to moment basis. He then adapts the production of the utterance elements (words, syllables, or phonemes) to those needs" (Lindblom, 1996, p. 1687; see also Lindblom, 1990).

Do these different speaking modes impact perception? Almost certainly. Hearers regularly benefit from speakers' automatic tendency to raise their voices in the presence of competing background noise (Lane & Tranel, 1971). Speakers also tend to place greater stress on rare words or when introducing a new topic (Berry, 1953; Fowler & Housum, 1987). More recently, Wright (1997) has shown that speakers modify their speech as a function of the estimated confusability of the word spoken. Specifically, words that sound similar to many high frequency words are articulated using vowels that define a larger, more expanded vowel space than equally familiar words that have very few similar-sounding words. Note however, that there is also the countering influence of speaker effort: Lindblom (1996, p. 1688) argues that "...phonetic form...is determined by the speaker's assumptions about the informational needs of the listener and by his own tacit propensity to make articulatory simplifications." Experienced speech perceivers likely take this information into account. When a speaker and listener do not share the same experiences and expectations regarding the content of the signal, communication can be expected to suffer.

In summary; it is misleading to conceive of the "interactive" aspect of naturalistic speech perception as "beginning strictly after peripheral auditory processing finishes." Rather, the raw acoustic signal itself is produced as a function of another speech user's knowledge about speech and language. In this sense, a "cognitive" influence on production has helped to determine perceptual outcome long before the signal even reaches the listener's cortical speech processing regions.

While we have taken the liberty of first presenting what we currently see as the most promising approaches towards future research in speech perception, these ideas have not developed in a vacuum. Next we briefly review past and current theoretical approaches to speech perception that have contributed to shaping the ideas already presented.

General Theoretical Approaches in the Study of Speech Perception

Motor Theory

The lack of invariance in the acoustic signal caused speculation for many years that perceptual data might somehow align better with facts of articulation (Liberman, Cooper, Harris, & MacNeilage, 1962). For example, Liberman et al. pointed out that the acoustic cues for a [d] in a syllable onset are quite different depending on the vowel that follows, yet an articulatory description of [d] as "an alveolar constriction" is comparable in each environment. It was proposed that coarticulatory effects are decoded by the listener's own knowledge of the effects of coarticulation on his/her own productions, and that it is through this mechanism that an abstract phoneme-based interpretation can be decoded. Liberman et al. argued that listeners make use of their own internalized articulator motor patterns to interpret heard speech. This became known as the "motor theory" of speech perception.

Motor theory also had associated with it strong claims about the existence of a specialized speech module (Liberman, 1970; Liberman & Mattingly, 1989). Data from categorical perception studies and duplex perception were commonly cited as evidence for this hypothesis, although, as has already been discussed, these interpretations are problematic.¹³ It was subsequently found, however, that low-level articulatory gestures map no less variably to perception than do acoustics. Production data charting the actual motion and activity of the articulators show that both within and across individual speakers, articulatory variability is very high (e.g., Harris, 1977; Johnson, Ladefoged & Lindau, 1993; MacNeilage, 1970), even when perceptual report is relatively stable. Also problematic are data indicating that not-yet-

¹³ There are other good reasons to nevertheless believe that speech is processed by the adult language user's nervous system in ways that are fundamentally different from other non-speech signals.

articulating infants appear to have the rudiments of speech perception, and the existence of individuals with severe voice disorders but normal perceptual capacities. Such findings resulted in the revision of the motor theory, such that the proposed motor correspondence no longer referred to externally measurable articulatory motions, but rather to the recovery of abstract sets of motor commands, or “intended phonetic gestures.” With this change, however, motor theory’s main hypothesis became extremely difficult, if not impossible to test. (See Liberman & Mattingly, 1985, for a discussion.)

Researchers, however, continue to explore the idea that the role of the perceiver as also a producer should contribute a potentially valuable information source to a model of speech perception, and even perhaps yield some strategies for improving automatic speech recognition (Rose, Schroeter, & Sondhi, 1996). Although it was primarily a theory of perception, aspects of motor theory can also be seen in current theories of articulatory production based on the timing and temporal overlap of abstractly defined gestures and goals (e.g., Browman & Goldstein, 1995; Saltzman & Munhall, 1989).

The Direct Realist Approach

The “direct realist” approach to speech perception is based on the legacy of Gibsonian ideas of “direct perception,” particularly in the areas of vision, and haptic sensation (e.g., Gibson, 1966). With reference to speech, direct realism is largely associated with the work of Fowler and colleagues (e.g., Fowler, 1986; 1996). The main point being made in the direct realist approach is that in delving only deeper into the intricacies of acoustic variation, the ecological function of speech perception is in danger of being ignored. What animals directly perceive is information about events in their environment, not the intermediate structure conveyed to some intervening medium such as the air. Explanation, it is argued, can be couched in terms of the “public aspect” of perception. Proponents of this view suggest that “there is no need to look inside the perceiver to explain how the percept acquires its motor character” (Fowler & Rosenblum, 1990, p. 743). In other words there is no need for positing the intermediate mental representations of a levels-of-processing model.

The “motor character” of the percept is referred to above, because ostensibly, the objects of perception in direct realism are still abstract gestures. The abstract gestures described by direct realist approaches span more than a single modality, however, thereby providing the potential basis for a type of parity relation between perception and production. Direct realism, however, does not accept motor theory’s claim that reference to one’s own specialized speech motor system is necessary for perception of speech. Speech is directly perceived like any other meaningful sound in a creature’s environment, in terms of a hypothesis about its source (Gaver, 1993) with perceptual mechanisms that are not necessarily unique to humans.

Although a direct realist perspective does provide a valuable reminder of the need to step back and consider perception in relation to its larger environment, to constitute a testable theory of speech perception, it will require considerably more specification. (For discussion see Fowler, 1996; Lindblom, 1996; Ohala, 1996; and Studdert-Kennedy, 1986.)

An Integration of Multiple Information Sources Approach

As we have already noted, information about the speech signal often comes from both auditory and visual channels. The two approaches already mentioned acknowledge that information from other sensory modalities can affect the perception of speech, but being more general theories, they are not of a form to make very specific predictions about how this information is used under particular sets of circumstances. The “fuzzy logical model of perception” (FLMP) proposed by Massaro (Massaro, 1987; 1989; 1998; Oden & Massaro, 1978) aims to model both auditory and visual influences under very

specific response constraints. Massaro's theory of speech perception is probably best characterized as a parameter-tailored version of a more general Bayesian model for general categorization of stimuli into mutually exclusive categories. The motivation for this model is to see if perception of bimodal speech can or cannot be explained in terms of a simple mixture or weighting of the two channels of information, auditory and visual. The fuzzy logical model assumes that the information conveyed in the two sensory input channels is independently evaluated before integration. The "fuzzy logic" aspect of this model simply involves the fact that Massaro chooses to frame his probabilities in terms of "partial truth values" meant to represent continuous feature values, (i.e., x is partially voiced). Following evaluation, the two sources of information are integrated and then compared to some type of (modality-independent feature-based) best "example" prototype in long term memory. The best match is determined to be the most likely response.

This model has been formalized so generally (in terms of basic axioms of formal probability) that some have commented (e.g., Crowder 1989; Cutting, Bruno, Brady & Moore, 1992) that FLMP in a sense works "too well." Indeed, by adjusting the probability and prior probability parameters on a case by case basis, FLMP can be tightly fit to a variety of existing response data from forced choice discrimination and goodness-ratings tasks involving stimulus items ranging along a single continuum between two perceptual categories.

In a number of papers, Massaro has used signal detection theory and FLMP to try to disentangle "pure" perceptual sensitivity for phonemic contrasts from different kinds of response "biases." Lexical influences are often mentioned by Massaro as being among these "biases" (Massaro, 1994). Future versions of FLMP will hopefully model the influence of these crucial "biases" in more detail and be applicable to a wider range of experimental paradigms beyond simple closed-set response choices involving a minimal number of contrasts. (See Sommers, Kirk, & Pisoni (1997) regarding how the use of open-set versus closed set tasks can affect accurate clinical assessment of speech perception skills.)

Bridging the Gap Between Speech Perception and Spoken Word Recognition

General Points Regarding Spoken Word Recognition

As may be apparent from the theoretical approaches outlined above, the traditional study of speech sound perception has typically minimized, or simply ignored, the possible influences of stored lexical and other linguistic knowledge. On the other hand, spoken word recognition models often assume as their input, the very output that models of speech perception have a difficult time accounting for, that is, sequences of idealized, symbolic phonemes, cleanly segmented at word boundaries. Instead, research in spoken word recognition has focused the bulk of its efforts on the impact of stored lexical representations (i.e., a "mental lexicon").

What sort of data should a comprehensive model of spoken word recognition account for? In addition to explaining how a word is correctly identified, it must account for behavioral patterns of misidentification and errors. It should explain how word frequency and phonetic similarity can influence spoken word recognition, and it will almost certainly require some kind of commitment as to how spoken word representations are stored and retrieved. In practical terms, a model of spoken word recognition must predict quantified measures of perceptual behavior including the relative speed and accuracy of responses as a function of the acoustic input provided. In the following sections we review--in highly abbreviated form--a selection of models that have each contributed some important insight into these processes.

Cohort Theory

The Cohort model (Marslen-Wilson & Welsh, 1978; Marslen-Wilson, 1987) was among the first to make detailed predictions about the time course of the recognition process in the context of acoustic-phonetic similarities between known words. Early versions of cohort theory conceived of the recognition process as a temporally unfolding “weeding-out” of a cohort of possible word candidates as each new bit of acoustic input becomes available. Recognition occurred, according to the model, when only a single word remained in the cohort. (See Figure 6 (a)).

 Insert Figure 6 about here

The cohort model has been studied extensively using the gating paradigm introduced by Grosjean (1980). The gating paradigm asks listeners to identify a word given successively longer segments of the signal (usually measured from its onset). Performance on this task provides some insight into the time course of word recognition and factors influencing the selection of lexical candidates (see also Salasoo & Pisoni, 1985).

The original cohort model assumed, however, that candidates were fatally eliminated from consideration as soon as any new bit of inconsistent acoustic-phonetic information was received. This assumption has since been relaxed to permit instances of successful identification in less than ideal listening conditions, or from a partially distorted production. Elimination of candidates is argued to be also possible via “top-down” contextual information. Word frequency effects are accounted for by having stored representations of familiar words possess elevated resting activation levels. This provision predicts faster and more accurate identification of high frequency items, and while this activation level can be gradually suppressed in the face of counter-evidence, higher frequency words will tend to remain in the active cohort longer than low frequency candidates. This activation-based account of frequency effects can be contrasted with that of earlier search models, such as Forster’s (1978), which assumed that word recognition took place within some type of frequency-ordered data structure with high-frequency words searched first.

Although cohort theory provides an intuitive hypothesis about how spoken words may be discriminated and recognized, it makes a number of uncorroborated predictions and leaves several key problems unaddressed. In particular, cohort theory does not incorporate any influence of cohort size during the time course of recognition. It also assumes knowledge of where one word starts and another begins. The way in which the cohort model deals with spoken non-words presented for lexical decision is also not ideal. It predicts that non-words will be identified as such immediately after they “diverge” from all known words in their structure; the data, however, suggest that information that occurs after the divergence point can influence the speed at which the item is identified as a non-word (Taft & Hambly, 1986). Also rather problematic for Cohort theory is Bard, Shillcock, and Altmann’s (1988) report that approximately 20% of successful identifications of words spoken in continuous speech are not recognized in a word-by-word gating task till one or more words following have been presented (see also Grosjean & Gee, 1987). That is, it is not guaranteed that a word will be recognized immediately upon the elimination of all other candidates; in many situations some delay is experienced. Findings such as these have encouraged exploration of other, more flexible ways of defining “cohort” candidate sets during word recognition.

The Neighborhood Activation Model

A large body of evidence suggests that words are recognized in the context of knowledge about other similar sounding words (Treisman, 1978). The Neighborhood Activation Model (NAM) (Luce & Pisoni, 1998) was designed to emphasize this aspect of word recognition. The term “neighborhood” simply refers to the set of words that are phonetically similar to a given word according to an arbitrarily defined criterion (see Figure 6 (b)).

NAM predicts that the probability of correctly identifying a word in noise will be equal to its frequency times the probability of correctly identifying all of its constituent phones, divided by the sum of: (1) the term in the numerator, plus (2) a term summing across all neighbors, the probability of identifying all of the target word’s segments as the segments contained in the neighbor, multiplied by the frequency of that neighbor. Existing corpora of phonetic transcriptions and frequency statistics in combination with behavioral data on the likelihood of auditory confusion between phones provide the parameters for this specialized version of R. D. Luce’s (1959) “choice rule.”

For identification of words in the clear, the confusion data probabilities are replaced by counts that define a neighborhood for a given word as the set of all known words that differ by one substitution, deletion, or addition of a segment from the transcription of the target word (Greenberg & Jenkins, 1964). The model is then used to make qualitative predictions about speed and accuracy measures for lexical decision and naming tasks as a function of word frequency, neighborhood size, and the mean frequency of the items in the neighborhood (see Luce & Pisoni, 1998).

To illustrate how this rule operates for identification accuracy in noise, consider two words with the same frequency, the same segment-based probability of correct identification, and also the same number of neighbors (each whose component segments are equally easy to identify accurately). If these values are changed, however, so that one of the words tends to have neighbors that are higher in frequency than the other, the equation predicts the first word to be more difficult to identify than the second. If the two words instead have equally frequent neighbors, but one word has many fewer neighbors total than the other, then that word is predicted to be easier to accurately recognize. Two words having equal frequency, equal number of neighbors, and equal average frequency of these neighbors can still be differentiated in this equation if they differ in the probability of correct identification of all constituent segments. Lastly, all other factors being equal, the segmental composition of a word’s neighbors can distinguish two words—the word whose neighbors contain easily identifiable individual segments is predicted to be easier to identify than the word whose neighbors contain segments that are often confused for other sounds.

NAM has been applied to a large target recognition set of highly familiar monosyllabic CVC English words whose possible neighborhoods were restricted to other highly familiar monosyllabic words. The combination of factors included in the model has been shown to be a significantly better predictor of identification than target word frequency alone.

NAM in its more common form assumes access to a segmented, phoneme-based representation of speech and has been criticized for lacking much of a dynamic temporal component. While not actually implemented to reflect this, NAM suggests, however, that a probability computation similar to the one just described unfolds continuously in time over the course of the input thereby leading to identification or misidentification. Lexical access is assumed to occur after this point, with more abstract representations, such as meaning, only then being activated and available for retrieval (see Auer & Luce, submitted).

The TRACE Model

The connectionist approaches to spoken word recognition that will be reviewed next involve computing activation values in parallel across different layers of computational units called “nodes.” Each layer of nodes usually represents (by fiat of the designer) a different level of abstract linguistic structure. Nodes may have their activation levels influenced by relationships defined with other nodes in terms of weighted connection strengths. Behavior of the network is determined by deploying an activation function and an update rule on its architecture. Certain types of connectionist models lend themselves to representing the temporal nature of speech perception in at least a semi-plausible manner and are thus better able to deal with some types of stimulus variability. Connectionist models have the attraction of both formal specification and potentially physical implementation. Artificial neural net systems are often pre-designed to accomplish a certain task (as is the case with TRACE, discussed below), but it is worth keeping in mind that the more interesting case is the self-organizing net that adapts itself via a learning process often (but not necessarily) provided by the designer (for discussion, see Grossberg, Boardman, & Cohen, 1997).

The TRACE model is an interactive activation neural-net model of speech perception (McClelland & Elman, 1986) which in its design drew considerably on McClelland and Rumelhart’s (1981) model of visual word recognition. TRACE uses three layers of nodes, one layer representing a feature level of analysis, one layer, a phoneme level, and the last layer, word categories (see Figure 6 (c)). Nodes within each layer represent particular features, phonemes, or lexical items. Nodes that are mutually consistent (e.g., the node for the “voicing feature” and the node representing the voiced phoneme [b]) are bidirectionally connected by excitatory connections. There are also bidirectional within-layer inhibitory connections, such that a highly activated unit will tend to send negative activation to its “competitor” within-level category nodes. There are, however, no between-level inhibitory links.

The TRACE model as described in McClelland and Elman (1986), uses “mock speech” input defined in terms of values assigned to seven acoustic feature dimensions.¹⁴ These feature values may take on a range of eight different values at every time-step of 25 ms and are thus able to reflect some of the coarticulation present in natural speech. The number of identifiable phoneme categories is simplified to fourteen, and the network’s vocabulary to approximately two hundred possible words. Inputs corresponding to immediate feature values are fed into a single time slice of the network (containing its own full set of all feature, phoneme and word nodes), and into subsequent time slices, one after another. The influence of input at one time-step on the activations at another time-step is implemented by having each phoneme node linked to the feature nodes of the previous and following time-step. Output behavior is determined by the maximum activation value within a particular level at a particular time-step. The simulated time course of recognition can be observed in this manner.

TRACE can be shown to simulate many of the classic effects in the literature, and is able to recover from and “fill in” noisy input (for revisions suggested to deal with some of the phoneme categorization data, see McClelland, 1991). One of TRACE’s inelegances however, is its reduplication of the entire network at each time step. Newer neural network designs suggest ways to avoid this, and also ways of beginning to deal with some of the temporal structure of the speech signal via involvement of memory resources (see Grossberg et al., 1997 on ARTPHONE for one recent discussion).

¹⁴ Elman & McClelland (1986) reports a separate version of the TRACE model (TRACE I) which utilizes input values derived from actual spectra in order to identify phonemes.

Among TRACE's weaknesses are its very limited lexicon and its failure to use information that can constrain the likelihood that all time slices represent plausible beginnings of words. These characteristics motivated the development of Shortlist (Norris, 1994).

Shortlist

Shortlist is also a connectionist model of spoken word recognition, but unlike TRACE, it does not make use of "top-down" connections per se. Norris (1994) has argued that such connections are redundant and undesirable. Shortlist is designed to isolate the process of candidate word generation, from the process of competition among these candidates. To accomplish this, Shortlist uses a single recurrent network (lacking top-down connections), to generate phoneme classifications from feature values given as input. The recurrent network feeds its output at each time-step into a separate single-layer interaction activation network containing a limited set of mutually inhibitory lexical hypothesis nodes (the "shortlist" or cohort). During the unfolding process of word recognition, lexical candidates are continually being generated by the recurrent network and replacing former candidates. (See Figure 6 (d).)

With this architecture, Shortlist implements the proposal that phonemes can be identified by two routes: (1) "bottom-up" acoustic-phonetic analysis, or (2) "top-down" via lexical knowledge. In other words, a current perceptual hypothesis can be "read out" of either the phoneme or lexical hypothesis level. Lexical hypotheses that are strong but partially inconsistent with the phoneme classifications are not able to (indirectly) inhibit the phoneme classification process at the "lower level" via top-down excitatory connections to competitor phoneme nodes. Norris (1994) criticized TRACE for permitting potential top-down lexical influences under all conditions, suggesting that "top-down effects may well be dependent on the quality of the stimulus and may only emerge when the stimulus is degraded in some way, either by low pass filtering or by the removal of phonetic cues" (Norris, 1994, p. 192). (Since, however, more often than not in everyday circumstances, the speech signal is less than optimal, this point remains at issue.)

McQueen, Norris, and Cutler (1994) have also suggested that Shortlist employs a "metrical segmentation strategy" which builds in the heuristic, suitable for finding the word boundaries of English in continuous speech, that stressed syllables tend to mark the beginning of content words. This strategy, when combined with other acoustic cues to word juncture (e.g., Nakatani & Dukes, 1977), may help to solve the word segmentation problem, at least in English.

The models thus far presented were designed with the primary goal of accurately representing certain well-documented phenomena of human spoken language processing behavior. Practicality of physical implementation and efficiency of design were not necessarily at the forefront of their concerns. The progress made towards addressing the problems of spoken language recognition from the perspective of this alternative set of priorities, has, however, yielded some insights about what components cannot be left out of even a minimal working system for the recognition of spoken language.

Conclusions from Continuous Automatic Speech Recognition

At this point, readers familiar with currently available commercial speech-to-text software applications may be wondering if any of the systems used in continuous automatic speech recognition by machines constitute models of human speech perception at least as well as some of the proposals reviewed so far. There are different types of automatic speech recognition (ASR) systems (some of which use neural-net type architectures). Many systems operate on series of power spectra, smoothed and averaged to bring out formant information and cancel out random noise. These are then compared to stored sets of canonical smoothed spectra, combinations of which are matched to phone or word

representations. These matches do not go on in vacuo, as it were—probabilistically encoded information about what phones typically occur together, and how probable various word combinations are in the language, also get figured in. This last mentioned “language module” or “long-term linguistic knowledge storage component” of ASR systems has turned out to be crucial to successful performance (Jelinek, 1997).

For ASR, within-speaker coarticulation has turned out to be the more tractable problem, but one that has been handled, not by assuming that words can be extracted phoneme by phoneme from the signal, but rather that stored knowledge about words, language and the world can be brought to bear on the process. Variability arising from environmental conditions and cross-speaker differences (e.g., regional dialect, age), has posed a much more difficult problem, partly countered by extensive training sessions, storage of different “talker” templates, and noise-canceling microphones.

ASR algorithms traditionally have not necessarily sought out the “most human way” of accomplishing a goal—no assumption has been made that this would be an optimal solution. Instead designers have balanced issues of computational speed and resources alongside issues of performance. Next discussed is a model of spoken word recognition that was among the first to try to address both practical design and implementation issues in addition to modeling actual human behavior.

Lexical Access from Spectra

Klatt’s (1979; 1986; 1989) “lexical access from spectra” model, or LAFS, is probably the best-known example of an ASR-derived model of human spoken word recognition. LAFS was designed with the ASR methods of the day in mind and was based on a system known as Harpy (Lowerre & Reddy, 1980), so it is somewhat dated, but its basic points are still relevant.

LAFS assumes that word recognition is a process of matching an incoming series of short-term spectra to the stored contents of a very large branching tree-structure network of temporally ordered spectra. The spectra used by LAFS in storage and comparison are not raw spectra but rather neural transforms generated by the peripheral auditory system. Inputs are not defined terms of any abstract intermediate linguistic unit except the algorithm for generating the power spectra.

The proposed method of constructing the large network of stored spectra was not very satisfactory from a psychological (or technical) point of view. Klatt’s proposed solution was to create a phoneme-based tree containing the entire vocabulary of the system, modify this tree according to traditional rules of generative phonology to create a phone-based tree reflecting all possible phonetic realizations of these words including coarticulatory influences across word and phoneme boundaries, and then to in turn convert this network into a spectra-based tree generated from spectra sequences selected to correspond to several thousand diphones. This “precompiled” decoding network was described as theoretically containing “all possible acoustic realizations of a word or a sequence of words in the language.”

While a working implementation of LAFS was never built due to technical limitations of the day, currently available continuous ASR systems incorporate some of its components. Flexible handling of inter-talker variability continues, however, to be a difficulty for even today’s systems. Klatt (1979) suggested that some type of on-line adaptation algorithm for modifying the set of stored spectral templates was needed for LAFS to deal with inter-talker variability—a current technical review of how ASR systems do something akin to this can be found in Zhao (1997).

LAFS is clearly a model of a “mature” recognition system—how new words would be dealt with was never addressed in detail. In fact, none of the models reviewed above really provides an understanding of how and why the perceptual categories each model assumes arise in the first place. In each case, the system is simply endowed with the perceptual units judged relevant and given an architecture that has developmental plausibility in only the loosest of terms.

Development of Speech Perception and Spoken Word Recognition

The developmental literature on infant discrimination of phonetic contrasts is quite large and we do not have room to do it justice here. For recent reviews see Aslin, Jusczyk, and Pisoni (1998), Jusczyk, (1993; 1997), Kuhl (1993), Vihman (1996), and Werker and Polka (1993).

Infants prior to about six to ten months of age show an impressive ability to make acoustic discriminations. This ability is not limited to speech sounds, and wide variety of contrasts used in languages both native and nonnative to the child can be discriminated. These findings have usually been demonstrated using the habituation-based high-amplitude sucking procedure or head-turn preference procedure (see the reviews listed above for details). Towards the last quarter of their first year, infants begin to show an influence of their native language in their discriminations, with responses resembling the categorical-like data from adults described earlier.

A long-standing question has been how children (and adults, for that matter) are able to perceptually segment words and potential words from a continuous speech stream. It has been proposed that sensitivity to stress patterns might be key. Jusczyk, Cutler, and Redanz (1993) have shown that by nine months, American infants show evidence of developing a preference for words with a strong-weak stress pattern—the most common pattern for content words in their native language (Cutler & Carter, 1987). The learning of language-specific phonotactic distributions, that is, where particular sounds tend to occur within utterances, may also aid in segmentation (Brent & Cartwright, 1996). Evidence suggesting that infants can acquire prenatal experience with native language prosody indicates that the groundwork for dealing with the word segmentation problem may be laid out even before birth (e.g., Decasper & Spence, 1986; Mehler, Jusczyk, Lambertz, Halsted, Bertoni, & Amiel-Tilson, 1988).

Assuming a solution to the segmentation problem, how then are young children’s “lexical experiences” stored and organized in long term memory? A dearth of practical experimental methods has resulted in a situation where there are large bodies of infant and adult speech perception data (albeit based on rather different operational definitions of what constitutes speech perception), but relatively little data from young children beyond infancy. It has been argued that young school-age children show the phoneme perception skills of an adult, but clearly, young children’s word recognition skills are much more limited than those of adults, even while their capacity for learning the sound forms of new words is at its height. A handful of studies have begun to examine how speech perception changes as a function of a growing lexicon (Charles-Luce & Luce, 1990, 1995; Dollaghan, 1994; Gerken, Murphy, & Aslin, 1995; Walley & Metsala, 1990). Further research in this area should help clarify how low-level auditory representations of speech interface with long-term lexical knowledge.

Some of these ideas have recently been fleshed out by Jusczyk (1993; 1997) into a theory of word recognition and phonetic structure acquisition (WRAPSA). WRAPSA accounts for developmental changes in discrimination sensitivity in terms of experience-dependent changes in “weighting-schemes” applied to the acoustic-phonetic input. WRAPSA adopts an account of these weighting shifts, analogous to proposals in the general categorization literature of cognitive psychology, that “shrinking or stretching” of certain perceptual dimensions relative to their corresponding physical dimensions occurs as a result of experience (e.g., Nosofsky, 1987).

What is stored in long-term memory when a speech event is experienced? For many years, the assumption that storage space in the brain was “at a premium,” favored the notion of abstract phonetic representations. “Memory space” could be used more efficiently, it was suggested, by storing a prototype (perhaps formed by averaging each new instance with the existing prototype), instead of individual experiences (“exemplars”) (e.g., see Posner, 1969).

A string of papers published over the last twenty years suggests, however, that massive reductions in information storage are not necessarily required or even desirable (e.g., Hintzman, 1986; Nosofsky, 1988). The discovery of methods for compressing large amounts of information in small amounts of physical space has helped to make the storage capacity problem less of an issue. People show evidence of storing more in their brains than simply a symbolic representation of speech. They also show evidence of being better able to correctly generalize in recognition tasks when variability in a particular category of interest has been previously experienced (recall previously discussed data on the perception of nonnative contrasts, for example). Exemplar-based models provide ways of accounting for such findings. The WRAPSA model discussed above argues for an exemplar-based learning and storage process, as have recent models of adult vowel identification (Johnson, 1997), and results from word recognition (e.g., Goldinger, 1996; 1998).

Current Trends and Some Future Directions

While we have covered a wide variety of topics, there are several core issues touched on in this chapter that we feel are fundamental to future progress of the field. Briefly, in review: firstly, we urge that serious attempts be made to address issues of stimulus variability now. If tackling these issues is postponed further, the problem will be compounded when issues of perceiving typical “lab speech” versus naturalistic speech are to be resolved. Variability in the speech signal may in fact represent a bonus rather than a liability, a clever training recipe for insuring the communicative robustness of speech, rather than an inconvenient complication (as it has sometimes been treated in the last fifty years of speech science research).

Second, there is something distinctly peculiar with the idea of trying to study speech perception isolated from the influences of long-term memory. The organization of long term memory representations of speech-relevant data and the means by which these structures are continually updated is extremely relevant to the perception of fluent speech and must be dealt with explicitly. An account of speech perception also needs to be situated in a position where it can be plausibly integrated with other aspects of cognitive functioning and development. An example of work with important clinical implications that is headed in this direction is research looking at the effects of age-related changes in short term memory, long term memory, and attention on speech perception and spoken word recognition (e.g., Sommers, 1996; 1997).

More work also needs to be done on audiovisual aspects of speech perception. That people can and do make use of visual information in perceiving speech categories has been firmly established. However, this account still requires considerable fleshing out on when and how this information is used and what kind of differences may exist between language populations and between individuals.

Lastly, speech perception does not necessarily involve matching to, or storing only abstract idealized canonical units. The capacity to store data relevant to speech has been seriously underestimated in the past, and this assumption has constrained the design of speech perception and spoken word recognition models. In order to account for a number of the findings reviewed in this chapter, new types of process models must be developed. To be productive, this effort will require the continued cooperation

of the research areas of sensory, cognitive, and developmental psychology, speech and hearing sciences, linguistics, and speech technology/engineering.

References

- Aslin, R. N., Jusczyk, P. W. & Pisoni, D. B. (1998). Speech and auditory processing during infancy: Constraints on and precursors to language. In W. Damon (Ed.), *Handbook of Child Psychology, Fifth Edition, Volume 2: Cognition, Perception, & Language*. New York: John Wiley & Sons.
- Aslin, R. N., Pisoni, D. B., Hennessy, B. L. & Perey, A. J. (1981). Discrimination of voice-onset time by human infants: New findings and implications for the effects of early experience. *Child Development*, *52*, 1135-1145.
- Auer, E. T. & Bernstein, L. E. (1997). Speech reading and the structure of the lexicon: Computationally modeling the effects of reduced phonetic distinctiveness on lexical uniqueness. *Journal of the Acoustical Society of America*, *102*, 3704-3710.
- Auer, E. T. & Luce, P. A. (submitted). Dynamic processing in spoken word recognition: The influence of paradigmatic and syntagmatic states. *Cognitive Psychology*.
- Bagley, W. C. The apperception of the spoken sentence: A study in the psychology of language. *American Journal of Psychology*, *1900-1901*, *12*, 80-130.
- Ballas, J. A. (1993). Common factors in the identification of an assortment of brief everyday sounds. *Journal of Experimental Psychology: Human Perception and Performance*, *19*, 250-267.
- Bard, E. G., Shillcock, R. C., & Altmann, G. E. (1988). The recognition of words after their acoustic offsets in spontaneous speech: Evidence of subsequent context. *Perception and Psychophysics*, *44*, 395-408.
- Berry (1953). Some statistical aspects of conversational speech. In W. Jackson (Ed.), *Communication Theory*. London, UK: Butterworth.
- 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*. Cambridge, MA: MIT Press.
- Best, C. T., McRoberts, G. W. & Sithole, N. N. (1988). Examination of perceptual reorganization for non-native speech contrasts: Zulu click perception by English-speaking adults and infants. *Journal of Experimental Psychology: Human Perception and Performance*, *14*, 345-360.
- Blessner, B. (1969). The inadequacy of a spectral description in relationship to speech perception. Presentation to Acoustical Society Meeting, November, San Diego, CA.
- Blessner, B. (1972). Speech perception under conditions of spectral transformation: I. Phonetic characteristics. *Journal of Speech and Hearing Research*, *15*, 5-41.
- Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: IV. Effects of perceptual learning on speech production. (1997). *Journal of the Acoustical Society of America*, *101*, 2299-2310.
- Brancazio, L. & Fowler, C. A. (1998). On the relevance of locus equations for production and perception stop consonants. *Perception and Psychophysics*, *60*, 24-50.

- Brent, M. R. & Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, *61*, 93-125.
- Broadbent, D. E. (1955). A note on binaural fusion. *Quarterly Journal of Experimental Psychology*, *7*, 46-47.
- Browman, C. P. & Goldstein, L. (1995). Dynamics and articulatory phonology. In *Mind as Motion*, R. F. Port & T. van Gelder (Eds.) Cambridge: MIT Press.
- Charles-Luce, J. & Luce, P. A. (1995). An examination of similarity neighbourhoods in young children's receptive vocabularies. *Journal of Child Language*, *22*, 727-735.
- Charles-Luce, J. & Luce, P. A. (1990). Similarity neighbourhoods of words in young children's lexicons. *Journal of Child Language*, *17*, 205-215.
- Chiba, T. & Kajiyama, M. (1941). *The Vowel: Its Nature and Structure*. Tokyo: Kaiseikan.
- Cole, R. A. & Rudnicky, A. I. (1983). What's new in speech perception? The research and ideas of William Chandler Bagley, 1874-1946. *Psychological Review*, *90*, 94-101.
- Connine, C. M., Titone, D & Wang, J. (1993). Auditory word recognition: Extrinsic and intrinsic effects of word frequency. *Journal of Experimental Psychology Learning, Memory, and Cognition*, *19*, 81-94.
- Crowder, R. G. (1989). Categorical perception of speech: A largely dead horse, surpassingly well kicked. *Behavioral and Brain Sciences*, *12*, 760.
- Cutler, A. & Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language*, *2*, 133-142.
- Cutting, J. E., Bruno, N., Brady, N. P. & Moore, C. (1992). Selectivity, scope, and simplicity of models: A lesson from fitting judgements of perceived depth. *Journal of Experimental Psychology: General*, *121*, 364-381.
- DeCasper, A. J. & Spence, M. J. (1986). Prenatal maternal speech influences newborns' perception of speech sounds. *Infant Behavior and Development*, *9*, 133-150.
- Delattre, P., Liberman, A., & Cooper, F. S. (1955). Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America*, *27*, 769-773.
- Denes, P. B. & Pinson, E. N. (1993). *The Speech Chain: The Physics and Biology of Spoken Language*. New York: W. H. Freeman.
- Dollaghan, C. A. (1994). Children's phonological neighbourhoods: Half empty or half full? *Journal of Child Language*, *21*, 257-271.
- Duffy, S. A. & Pisoni, D. B. (1992). Comprehension of synthetic speech produced by rule: A review and theoretical interpretation. *Language and Speech*, *35*, 351-389.

- Eimas, P. D. (1963). The relation between identification and discrimination along speech and non-speech continua. *Language and Speech*, 6, 206-217.
- Eimas, P. D. (1985). The perception of speech in early infancy. *Scientific American*, January, 46-52.
- Eimas, P. D., Siqueland, E. R., Jusczyk, P. W. & Vigorito, J. (1971). Speech perception in infants. *Science*, 171, 303-306.
- Elman, J. L. (1989). Connectionist approaches to acoustic/phonetic processing. In W. D. Marslen-Wilson (Ed.), *Lexical Representation and Access*. Cambridge MA: MIT Press.
- Elman, J. L. & McClelland, J. L. (1986). Exploiting lawful variability in the speech waveform. In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and variability in speech processes*. Hillsdale, NJ: Erlbaum.
- Fant, G. (1960). *Acoustic theory of speech production, with calculations based on X-ray studies of Russian articulations*. Gravenhage: Mouton.
- Fisher, C. G. (1968). Confusions among visually perceived consonants. *Journal of Speech and Hearing Research*, 11, 796-804.
- Flanagan, J. L. (1972). *Speech Analysis, Synthesis & Perception, 2nd Edition*. New York: Springer-Verlag.
- Fletcher, H. (1929). *Speech and Hearing*. New York: Van Nostrand.
- Forster, K. I. (1978). Accessing the mental lexicon. In E. Walker (Ed.), *Explorations in the Biology of Language*. Montgomery, VT: Bradford.
- Foulke, E. (1971). The perception of time compressed speech. In D. L. Horton & J. J. Jenkins (Eds.), *The Perception of Language*. Columbus, Ohio: Charles E. Merrill Publishing.
- Fowler, C. A. (1986). An event approach to the study of speech perception from a direct realist perspective. *Journal of Phonetics*, 14, 3-28.
- Fowler, C. A. (1994). Invariants, specifiers, cues: An investigation of locus equations as information for place of articulation. *Perception and Psychophysics*, 55, 597-610
- Fowler, C. A. (1996). Listeners do hear sounds, not tongues. *Journal of the Acoustical Society of America*, 99, 1730-1741.
- Fowler, C. A. & Housum, J. (1987). Talker's signaling of "new" and "old" words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language*, 26, 489-504.
- Fowler, C. A. & Rosenblum, L. D. (1990). Duplex perception: A comparison of monosyllables and slamming doors. *Journal of Experimental Psychology: Perception and Performance*, 16, 742-754.

- Fujisaki, H. & Kawashima, T. (1968). The influence of various factors on the identification and discrimination of synthetic speech sounds. *Reports of the 6th International Congress on Acoustics*. Tokyo.
- Fry, D. B. (1979). *The Physics of Speech*. Cambridge UK: Cambridge University Press.
- Fry, D. B., Abramson, A. S., Eimas, P. D. & Liberman, A. M. (1962). The identification and discrimination of synthetic vowels. *Language and Speech*, 5, 171-189.
- Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, 6, 110-125.
- Gaver, W. W. (1993). What in the world do we hear?: An ecological approach to auditory event perception. *Ecological Psychology*, 5, 1-29.
- Gerken, L., Murphy, W. D. & Aslin, R. N. (1995). Three- and four-year-olds' perceptual confusions for spoken words. *Perception and Psychophysics*, 57, 475-486.
- Gerstman, L. J. (1968). Classification of self-normalized vowels. *IEEE Transactions on Audio and Electroacoustics*, AU-16, 1, 78-80.
- Gibson, J. J. (1966). *The Senses Considered as Perceptual Systems*. Boston, MA: Houghton Mifflin.
- Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 22, 1166-1183.
- Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, 105, 251-279.
- Goto, H. (1971). Auditory perception by normal Japanese adults of the sounds 'l' and 'r'. *Neuropsychologia*, 9, 317-323.
- Greenberg, J. H. & Jenkins, J. J. (1964). Studies in the psychological correlates of the sound system of American English. *Word*, 20, 157-177.
- Green, K. P., Kuhl, K. P., Meltzoff, A. N., & Stevens, E. B. (1991). Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect. *Perception and Psychophysics*, 50, 524-536.
- Grosjean, F. (1985). The recognition of words after their acoustic offsets: Evidence and implications. *Perception and Psychophysics*, 38, 299-310.
- Grosjean, F. & Frauenfelder, U. H. (1997). (Eds.) *A Guide to Spoken Word Recognition Paradigms*. Hove, UK : Psychology Press.
- Grosjean, F. & Gee, J. P. (1987). Prosodic structure and spoken word recognition. *Cognition*, 25, 135-155.
- Grossberg, S., Boardman, I. & Cohen, M. (1997). Neural dynamics of variable-rate speech categorization. *Journal of Experimental Psychology: Human Perception and Performance*, 23, 481-503.

- Hall, M. D. & Pastore, R. E. (1992). Musical duplex perception: Perception of figurally good chords with subliminal distinguishing tones. *Journal of Experimental Psychology Human Perception and Performance*, 18, 752-762.
- Hardison, D. M. (1996). Bimodal speech perception by native and nonnative speakers of English: Factors influencing the McGurk effect. *Language Learning*, 46, 3-73.
- Harris, K. S. (1977). The study of articulatory organization: Some negative progress. *Haskins Laboratories Status Report on Speech Research*, 50, Apr-Jun, 13-20.
- Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, 97, 3099-3111.
- Hintzman, D. (1986). Schema abstraction in a multiple-trace model. *Psychological Review*, 93, 411-428.
- Jelinek, F. (1997). *Statistical Methods for Speech Recognition*. Cambridge, MA: MIT Press.
- Johnson, K. (1997). *Acoustic and Auditory Phonetics*. Cambridge: Blackwell.
- Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In *Talker Variability in Speech Processing*. K. Johnson & J. W. Mullennix (Eds.) San Diego: Academic Press.
- Johnson, K., Ladefoged, P. & Lindau, M. (1993). Individual differences in vowel production. *Journal of the Acoustical Society of America*, 94, 701-714.
- Jusczyk, P. W. (1993). From general to language-specific capacities: The WRAPSA model of how speech perception develops. *Journal of Phonetics*, 21, 3-28.
- Jusczyk, P. W. (1997). *The Discovery of Spoken Language*. Cambridge, MA: MIT Press.
- Jusczyk, P. W., Cutler, A., & Redenz, N. J. (1993). Infants' preference for the predominant stress patterns of English words. *Child Development*, 64, 675-687.
- Kent, R. D. (1997). *The Speech Sciences*. San Diego : Singular Pub. Group.
- Takehi, K. (1992). Adaptability to differences between talkers in Japanese monosyllabic perception. In *Speech Perception, Production, and Linguistic Structure*. Tohkura, Y., Vatikiotis-Bateson, E., Sagisaka, T. (Eds.) Amsterdam: IOS Press. (pp.135-142).
- Kewley-Port, D. (1983). Time-varying features as correlates of place of articulation in stop consonants. *Journal of the Acoustical Society of America*, 73, 322-335.
- Kewley-Port, D. & Luce, P. A. (1984). Time-varying features of initial stop consonants in auditory spectra: A first report. *Perception and Psychophysics*, 35, 353-360.
- Kewley-Port, D., Pisoni, D. B., & Studdert-Kennedy, M. (1983). Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants. *Journal of the Acoustical Society of America*, 73, 1779-1793.

- 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. Perkell & D. H. Klatt (Eds.), *Invariance and Variability in Speech Processes*. Hillsdale, NJ: Erlbaum.
- Klatt, D. H. (1989). Review of selected models of speech perception. In W. D. Marslen-Wilson (Ed.), *Lexical Representation and Process*. Cambridge, MA: MIT Press.
- Kluender, K. R. (1994). Speech perception as a tractable problem. In M. A. Gernsbacher (Ed.), *The Handbook of Psycholinguistics*. San Diego, CA: Academic Press.
- Kluender, K. R., Diehl, R. L. & Kileen, P. R. (1987). Japanese quail can learn phonetic categories. *Science*, 237, 1195-1197.
- Kuhl, P. K. (1993). Early linguistic experience and phonetic perception: Implications for theories of developmental speech perception. *Journal of Phonetics*, 21, 125-139.
- Kuhl, P. K. & Miller, J. D. (1975). Speech Perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants. *Science*, 190, 69-72.
- Kuhl, P. K. & Padden, D. M. (1983). Enhanced discriminability at the phonetic boundaries for the place feature in macaques. *Journal of the Acoustical Society of America*, 73, 1003-1010.
- Ladefoged, P. (1996). *Elements of Acoustic Phonetics, 2nd Edition*. Chicago: Chicago University Press.
- Ladefoged, P. & Broadbent, D. E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America*, 29, 98-104.
- Ladefoged, P. & Maddieson, I. (1996). *The Sounds of the World's Languages*. Cambridge, MA: Blackwell.
- Lambert, D., Cummings, K., Rutledge, J. & Clements, M. (1994). Synthesizing multistyle speech using the Klatt synthesizer. *Journal of the Acoustical Society of America*, 95, 2979-XXXX
- Lane, H. & Tranel, B. (1971). The Lombard sign and the role of hearing in speech. *Journal of Speech and Hearing Research*, 14, 677-709.
- Liberman, A. M. (1970). Some characteristics of perception in the speech mode. In D. A. Hamburg (Ed.), *Perception and Its Disorders, Proceedings of A. R. N. M. D.* Baltimore: Williams and Wilkins.
- Liberman, A. M., Cooper, F. S., Harris, K. S. & MacNeilage, P. F. (1963). A motor theory of speech perception. In *Proceedings of the Speech Communication Seminar, Stockholm 1962*. Stockholm: Royal Institute of Technology, D3.
- Liberman, A. L., Cooper, F. S., Shankweiler, D. P. & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, 74, 431-461.

- Lieberman, 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.
- Lieberman, A. M., Harris, K. S., Hoffman, H. S. & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54, 358-368.
- Lieberman, A. M. & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, 21, 1-36.
- Lieberman, A. M. & Mattingly, I. G. (1989). A specialization for speech perception. *Science*, 243, 489-494.
- Lieberman, I. Y., Shankweiler, D., Fischer, F. W., & Carter, B. (1974). Explicit syllable and phoneme segmentation in the young child. *Journal of Experimental Child Psychology*, 18, 201-212.
- Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In W. J. Hardcastle & A. Marchal (Eds.), *Speech Production and Speech Modelling*. The Netherlands: Kluwer Academic.
- Lindblom, B. (1996). Role of articulation in speech perception: Clues from production. *Journal of the Acoustical Society of America*, 99, 1683-1692.
- Lippman, R. P. (1997). Speech recognition by machines and humans. *Speech Communication*, 22, 1-15.
- Lisker, L. & Abramson, A. D. (1967). The voicing dimension: Some experiments in comparative phonetics. *Proceedings of the 6th International Congress of Phonetic Sciences*. Prague: Academia.
- Lively, S. E., Logan, J. D., & Pisoni, D. B. (1993). Training Japanese listeners to identify English /t/ and /l/: II. The role of phonetic environment and talker variability in learning new perceptual categories. *Journal of the Acoustical Society of America*, 94, 1242-1255.
- Logan, J. S., Greene, B. G. & Pisoni, D. B. (1989). Segmental intelligibility of synthetic speech produced by rule. *Journal of the Acoustical Society of America*, 86, 566-581.
- Logan, J. S., Lively, S. E. & Pisoni, D. B. (1991). Training Japanese listeners to identify English /t/ and /l/: A first report. *Journal of the Acoustical Society of America*, 89, 874-886.
- Lowerre, B. T. & Reddy, D. R. (1980). The harpy speech understanding system. In W. A. Lea (Ed.), *Trends in Speech Recognition*. Englewood-Cliffs: Prentice Hall.
- Luce, P. A., Feustal, T. C., & Pisoni, D. B. (1983). Capacity demands in short-term memory for synthetic and natural speech. *Human Factors*, 25, 17-32.
- Luce, P. A. & Pisoni, D. B. (1998). Recognizing Spoken Words: The Neighborhood Activation Model. *Ear and Hearing*, 19, 1-39.
- Luce, R. D. (1959). *Individual Choice Behavior*. New York: Wiley.

- MacNeilage, P.F. (1970). Motor control of serial ordering of speech. *Psychological Review*, 77, 182-196.
- Mann, V. A. & Repp, B. H. (1980). Influence of vocalic context on perception of the [Σ]-[s] distinction. *Perception and Psychophysics*, 28, 213-228.
- Marslen-Wilson, W.G. (1987). Functional parallelism in spoken word recognition. *Cognition*, 25, 71-102.
- Marslen-Wilson, W. G. & Welsh, A. (1978). Processing interactions and lexical access during word-recognition in continuous speech. *Cognitive Psychology*, 10, 29-63.
- Massaro, D. W. (1987). *Speech perception by ear and eye. A paradigm for psychological inquiry*. Hillsdale NJ: Erlbaum.
- Massaro, D. W. (1989). Testing between the TRACE model and the fuzzy logical model of speech perception. *Cognitive Psychology*, 21, 398-421.
- Massaro, D. W. (1994). Psychological aspects of speech perception: Implications for research and theory. In M. A. Gernsbacher (Ed.), *Handbook of Psycholinguistics*. San Diego: Academic Press.
- Massaro, D. W. (1998). *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*. Cambridge, MA: MIT Press.
- Massaro, D. W. & Oden, G. C. (1995). Independence of lexical context and phonological information in speech perception. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21, 1053-1064.
- McClelland, J. L. (1991). Stochastic interactive processes and the effect of context on perception. *Cognitive Psychology*, 23, 1-44.
- McClelland, J. L. & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18, 1-86.
- McClelland, J. L. & Rumelhart, D. E. (1981). An interactive model of context effects in letter perception: An account of basic findings. *Psychological Review*, 88, 375-407.
- McGurk, H. & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264, 746-748.
- McQueen, J. M. (1991). The effect of the lexicon on phonetic categorization: Stimulus quality in word-final ambiguity. *Journal of Experimental Psychology: Human Perception and Performance*, 17, 433-443.
- McQueen, J. M., Norris, D. & Cutler, A. (1994). Competition in spoken word recognition: Spotting words in other words. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 20, 621-638.
- Mehler, J., Jusczyk, P. W., Lambertz, G., Halsted, H., Bertoincini, J. & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition*, 29, 144-178.
- Miller, G. A. (1956). The perception of speech. In M. Halle (Ed.). *For Roman Jakobson: Essays on the occasion of his sixtieth birthday, 11 October 1956*. The Hague, Mouton.

- Miller, J. D., Wier, C. C., Pastore, R., Kelly, W. J., & Dooling, R. J. (1976). Discrimination and labeling of noise-buzz sequences with varying noise-lead times: An example of categorical perception. *Journal of the Acoustical Society of America*, *60*, 410-417.
- Miller, G. A. & Licklider, J. C. R. (1950). The intelligibility of interrupted speech. *Journal of the Acoustical Society of America*, *22*, 167-173.
- Miller, G. A. & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America*, *27*, 338-352..
- Miyawaki, K., Strange, W., Verbrugge, R., Liberman, A. M., Jenkins, J. J., & Fujimura, O. (1975). An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English. *Perception and Psychophysics*, *18*, 331-340.
- Mullennix, J. W. (1997). On the nature of perceptual adjustments to voice. In K. Johnson & J. W. Mullennix (Eds.), *Talker Variability in Speech Processing*. San Diego: Academic Press.
- Mullennix, J. W. & Pisoni, D. B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics*, *47*, 379-390.
- Munhall, K. G., Gribble, P., Sacco, L. & Ward, M. (1996). Temporal constraints on the McGurk effect. *Perception and Psychophysics*, *58*, 351-362.
- Nakatani, L. & Dukes, K. (1977). Locus of segmental cues for word juncture. *Journal of the Acoustical Society of America*, *62*, 714-719.
- Nearey, T. M. (1990). The segment as a unit of speech perception. *Journal of Phonetics*, *18*, 347-373.
- Norris, D. (1994). Shortlist: A connectionist model of continuous recognition. *Cognition*, *52*, 189-234.
- Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *13*, 87-108.
- Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *14*, 700-708.
- Nygaard, L. C. & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception and Psychophysics*, *60*, 355-376.
- Nygaard, L. C., Sommers, M. S. & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, *5*, 42-46.
- Oden, G. C. & Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review*, *85*, 172-191.
- Ohala, J. J. (1996). Speech perception is hearing sounds, not tongues. *Journal of the Acoustical Society of America*, *99*, 1718-1725.

- 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.
- Pastore, R. E., Schmuckler, M. A., Rosenblum, L., & Szczesiul, R. (1983). Duplex perception with musical stimuli. *Perception and Psychophysics*, 33, 469-474.
- Peterson, G. E. & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, 24, 175-184.
- Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception and Psychophysics*, 13, 253-260.
- Pisoni, D. B. (1977). Identification and discrimination of the relative onset of two-component tones: Implications for the perception of voicing in stops. *Journal of the Acoustic Society of America*, 61, 1352-1361.
- Pisoni, D. B. (1997a). Perception of synthetic speech. In *Progress in Speech Synthesis*. J. P. H. Van-Santen, R. W. Sproat, J. P. Olive & J. Hirschberg (Eds.) New York: Springer-Verlag.
- Pisoni, D. B. (1997b). Some thoughts on "normalization" in speech perception. In K. Johnson & J. W. Mullennix (Eds.) *Talker Variability in Speech Processing*. San Diego: Academic Press.
- Pisoni, D. B. & Lazarus, J. H. (1974). Categorical and non-categorical modes of speech perception along the voicing continuum. *Journal of the Acoustical Society of America*, 55, 328-333.
- Pisoni, D. B., Nusbaum, H. C. & Greene, B. G. (1985). Perception of synthetic speech generated by rule. *Proceedings of the IEEE*, 73, 1665-1676.
- Pitt, M. A. (1995). The locus of the lexical shift in phoneme identification. *Journal of Experimental Psychology Learning, Memory, and Cognition*, 21, 1037-1052.
- Pitt, M. A. & Samuel, A. G. (1993). An empirical and meta-analytic evaluation of the phoneme identification task. *Journal of Experimental Psychology: Human Perception and Performance*, 19, 699-725.
- Posner, M. I. (1969). Abstraction and the process of recognition. In J. T. Spence & G. H. Bower (Eds.), *The Psychology of Learning and Motivation*. New York: Academic Press.
- Ralston, J. V., Pisoni, D. B., & Mullennix, J. W. (1995). Perception and comprehension of speech. In *Applied Speech Technology*. A. Syrdal, R. Bennett & S. Greenspan (Eds.) CRC Press: Boca Raton, FL.
- Rand, T. C. (1974). Dichotic release from masking for speech. *Journal of the Acoustical Society of America*, 55, 678-680.
- Remez, R. E., Rubin, P. E., Pisoni, D. B. & Carrell, T. (1981). Speech perception without traditional speech cues. *Science*, 212, 947-950.

- Remez, R. E., Rubin, P. E., Berns, S. M., Pardo, J. S., & Lang, J. M. (1994). On the perceptual organization of speech. *Psychological Review*, *101*, 129-156.
- Repp, B. H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin*, *92*, 81-110.
- Rose, R. C., Schroeter, J. & Sondhi, M. M. (1996). The potential role of speech production models in automatic speech recognition. *Journal of the Acoustical Society of America*, *99*, 1699-1709.
- Rossing, T. D. (1990). *The Science of Sound, 2nd Edition*. Reading, MA: Addison Wesley.
- Sachs, R. M. (1969). Vowel identification and discrimination in isolation vs. word context. *Quarterly progress report No. 93*. Cambridge, MA: Research Laboratory of Electronics, MIT.
- Salasoo, A. & Pisoni, D. B. (1985). Interaction of knowledge sources in spoken word identification. *Journal of Memory and Language*, *24*, 210-231.
- Saltzman, E. & Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, *1*, 333-382.
- Sekiyama, K. (1997). Cultural and Linguistic Factors in Audiovisual Speech Processing: The McGurk Effect in Chinese Subjects. *Perception and Psychophysics*, *59*, 73-80.
- Sekiyama, K. & Tohkura, Y. (1993). Inter-language differences in the influence of visual cues in speech perception. *Journal of Phonetics*, *21*, 427-444.
- Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J. et al. (1995). Speech recognition with primarily temporal cues. *Science*, *270*, 303-304.
- Sommers, M. S. (1996). The structural organization of the mental lexicon and its contribution to age-related declines in spoken-word recognition. *Psychology and Aging*, *11*, 333-341.
- Sommers, M. S. (1997). Stimulus variability and spoken word recognition. II. The effects of age and hearing impairment. *Journal of the Acoustical Society of America*, *101*, 2278-2288.
- Sommers, M. S., Kirk, K. I. & Pisoni, D. B. (1997). Some considerations in evaluating spoken word recognition by normal-hearing, noise-masked normal-hearing, and cochlear implant listeners. I: The effects of response format. *Ear and Hearing*, *18*, 89-99.
- Stevens, K. N. (1968). On the relations between speech movements and speech perception. *Zeitschrift fur Phonetik, Sprachwissenschaft und Kommunikationsforschung*, *21*, 102-106.
- Stevens, K. N. (1996). Understanding variability in speech: A requisite for advances in speech synthesis and recognition. *2aSC3 in Session 2aSC – Speech Communication: Speech Communication for the next decade: New directions of research, technological development, and evolving applications*. Acoustical Society of America and Acoustical Society of Japan, Third Joint Meeting, December 1996, Honolulu, Hawaii.

- Stevens, K. N. & Blumstein, S. (1981). The search for invariant acoustic correlates of phonetic features. In *Perspectives on the Study of Speech*. P. D. Eimas & J. L. Miller (Eds.), Hillsdale, NJ: Lawrence Erlbaum.
- Stevens, K. N. & House, A. S. (1955). Development of a quantitative description of vowel articulation. *Journal of the Acoustical Society of America*, 27, 484-493.
- Strange, W., Jenkins, J. J. & Johnson, T. L. (1983). Dynamic specification of coarticulated vowels. *Journal of the Acoustical Society of America*, 74, 695-705.
- Strom, N. (1997). Speaker modeling for speaker adaptation in automatic speech recognition. In K. Johnson and J. W. Mullennix (Eds.), *Talker Variability in Speech Processing*. San Diego: Academic Press.
- Studdert-Kennedy, M. (1986). Two cheers for direct realism. *Journal of Phonetics*, 14, 99-104.
- Sumby, W. H. & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, 26, 212-215.
- Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing By Eye: The Psychology of Lip-reading*. Hillsdale, NJ: Lawrence Erlbaum.
- Sussman, H. M. & Shore, J. (1996). Locus equations as phonetic descriptors of consonantal place of articulation. *Perception and Psychophysics*, 58, 936-946.
- Sussman, H. M., McCaffrey, H. A., & Matthews, S. A. (1991). An investigation of locus equations as a source of relational invariance for stop place categorization. *Journal of the Acoustical Society of America*, 90, 1309-1325.
- Sussman, H. M., Hoemeke, K. A., & Ahmed, F. S. (1993). A cross-linguistic investigation of locus equations as a phonetic descriptor for place of articulation. *Journal of the Acoustical Society of America*, 94, 1256-1268.
- Taft, M. & Hambly, G. (1986). Exploring the Cohort Model of spoken word recognition. *Cognition*, 22, 259-282.
- Treisman, M. (1978). Space or lexicon? The word frequency effect and the error response frequency effect. *Journal of Verbal Learning and Verbal Behavior*, 17, 37-59.
- Trout, J. D. (1998). The biological basis of speech: What to infer from talking to the animals. Manuscript submitted for publication.
- Vihman, M. M. (1996). *Phonological Development: The Origins of Language in the Child*. Cambridge, MA: Blackwell.
- Walley, A. C. & Metsala, J. L. (1990). The growth of lexical constraints on spoken word recognition. *Perception and Psychophysics*, 47, 267-280.
- Warren, R. (1970). Perceptual restoration of missing speech sounds. *Science*, 167, 392-393.

- Werker, J. F. (1992). Cross-language speech perception: developmental change does not involve loss. In J. Goodman & H. C. Nusbaum (Eds.), *Speech Perception and Spoken Word Recognition*. Cambridge, MA: MIT Press.
- Werker, J. F. & Polka, L. (1993). Developmental changes in speech perception: Challenges and new directions. *Journal of Phonetics*, 21, 83-101.
- Wright, R. (1997). Lexical competition and reduction in speech: A preliminary report. In *Research on Spoken Language Processing Progress Report No. 21* (pp. 471-485). Bloomington, IN: Speech Research Laboratory, Indiana University.
- Zhao Y. (1997). Overcoming speaker variability in automatic speech recognition: The speaker adaptation approach. In K. Johnson & J. W. Mullennix (Eds.), *Talker Variability in Speech Processing*. San Diego: Academic Press.

Figure captions

Figure 1. Acoustic characteristics of the utterance “You are not walking on my street,” spoken by a female adult. (a) Expanded detail of the selected section of full waveform as shown in (b). (c) A wide-band spectrogram of the entire utterance. (d) A narrow-band spectrogram of the same utterance. (e) An f_0 pitch track of the utterance. (f) A single power spectra from point X in waveform.

Figure 2. An illustration of the effects of co-articulation. The sentence is: “You lose your yellow roses early.”

Figure 3. (a) A sagittal view of the human vocal tract. (b) An average vowel space for a male speaker of American English.

Figure 4. (a) Sample identification and discrimination functions relevant to the discussion of categorical perception of voiced English stop consonants. (b) Typical stimuli and the resulting percepts in a duplex perception experiment. (c) Visual and auditory inputs that result in the “fused” McGurk percept.

Figure 5. Six different instances of the utterance “You ought to keep one at home.”

Figure 6. The basic structure of the five models of spoken word recognition discussed in this chapter, simplified for illustrative purposes.