Attention, Perception, & Psychophysics
2010, 72 (5), 1218-1227
doi:10.3758/APP.72.5.1218

Speech perception as categorization

Lori L. Holt
Carnegie Mellon University, Pittsburgh, Pennsylvania

and

Andrew J. Lotto
University of Arizona, Tucson, Arizona

Speech perception (SP) most commonly refers to the perceptual mapping from the highly variable acoustic speech signal to a linguistic representation, whether it be phonemes, diphones, syllables, or words. This is an example of *categorization*, in that potentially discriminable speech sounds are assigned to functionally equivalent classes. In this tutorial, we present some of the main challenges to our understanding of the categorization of speech sounds and the conceptualization of SP that has resulted from these challenges. We focus here on issues and experiments that define open research questions relevant to phoneme categorization, arguing that SP is best understood as perceptual categorization, a position that places SP in direct contact with research from other areas of perception and cognition.

Spoken syllables may persist in the world for mere tenths of a second. Yet, as adult listeners, we are able to gather a great deal of information from these fleeting acoustic signals. We may apprehend the physical location of the speaker, the speaker's gender, regional dialect, age, emotional state, or identity. These spatial and indexical factors are conveyed by the acoustic speech signal in parallel with the linguistic message of the speaker (Abercrombie, 1967). Although these factors are of much interest in their own right, speech perception (SP) most commonly refers to the perceptual mapping from acoustic signal to some linguistic representation, such as phonemes, diphones, syllables, words, and so forth.[1]

Most of the research in the field of SP has focused on the mapping from the acoustic speech signal to phonemes, the smallest linguistic unit that changes meaning within a particular language (e.g., /r/ and /l/ as in *rake* vs. *lake*), with the often implicit assumption that phoneme representations are a necessary step in the comprehension of spoken language. The transformation from acoustics to phonemes occurs so rapidly and automatically that it mostly escapes our notice (Näätänen & Winkler, 1999). Yet this apparent ease masks the complexity of the speech signal and the remarkable challenges inherent in phoneme perception.

As a starting point, one might presume that phoneme perception is accomplished by detecting characteristics in the acoustic signal that correspond to each phoneme or by comparing a phoneme template in memory with segments of the incoming signal. In fact, this was the presumption in the early days of SP, starting in the 1940s (see Liberman, 1996), and it led to the hope that machine speech recognition was on the horizon. However, it became clear rather quickly that SP was not a simple detection or match-to-pattern task (Liberman, Delattre, & Cooper, 1952). Although there has been a wealth of studies documenting the acoustic "cues" that can signal the identity of different phonemes (see Stevens, 2000, for a review), there is significant variability in the relationship of these cues to the intended phonemes of a speaker and the perceived phonemes of a listener. The variability is due to a multitude of sources, including differences in speaker anatomy and physiology (Fant, 1966), differences in speaking rate (Gay, 1978; Miller & Baer, 1983), effects of the surrounding phonetic context (Kent & Minifie, 1977; Öhman, 1966), and effects of the acoustic environment such as noise or reverberation (Houtgast & Steeneken, 1973). The end result of all of these sources of variability is that there appear to be few or no invariant acoustic cues to phoneme identity (Cooper, Delattre, Liberman, Borst, & Gerstman, 1952; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; but see Blumstein & Stevens, 1981, for a possible exception). This means that listeners cannot accomplish SP by simply detecting the presence or absence of cues.

In place of a simple match-to-sample or detection approach, SP is now often conceived of as a complex categorization task accomplished within a highly multidimensional space. One can conceptualize a segment of the speech signal as a point in this space representing values across multiple acoustic dimensions. In most cases, the dimensions of this space are continuous acoustic variables such as fundamental frequency, formant frequency, formant transition duration, and so forth. That is,

L. L. Holt, lholt@andrew.cmu.edu

speech stimuli are represented by continuous values, as opposed to binary values of the presence or absence of some feature. SP is the process that maps from this space onto representations of phonemes or linguistic features that subsequently define the phoneme (Jakobson, Fant, & Halle, 1952). This is an example of *categorization*, in that potentially discriminable sounds are assigned to functionally equivalent classes (Massaro, 1987).

An early example of such an acoustic space representation for phoneme classes is present in Peterson and Barney (1952), where vowel productions by adult males and females and children were displayed in terms of first and second formant ($F1$ and $F2$) frequencies. This simple distribution map demonstrates that exemplars of particular phonemes tend to cluster together in acoustic space (e.g., instances of the vowel /i/ as in *heat* tend to have low $F1$s and high $F2$s), but there is a tremendous amount of overlap among the distributions of different vowels owing to variability in speech productions (see also Hillenbrand, Getty, Clark, & Wheeler, 1995, for an update on these vowel measures, and Lisker & Abramson, 1964, for overlap in consonant voicing distributions). Presumably, listeners have to determine boundaries in order to parse these acoustic spaces and perceive the intended phonemes despite acoustic variability. Whereas there are a few auditory perceptual discontinuities that may aid in parsing acoustic space into categories in some cases (Holt, Lotto, & Diehl, 2004; Pisoni, 1977; Steinschneider et al., 2005), for the vast majority of cases listeners must determine the boundaries among phoneme categories on the basis of their experience with the language.

Unfortunately, even a perceptual categorization approach to SP does not provide easy answers to many of the questions regarding phoneme perception. In this tutorial, we present some of the main challenges to our understanding of the categorization of speech sounds, as well as the development of our conceptualization of SP that has resulted from these challenges. Because it is not possible to exhaustively review 60+ years of research and theory here, we focus on issues and experiments that define open research questions.

Challenges of Speech Sound Categorization

A major problem of mapping from multidimensional acoustic distributions to phonemes is that some of the variability in the acoustic input space is relevant to the linguistic message, some of the variability is related to characteristics of the speaker, and some of the variability is noise. To further complicate things, variation on any particular acoustic dimension could be the result of any of these sources, depending on the context. The pitch (fundamental frequency, $f0$) of the vowel in the utterance /ba/, for example, may be linguistically insignificant as it varies with the sex and age of the speaker (Klatt & Klatt, 1990), but relative pitch does serve as a linguistically reliable cue to /ba/ versus /pa/, with /pa/ having a higher pitch relative to /ba/ (House & Fairbanks, 1953).

Voice pitch is one of as many as 16 cues that can distinguish /ba/ from /pa/ (Lisker, 1986). Whereas any of these multiple cues may be informative for the speech categorization, the perceptual effectiveness of each cue varies. For example, when categorizing consonants such as /b/, /d/, and /g/, American English listeners make greater use of differences in formant transitions as opposed to frequency information in the noise burst that precedes the transitions even though both cues reliably covary with the consonants (Francis, Baldwin, & Nusbaum, 2000). Of significance, listeners' relative reliance on particular acoustic cues changes across development (see, e.g., Nittrouer, 2004) and varies depending on the listener's native language (e.g., Iverson et al., 2003). Thus, establishing the mapping from an acoustic *input* space to a *perceptual* space is a developmental process that depends on language experience.

For several months after birth, normal-hearing infants appear to parse the speech input space in the same manner (see Kuhl, 2004, and Werker & Tees, 1999, for reviews). No matter the linguistic environment in which they are developing, the basic characteristics of the human auditory system's response to speech signals dictates perception. Since speech sounds must be discriminably different enough from one another to reliably convey meaning, languages have evolved inventories of speech sounds that exploit basic human auditory function (Diehl & Lindblom, 2004; Lindblom, 1986). Thus, young infants tend to discriminate nearly any speech distinction they are presented (Kuhl, 2004). However, by the first birthday, experience with the regularities of the native language restructures the perceptual space to which speech input maps (Werker & Tees, 1984). By this time, infants developing in English-speaking environments perceive the same sounds differently, for example, than do infants developing in Swedish-speaking environments (Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992). Infants appear to have parsed the perceptual space, finding regularity relevant to the native language amid considerable acoustic variability across other dimensions.

These changes have been described as a "warping" of the perceptual space (Kuhl et al., 2008). If we imagine perceptual space as a multidimensional topography, the perceptual landscape can be described as relatively flat in early infancy, with any discontinuities arising from discontinuities in human auditory processing. With experience with the native language environment, the perceptual space is warped to reflect the regularities of the native speech input space (Kuhl, 2000; Spivey, 2007), and infants begin to perceive speech relative to the characteristics of the native language rather than solely according to psychoacoustic properties. The groundwork for reorganizing the perceptual space according to the regularities of the native language input thus begins in infancy (see Kuhl, 2000), although development of speech categories continues through childhood (see Walley, 2005). Although the development of speech categories is now widely documented, research is just beginning to uncover the learning mechanisms that guide this experience-dependent process.

A natural question that arises is, How does the initial categorization parsing based on one's native language affect the ability to learn a second language? A popular example of this issue comes from comparing English, which distinguishes /r/ from /l/, and Japanese, which does not use /r/ and /l/ to distinguish meaning and instead possesses a single lateral flap (Ladefoged & Maddieson, 1996), which overlaps with /r/ and /l/ in an acoustic space defined by the onset frequencies of the second (*F*2) and third (*F*3) formants (Lotto, Sato, & Diehl, 2004). Thus, English listeners must parse the perceptual space to best capture the linguistically relevant acoustic variability distinguishing /r/ from /l/, whereas Japanese listeners need not parse the space in quite the same manner, because variability in this region of the perceptual space is not relevant to Japanese (Best & Strange, 1992). Once the perceptual system commits to a parse of the perceptual space, there are long-term consequences for SP; the experience that we have with the sounds of our native language fundamentally shapes how we hear speech. Specifically, between-category sensitivity (e.g., an English listener distinguishing the consonants of *rock* and *lock*) is preserved, whereas within-category sensitivity (distinguishing two acoustically different instances of *rock*) is attenuated (Kuhl et al., 1992; Werker, 1994). This surely benefits our ability to communicate in a native language, but it has consequences for adults' perception and acquisition of nonnative speech categories.

An example of this is the difficulty native Japanese listeners have in perceiving English /r/ versus /l/ (Goto, 1971; Miyawaki et al., 1975). Although Japanese adults can improve their English /r/–/l/ perception and production (e.g., Bradlow, Nygaard, & Pisoni, 1999; Bradlow & Pisoni, 1999; Logan, Lively, & Pisoni, 1991; McCandliss, Fiez, Protopapas, Conway, & McClelland, 2002), it may take decades of English experience for native Japanese listeners to approach native levels of perceptual performance with English /r/–/l/ (Flege, Yeni-Komshian, & Liu, 1999), and even then, there are large individual differences in achievement (see, e.g., Slevc & Miyake, 2006). Native Japanese listeners' perceptual space has been tuned for the regularities of Japanese, and this organization is not entirely compatible with the speech input space of English.

The phenomenon of difficulty in perceiving nonnative speech categories demonstrates that speech is perceived through the lens of native language categories. Indeed, electrophysiological evidence suggests that the influence of categorization on SP is evident at very early stages of stimulus processing (e.g., Näätänen et al., 1997; Sharma & Dorman, 2000; Winkler et al., 1999; Zhang, Kuhl, Imada, Kotani, & Pruitt, 2001). The difficulties are greatest for nonnative sounds similar to native categories (Best, 1994; Flege, 1995; Harnsberger, 2001), suggesting that the warping of the perceptual space by the first language especially influences SP of acoustically similar nonnative sounds. Although the difficulties appear to be related to the age of category acquisition (Lenneberg, 1967), with adults having greater perceptual difficulty than younger listeners, much evidence suggests that this is related more to the length and degree of immersion in the second language environment than to maturation (e.g., Flege, 1995; Flege et al., 1999). Moreover, the perceptual changes introduced by parsing the perceptual space seem not to involve a loss of auditory sensitivity, since with sensitive measures adults can demonstrate an ability to distinguish difficult nonnative speech categories (Werker & Tees, 1984).

Categorization, Not Categorical Perception

It is important to distinguish the description of SP as *categorization* from the notion that SP is *categorical*. Opening almost any perception or cognition textbook to the section on speech, one is likely to find an illustration displaying perhaps the best-known pattern of SP outside the field, categorical perception (CP; see Wolfe et al., 2008). In a typical CP experiment, a series of speech sounds varying in equal physical steps along some acoustic dimension is presented to listeners, whose task is to classify them as two or more phonemes. Typically, the proportion of each category response does not vary gradually with the change in acoustic parameters. Instead, there is an abrupt shift from consistent labeling of the stimuli as one phoneme to consistent labeling as a competing phoneme across a small change in the acoustics. This is one of three hallmarks of the phenomenon of CP. A second defining characteristic of CP is the pattern of discrimination across the acoustic speech series. When listeners discriminate pairs of stimuli along the series, the resulting function is discontinuous. Discrimination is nearly perfect for stimuli that lie on opposite sides of the sharp identification/categorization boundary, whereas discrimination is very poor for pairs of stimuli that are equally acoustically distinct but lie on the same side of the identification/categorization boundary. The final characteristic of CP is that identification/categorization performance predicts discrimination performance; speech sounds that are given the same label (e.g., "ba") are difficult to discriminate, whereas those given different labels are discriminated with high accuracy (see Harnad, 1987; Studdert-Kennedy, Liberman, Harris, & Cooper, 1970).

CP was formerly thought to be a peculiarity of SP (Liberman, 1957; Liberman, Harris, Hoffman, & Griffith, 1957) and was among several perceptual phenomena that have had great impact on speech theories. Its interpretation served to ignite debates over the objects of SP and the mechanisms that support their processing (see Diehl, Lotto, & Holt, 2004, for a review). However, CP has since been observed for perception of human faces (Beale & Keil, 1995) and facial expressions (Bimler & Kirkland, 2001), music intervals (see Krumhansl, 1991, for a review), and artificial stimuli that participants learn to categorize in laboratory tasks (Livingston, Andrews, & Harnad, 1998). It is observed in the behavior of nonhuman animals as well (see Kluender, Lotto, & Holt, 2005, for a review). Moreover, the prototypical pattern of CP is not observed for all speech sounds. Its patterns are much weaker for vowels than for stop consonants like /b/ and /p/, for example (Pisoni, 1973), and sensitive methods for measuring discrimination or discrimination training can cause the peaks in discrimination at the boundaries to disappear even for consonants (Carney, Widin, & Viemeister, 1977; Samuel, 1977). Rather than a speech-specific phenomenon, CP is a far more general characteristic of how perceptual systems respond to experience with regularities in the environment (Damper &

Harnad, 2000) and, perhaps, of how time-varying signals are accommodated in perceptual memory (Mirman, Holt, & McClelland, 2004). Thus, the theoretical implications associated with CP (such as the proposition that it is a speech-specific phenomenon or that it is a qualitatively different sort of perceptual process) have not withstood empirical scrutiny.

However, although much of the controversy about the interpretation of CP has settled, CP has left an indelible mark on thinking about SP (perhaps especially among those outside the immediate field of SP). The sharp identification functions of CP are characterized by their steep boundary, but also by the relative flatness of the function within categories giving the appearance that, within a speech category, tokens are equivalent and that their acoustic variability is uninformative to the perceptual system. The classic CP pattern of responses suggests that the mapping from acoustics to speech label is discrete, such that acoustically variable instances of /ba/, for example, are mapped to "ba" irrespective of the acoustic nuances of a particular /ba/, its speaker, or its context.

Relatedly, one of the ways in which CP has left its mark is that descriptions of SP tend to describe speech *identification* instead of speech *categorization*. On the face of it, this seems a small difference, especially since these terms are often used interchangeably in the SP literature. However, *identification* (at least as it is used in other categorization literatures) is a decision about an object's unique identity that requires discrimination between similar objects. *Categorization*, on the other hand, reflects a decision about an object's type or kind requiring generalization across the perceptually discriminable physical variability of a class of objects (Palmieri & Gauthier, 2004). Whereas CP, with its suggested insensitivity to intracategory variability, is consistent with *identification*, there is much evidence that the facts of SP are better captured by *categorization*.

For example, when one exploits measures more continuous than the binary responses typical of CP tasks (e.g., was that sound /ba/ or /da/?), listeners' behavior suggests the rich internal structure of speech categories. Listeners rate some exemplars as "better" instances of a speech category than others (e.g., Iverson & Kuhl, 1995; Kuhl, 1991; Volaitis & Miller, 1992). Eyetracking paradigms further reveal that fine-grained acoustic details of an utterance affect its categorization (e.g., McMurray, Aslin, Tanenhaus, Spivey, & Subik, 2008; McMurray, Tanenhaus, & Aslin, 2002). It seems that the appearance of phonetic homogeneity in CP is largely a result of the binary response labels of CP identification tasks (Lotto & Holt, 2000). Furthermore, SP is affected by the familiarity of the voice that utters a token (Nygaard & Pisoni, 1998), suggesting that fine-grained acoustic details are retained in addition to phonemic labels. This more detailed information persists to influence word-level knowledge (Hawkins, 2003; McMurray et al., 2002) and memory (Goldinger, 1996, 1998). It appears that SP is not completely based on discrete, arbitrary labels such as phonemes (Lotto & Holt, 2000). Therefore, it is likely to be more productive to consider the mapping from the multidimensional input space to a perceptual space that has been studied by SP research as *categorization* rather than as *categorical*.

If SP is really a case of perceptual categorization, then our understanding of speech communication could benefit from what we know about general categorization processes. In fact, many of the models that have been successful for visual categorization have been applied to speech sound categorization, including classic prototype (Samuel, 1982), decision bound (Maddox, Molis, & Diehl, 2002; Nearey, 1990), and exemplar (Johnson, 1997) models. However, although perceptual categorization has long been studied in the cognitive sciences (see, e.g., Cohen & Lefebvre, 2005, for a review), the categorization challenges presented by speech signals are somewhat different from those for the visual categories that are more often studied: The speech input space is composed of mostly continuous acoustic dimensions that must be parsed into categories; there is typically no single cue that is necessary or sufficient for defining category membership; speech category exemplars are inherently temporal in nature, thereby limiting side-by-side comparisons; and information for speech categories is spread across time, thus creating segmentation issues. The evidence that exists suggests that these differences matter in understanding SP (Mirman et al., 2004). Unfortunately, the literature available to guide our understanding of the processes, abilities, and constraints of general auditory categorization is quite limited (but see Goudbeek, Smits, Swingley, & Cutler, 2005; Goudbeek, Swingley, & Smits, 2009; Guenther, Husain, Cohen, & Shinn-Cunningham, 1999; Holt & Lotto, 2006; Holt et al., 2004; Mirman et al., 2004; Wade & Holt, 2005a). Further research in auditory cognition will be needed in order to discover how auditory categorization and learning, in general, advance and limit SP (see Holt & Lotto, 2008).

The Adaptive Nature of Speech Categorization

The preceding description of SP as perceptual categorization illustrates some of the complexities in mapping from acoustics to phonemes. The reader may at this point find these complexities to be challenging but not particularly daunting. However, there is an additional level of complexity to phoneme categorization that has kept researchers busy for 60+ years. The problem was summed up well years ago by Repp and Liberman (1987) when they said that "phonetic categories are flexible" (p. 90). That is, phonetic categorization is extremely context sensitive.

One way in which context influences SP is that how speech sounds are labeled changes as a function of both the overall makeup of the stimulus set and the surrounding phonetic context. Even in classic CP tasks, the range of stimulus exemplars presented during the CP task influences the observed position of the category boundary along the stimulus series (Brady & Darwin, 1978; Rosen, 1979). The presence of comparison categories available in a task (/r/ and /l/ vs. /r/ and /l/ and /w/, for example) also influences the mapping to speech categories (Ingvalson, 2008). Thus, identical signals may be categorized as different speech sounds, depending on the characteristics of the other signals in the set in which they appear.

Adjacent phonetic context also strongly influences how a particular acoustic speech signal is categorized. For example, a syllable may be perceived as a /ga/ when preceded by the syllable /al/, but as a /da/ when preceded by /ar/ (Mann, 1980). Context dependence in SP is even observed "backward" in time, such that sounds that follow a target speech sound may influence how listeners categorize the target (e.g., Mann & Repp, 1980). The rate of speech (Miller & Liberman, 1979; Summerfield, 1981) or the acoustic characteristics of voice that produce a preceding sentence also influence how speech is categorized. Ladefoged and Broadbent (1957) demonstrated that they could shift a perceived target word from "bit" to "bet" by changing the acoustics of a preceding carrier phrase (e.g., raising or lowering the $F1$ frequencies in the phrase "Please say what this word is"). Even nonspeech contexts that mimic spectral or temporal characteristics of speech signals, but are not perceived as speech, influence speech categorization (e.g., Holt, 2005; Lotto & Kluender, 1998; Wade & Holt, 2005b). The fact that nonspeech signals shift the mapping from speech acoustics to perceptual space demonstrates that general auditory processes are involved in relating speech signals and their contexts. Effects of context also occur at multiple levels. SP can be shifted by phonotactic (Pitt & McQueen, 1998; Samuel & Pitt, 2003), lexical (Magnuson, McMurray, Tanenhaus, & Aslin, 2003; McClelland & Elman, 1986), and semantic (Borsky, Tuller, & Shapiro, 1998; Connine, 1987) context, indicating the possibility of an influence of feedback from higher level representations onto speech categorization (see McClelland, Mirman, & Holt, 2006, and Norris, McQueen, & Cutler, 2000, for reviews and debate).

So what are the cues that allow listeners to reliably map from speech input to perception of native language categories? This is a difficult question to answer, because, as described above, the "cues" for SP change radically with task and context. This fact has long been acknowledged in the literature and studied as, for example, trading relations—examining how specific acoustics cues "trade" off one another to be more or less dominant in signaling particular speech categories (e.g., Oden & Massaro, 1978; Repp, 1982). However, our attempts to relate a set of cues as the definitive signals of speech categories ultimately may be misplaced, precisely because of the inherent flexibility of SP. Listeners have exquisite sensitivity to the regularity present in acoustic signals, including speech, and appear to dynamically adjust perception to characteristics of this regularity. Moreover, the nature of this regularity appears to be task dependent; the same speech stimulus set is perceived quite differently as the task varies. This suggests that the "cues" of speech categorization, to some extent, are determined online.

Perhaps the most convincing demonstrations of the flexibility of SP come from studies demonstrating that listeners can maintain veridical perception in the face of radical distortions of the speech signal. The upshot of this work is that there do not appear to be acoustic dimensions or features that are absolutely necessary for SP. Listeners can understand a signal of three sine waves following the center frequencies of the first three formants in so-called sine-wave speech, despite the loss of the harmonic structure and fine-grained acoustic detail (Remez, Rubin, Pisoni, & Carrell, 1981). In this case, the spectral envelope defined by the formant frequencies and the temporal envelope defined by the changes in the overall amplitude of the signal across time are maintained. However, listeners can also maintain veridical SP when the spectral envelope and harmonic structure are distorted, as in the case of noise-vocoded speech (Davis, Johnsrude, Hervais-Adelman, Taylor, & McGettigan, 2005; Hervais-Adelman, Davis, Johnsrude, & Carlyon, 2008; Shannon, Zeng, Kamath, Wygonksi, & Ekelid, 1995). This distortion involves dividing the signal into a small number of frequency bands and replacing acoustic information in those bands with noise that maintains the slow amplitude changes (typically less than 50 Hz) of the frequency band.

Noise-vocoded speech is similar, in some aspects, to the signal presented to listeners with cochlear implants, particularly in its destruction of frequency resolution and harmonic detail. The amazing perceptual performance of some listeners with cochlear implants is one of the most remarkable demonstrations of SP flexibility. Despite the major differences in the signal conveyed by a cochlear implant versus ordinary auditory processing, some implanted listeners achieve normal-level SP for sounds presented in quiet (e.g., Wilson & Dorman, 2007). With some training, normal-hearing listeners can also achieve reasonably good SP performance with severely time-compressed (Dupoux & Green, 1997; Pallier, Sebastian-Gallés, Dupoux, Christophe, & Mehler, 1998), spectrally shifted (Fu & Galvin, 2003), or highly synthetic (Greenspan, Nusbaum, & Pisoni, 1988) speech signals. One can even divide the signal into 50-msec chunks, reverse each of these chunks in time (so that the chunks maintain their order, but are each reversals of original chunks), and maintain nearly 100% intelligibility (Saberi & Perrott, 1999). We can maintain normal conversations on phones with bandwidths between 300 and 3000 Hz, suggesting that all of the important information in speech is in this frequency band. But, listeners can achieve nearly 90% correct categorization performance for consonants when the signal is filtered to contain information only below 800 Hz and above 4000 Hz (Lippmann, 1996).

What does this mean for SP? It is common in the literature to see a constellation of acoustic cues associated with a speech category. This makes sense in many cases, because the task is constant, acoustics are relatively unambiguous, context is neutral, and perception is consistent. However, given the flexibility of SP detailed above, it is clear that we cannot hope to provide a definitive a priori description of the acoustic cues and dimensions that will be mapped to particular phonemes. A major challenge for SP researchers is to determine what kinds of processes allow listeners to maintain consistent perceptual performance in the face of varying acoustics and listening conditions.

Speech Communication = Speech Categorization? Perhaps Not in the Wild

Most models of language presume a mapping from acoustics to phoneme, with phonemes mapping to higher level language representations such as words (e.g., McClelland & Elman, 1986; Norris et al., 2000). However, it is worth keeping in mind that the evidence for speech categorization as a *necessary* stage of processing in everyday speech communication is not incredibly strong. For example, Broca's aphasia (which is produced by diffuse damage to the left frontal regions of the brain causing severe motor speech deficits while leaving speech recognition intact; Goodglass, Kaplan, & Barresi, 2001) may leave listeners impaired on SP tasks like classic syllable identification and discrimination CP (Blumstein, 1995), but this deficit doubly dissociates from impairments on speech *recognition* (e.g., comprehending words; Miceli, Gainotti, Caltagirone, & Masullo, 1980). Thus, the kinds of tasks that require listeners to make explicit use of phonetic information may tap differentially into processes such as attention, executive processing, or working memory in comparison with ordinary speech communication (see Hickok & Poeppel, 2007).

Spoken language possesses information and regularity at multiple levels. A single utterance of *cupcake*, for example, conveys indexical characteristics of the speaker's gender, whether she is familiar to the listener, her emotion, and her sociolinguistic background. It conveys information for the phonetic categories /kʌpkek/. Moreover, we recognize it as a real English word and link it to our semantic knowledge of cupcakes. This brief acoustic signal conveys much potential information.

It is important to remember, however, that the tasks we use to study SP differentially tap into this information. The kinds of identification and discrimination tasks that create canonical CP data highlight phonetic-level processing in identifying and differentiating /kʌ/ versus /gʌ/, whereas a lexical decision task highlights word-level knowledge of "cupcake." Moreover, listeners make greater use of fine phonetic detail when nonwords outnumber words in a stimulus set, but lexical influences predominate when the task is biased toward word recognition with a greater proportion of words (Mirman, McClelland, Holt, & Magnuson, 2008). In SP research, the kinds of tasks and stimulus sets that we present shape the perceptual processing that we observe.

Everyday speech perception "in the wild" is likely to tap into a broader set of processes than those captured in individual laboratory tasks. It is important to note that this is not to suggest that adult (or even infant or animal) listeners *cannot* categorize speech; there is abundant evidence that they can. Rather, these data suggest that the cognitive and perceptual processes involved in speech categorization and those in online perception of fluent speech may not be one and the same. Although this possibility is not always acknowledged in research in SP, it is significant to our ultimate understanding of how SP relates to spoken language more generally.

Open Questions: Speech Perception in an Auditory Cognitive Neuroscience Framework

At first blush, the caveat above would seem to diminish the importance of studying and understanding speech categorization. On the contrary, however, the 60+ year history of SP research and its documentation of the multidimensional acoustic cues that covary with speech categories have provided what might be an unparalleled understanding of a natural, complex, ecologically valid perceptual categorization space (Kluender, 1994). Even the perceptual dimensions of faces—another prominent ecologically relevant perceptual category space—have not been studied in this detail. What is more, categorization within the highly multidimensional "speech space" (to compare to the "face space" considered in visual face categorization; Valentine, 1991) is completely dependent on experience with a native language. Perhaps no other domain is so rich in its potential for understanding perceptual categorization.

There remains much to learn. Beyond informing our understanding of perceptual categorization and auditory processing, generally speaking, SP extends to many core areas of cognitive science. As categorization, SP offers a platform from which to investigate development (Kuhl, 2004), learning (Holt, Lotto, & Kluender, 1998), adult plasticity (McClelland, 2001), and the prospect of critical periods in human learning (Flege, 1995). The multiple sources of information that covary with the acoustic speech signal provide an opportunity for understanding cross-modal integration (Massaro, 1998) and the role of feedback in language processing (McClelland et al., 2006). Classic issues of cognitive science such as working memory (Frankish, 2008), attention (Francis & Nusbaum, 2002), and the interplay of production and perception (Galantucci, Fowler, & Turvey, 2006) are all pieces of the puzzle in understanding SP. Moreover, the special status of speech as a human communication signal provides an opportunity for even further significant extensions. Research is just beginning to uncover how social cues support speech category acquisition (Kuhl, 2007) and how personality variables may predict the degree to which information in the speech signal is integrated (Stewart & Ota, 2008).

Studying SP informs us also about the general characteristics of auditory perception and cognition. Our understanding of auditory processing has come largely from studies of simple sounds such as tones, clicks, and noise bursts. By contrast, speech is much more like the complex sounds that our auditory systems have evolved to process (Lewicki, 2002; Smith & Lewicki, 2006). As such, it is perhaps even better situated to reveal the nature of relatively poorly understood (at least in comparison with vision) processes of auditory perception and cognition. Already, studying speech categorization has provided information about the kinds of processing that the auditory system must accomplish (e.g., Holt, 2005). SP, with its complex, multidimensional input space and experience-dependent perceptual space, can reveal characteristics of general auditory processing that are just not apparent with simple acoustic stimuli.

Summary

SP is traditionally studied as the mapping from acoustics to phonemes. We have argued here that this process is best understood as one of perceptual categorization, a position that places SP in direct contact with research from other areas of perception and cognition. Whereas the study of SP has long been relegated to the periphery of cognitive science as a "special" perceptual system that can tell us little about general issues of human behavior, the latest research in SP guides us away from the classic way of thinking about SP, to consider categorization rather than identification, the regularity that exists amidst variable speech acoustics as a source of rich information, and the online adaptive nature of speech categorization. These issues place SP in a central position in the cognitive and perceptual sciences.

AUTHOR NOTE

The authors were supported by collaborative awards from the National Science Foundation (BCS0746067) and the National Institutes of Health (R01DC004674). Correspondence concerning this article should be addressed to L. L. Holt, Department of Psychology, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213 (e-mail: lholt@andrew.cmu.edu).

REFERENCES

Abercrombie, D. (1967). *Elements of general phonetics*. Chicago: Aldine.

Beale, J. M., & Keil, F. C. (1995). Categorical effects in the perception of faces. *Cognition*, **57**, 217-239.

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

Best, C. T., & Strange, W. (1992). Effects of phonological and phonetic factors on cross-language perception of approximants. *Journal of Phonetics*, **20**, 305-330.

Bimler, D., & Kirkland, J. (2001). Categorical perception of facial expressions of emotion: Evidence from multidimensional scaling. *Cognition & Emotion*, **15**, 633-658.

Blumstein, S. E. (1995). The neurobiology of the sound structure of language. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 915-929). Cambridge, MA: MIT Press.

Blumstein, S. E., & Stevens, K. N. (1981). Phonetic features and acoustic invariance in speech. *Cognition*, **10**, 25-32.

Borsky, S., Tuller, B., & Shapiro, L. P. (1998). "How to milk a coat": The effects of semantic and acoustic information on phoneme categorization. *Journal of the Acoustical Society of America*, **103**, 2670-2676.

Bradlow, A. R., Nygaard, L. C., & Pisoni, D. B. (1999). Effects of talker, rate, and amplitude variation on recognition memory for spoken words. *Perception & Psychophysics*, **61**, 206-219.

Bradlow, A. R., & Pisoni, D. B. (1999). Recognition of spoken words by native and non-native listeners: Talker-, listener- and item-related factors. *Journal of the Acoustical Society of America*, **106**, 2074-2085.

Brady, S. A., & Darwin, C. J. (1978). Range effect in the perception of voicing. *Journal of the Acoustical Society of America*, **63**, 1556-1558.

Carney, A. E., Widin, G. P., & Viemeister, N. F. (1977). Noncategorical perception of stop consonants differing in VOT. *Journal of the Acoustical Society of America*, **62**, 961-970.

Cohen, H., & Lefebvre, C. (2005). *Handbook of categorization in cognitive science*. Amsterdam: Elsevier.

Colin, C., & Radeau, M. (2003). Les illusions McGurk dans la parole: 25 ans de recherches [The McGurk illusions in speech: 25 years of research]. *L'Année Psychologique*, **103**, 497-542. [With summary in English] doi:10.3406/psy.2003.29649

Connine, C. M. (1987). Constraints on interactive processes in auditory word recognition: The role of sentence context. *Journal of Memory & Language*, **16**, 527-538. doi:10.1016/0749-596X(87)90138-0

Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M., & Gerstman, L. J. (1952). Some experiments on the perception of synthetic speech sounds. *Journal of the Acoustical Society of America*, **24**, 597-606.

Damper, R. I., & Harnad, S. R. (2000). Neural network models of categorical perception. *Perception & Psychophysics*, **62**, 843-867.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology: General*, **134**, 222-241.

Diehl, R. L., & Lindblom, B. (2004). Explaining the structure of feature and phoneme inventories. In S. Greenberg, W. Ainsworth, A. Popper, & R. Fay (Eds.), *Speech processing in the auditory system* (pp. 101-162). New York: Springer.

Diehl, R. L., Lotto, A. J., & Holt, L. L. (2004). Speech perception. *Annual Review of Psychology*, **55**, 149-179.

Dupoux, E., & Green, K. (1997). Perceptual adjustment to highly compressed speech: Effects of talker and rate changes. *Journal of Experimental Psychology: Human Perception & Performance*, **23**, 914-927.

Fant, G. (1966). *A note on vocal tract size factors and non-uniform F-pattern scalings*. (Speech Transmission Laboratory Quarterly Project Status Report No. 4, pp. 22-30). Stockholm: Royal Institute of Technology.

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 233-277). Baltimore: York Press.

Flege, J. E., Yeni-Komshian, G. H., & Liu, S. (1999). Age constraints on second-language acquisition. *Journal of Memory & Language*, **41**, 78-104.

Francis, A. L., Baldwin, K., & Nusbaum, H. C. (2000). Effects of training on attention to acoustic cues. *Perception & Psychophysics*, **62**, 1668-1680.

Francis, A. L., & Nusbaum, H. C. (2002). Selective attention and the acquisition of new phonetic categories. *Journal of Experimental Psychology: Human Perception & Performance*, **28**, 349-366.

Frankish, C. (2008). Precategorical acoustic storage and the perception of speech. *Journal of Memory & Language*, **58**, 815-836.

Fu, Q.-J., & Galvin, J. J., III (2003). The effects of short-term training for spectrally mismatched noise-band speech. *Journal of the Acoustical Society of America*, **113**, 1065-1072.

Galantucci, B., Fowler, C. A., & Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review*, **13**, 361-377.

Gay, T. (1978). Effect of speaking rate on vowel formant movements. *Journal of the Acoustical Society of America*, **63**, 223-230.

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.

Goodglass, H., Kaplan, E., & Barresi, B. (2001). *The assessment of aphasia and related disorders* (3rd ed.). Philadelphia: Lippincott Williams & Wilkins.

Goto, H. (1971). Auditory perception by normal Japanese adults of the sounds "L" and "R." *Neuropsychologia*, **9**, 317-323.

Goudbeek, M., Smits, R., Swingley, D., & Cutler, A. (2005). Acquiring auditory and phonetic categories. In H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science* (pp. 497-514). Amsterdam: Elsevier.

Goudbeek, M., Swingley, D., & Smits, R. (2009). Supervised and unsupervised learning of multidimensional acoustic categories. *Journal of Experimental Psychology: Human Perception & Performance*, **35**, 1913-1933.

Greenspan, S. L., Nusbaum, H. C., & Pisoni, D. B. (1988). Perceptual learning of synthetic speech produced by rule. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **14**, 421-433.

Guenther, F. H., Husain, F. T., Cohen, M. A., & Shinn-Cunningham, B. G. (1999). Effects of categorization and discrimination training on auditory perceptual space. *Journal of the Acoustical Society of America*, **106**, 2900-2912.

Harnad, S. (1987). *Categorical perception: The groundwork of cognition*. New York: Cambridge University Press.

Harnsberger, J. D. (2001). On the relationship between identification and discrimination of non-native nasal consonants. *Journal of the Acoustical Society of America*, **110**, 489-503.

Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics*, **31**, 373-405.

Hervais-Adelman, A., Davis, M. H., Johnsrude, I. S., & Carlyon, R. P. (2008). Perceptual learning of noise vocoded words: Effects of feedback and lexicality. *Journal of Experimental Psychology: Human Perception & Performance*, **34**, 460-474.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, **8**, 393-402.

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.

Holt, L. L. (2005). Temporally nonadjacent nonlinguistic sounds affect speech categorization. *Psychological Science*, **16**, 305-312.

Holt, L. L., & Lotto, A. J. (2006). Cue weighting in auditory categorization: Implications for first and second language acquisition. *Journal of the Acoustical Society of America*, **119**, 3059-3071.

Holt, L. L., & Lotto, A. J. (2008). Speech perception within an auditory cognitive science framework. *Current Directions in Psychological Science*, **17**, 42-46.

Holt, L. L., Lotto, A. J., & Diehl, R. L. (2004). Auditory discontinuities interact with categorization: Implications for speech perception. *Journal of the Acoustical Society of America*, **116**, 1763-1773.

Holt, L. L., Lotto, A. J., & Kluender, K. R. (1998). Incorporating principles of general learning in theories of language acquisition. In M. C. Gruber, D. Higgins, K. S. Olson, & T. Wysocki (Eds.), *Chicago Linguistic Society—Vol. 34: The panels* (pp. 253-268). Chicago: Chicago Linguistic Society.

House, A. S., & Fairbanks, G. (1953). The influence of consonant environment upon the secondary acoustical characteristics of vowels. *Journal of the Acoustical Society of America*, **25**, 105-113.

Houtgast, T., & Steeneken, H. J. M. (1973). The modulation transfer function in room acoustics as a predictor of speech intelligibility. *Journal of the Acoustical Society of America*, **54**, 557.

Ingvalson, E. M. (2008). *Predicting F3 usage in* /r–l/ *perception and production by native Japanese speakers*. Unpublished doctoral dissertation, Carnegie Mellon University.

Iverson, P., & Kuhl, P. K. (1995). Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. *Journal of the Acoustical Society of America*, **97**, 553-562.

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., & Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition*, **87**, B47-B57.

Jakobson, R. C., Fant, G. M., & Halle, M. (1952). *Preliminaries to speech analysis: The distinctive features and their correlates* (Tech. Rep. No. 13). Cambridge, MA: MIT, Acoustics Laboratory.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 145-165). San Diego: Academic Press.

Kent, R. D., & Minifie, F. D. (1977). Coarticulation in recent speech production models. *Journal of Phonetics*, **5**, 115-133.

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, **87**, 820-857.

Kluender, K. R. (1994). Speech perception as a tractable problem in cognitive science. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 173-217). San Diego: Academic Press.

Kluender, K. R., Lotto, A. J., & Holt, L. L. (2005). Contributions of nonhuman animal models to understanding human speech perception. In S. Greenberg & W. Ainsworth (Eds.), *Listening to speech: An auditory perspective* (pp. 203-220). New York: Oxford University Press.

Krumhansl, C. L. (1991). Music psychology: Tonal structures in perception and memory. *Annual Review of Psychology*, **42**, 277-303.

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics*, **50**, 93-107.

Kuhl, P. K. (2000). Language, mind, and brain: Experience alters perception. In M. S. Gazzaniga (Ed.), *The new cognitive neurosciences* (2nd ed., pp. 99-115). Cambridge, MA: MIT Press.

Kuhl, P. K. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience*, **5**, 831-843.

Kuhl, P. K. (2007). Is speech learning "gated" by the social brain? *Developmental Science*, **10**, 110-120.

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2008). Phonetic learning as a pathway to language: New data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society B*, **363**, 979-1000.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, **255**, 606-608.

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*. Oxford: Blackwell.

Lenneberg, E. H. (1967). *Biological foundations of language*. New York: Wiley.

Lewicki, M. S. (2002). Efficient coding of natural sounds. *Nature Neuroscience*, **5**, 356-363.

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

Liberman, A. M. (1996). *Speech: A special code*. Cambridge, MA: MIT Press.

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

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

Liberman, 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.

Lindblom, B. (1986). Phonetic universals in vowel systems. In J. J. Ohala & J. J. Jaeger (Eds.), *Experimental phonology* (pp. 13-44). Orlando, FL: Academic Press.

Lippmann, R. P. (1996). Accurate consonant perception without mid-frequency speech energy. *IEEE Transactions on Speech & Audio Processing*, **4**, 66.

Lisker, L. (1986). "Voicing" in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees. *Language & Speech*, **29**, 3-11.

Lisker, L., & Abramson, A. S. (1964). A cross-linguistic study of voicing in initial stops: Acoustical measurements. *Word*, **20**, 384-422.

Livingston, K. R., Andrews, J. K., & Harnad, S. (1998). Categorical perception effects induced by category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **24**, 732-753.

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *Journal of the Acoustical Society of America*, **89**, 874-886.

Lotto, A. J., & Holt, L. L. (2000). The illusion of the phoneme. In S. J. Billings, J. P. Boyle, & A. M. Griffith (Eds.), *Chicago Linguistic Society—Vol. 35: The panels* (pp. 191-204). Chicago: Chicago Linguistic Society.

Lotto, A. J., & Kluender, K. R. (1998). General contrast effects in speech perception: Effect of preceding liquid on stop consonant identification. *Perception & Psychophysics*, **60**, 602-619.

Lotto, A. J., Sato, M., & Diehl, R. L. (2004). Mapping the task for the second language learner: The case of Japanese acquisition of /r/ and /l/. In J. Slifka, S. Manuel, & M. Matthies (Eds.), *From sound to sense: 50+ years of discoveries in speech communication* [Online conference proceedings]. Cambridge, MA: MIT.

Maddox, W. T., Molis, M. R., & Diehl, R. L. (2002). Generalizing a neuropsychological model of visual categorization to auditory categorization of vowels. *Perception & Psychophysics*, **64**, 584-597.

Magnuson, J. S., McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2003). Lexical effects on compensation for coarticulation: The ghost of Christmash past. *Cognitive Science*, **27**, 285-298.

Mann, V. A. (1980). Influence of preceding liquid on stop-consonant perception. *Perception & Psychophysics*, **28**, 407-412.

Mann, V. A., & Repp, B. H. (1980). Influence of vocalic context on perception of the [ʃ]-[s] distinction. *Perception & Psychophysics*, **28**, 213-228.

Massaro, D. W. (1987). Categorical partition: A fuzzy-logical model of categorization behavior. In S. Harnad (Ed.), *Categorical perception: The groundwork of cognition* (pp. 254-283). Cambridge: Cambridge University Press.

Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA: MIT Press, Bradford Books.

McCandliss, B. D., Fiez, J. A., Protopapas, A., Conway, M., & McClelland, J. L. (2002). Success and failure in teaching the [r]–[l] contrast to Japanese adults: Tests of a Hebbian model of plasticity and stabilization in spoken language perception. *Cognitive, Affective, & Behavioral Neuroscience*, **2**, 89-108.

McClelland, J. L. (2001). Failures to learn and their remediation: A Hebbian account. In J. L. McClelland & R. S. Siegler (Eds.), *Mechanisms of cognitive development: Behavioral and neural perspectives* (pp. 97-122). Mahwah, NJ: Erlbaum.

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

McClelland, J. L., Mirman, D., & Holt, L. L. (2006). Are there interactive processes in speech perception? *Trends in Cognitive Sciences*, **10**, 363-369.

McMurray, B., Aslin, R. N., Tanenhaus, M. K., Spivey, M. J., & Subik, D. (2008). Gradient sensitivity to within-category variation in words and syllables. *Journal of Experimental Psychology: Human Perception & Performance*, **34**, 1609-1631.

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition*, **86**, B33-B42.

Miceli, G., Gainotti, G., Caltagirone, C., & Masullo, C. (1980). Some aspects of phonological impairment in aphasia. *Brain & Language*, **11**, 159-169.

Miller, J. L., & Baer, T. (1983). Some effects of speaking rate on the production of /b/ and /w/. *Journal of the Acoustical Society of America*, **73**, 1751-1755.

Miller, J. L., & Liberman, A. M. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception & Psychophysics*, **25**, 457-465.

Mirman, D., Holt, L. L., & McClelland, J. L. (2004). Categorization and discrimination of nonspeech sounds: Differences between steady-state and rapidly-changing acoustic cues. *Journal of the Acoustical Society of America*, **116**, 1198-1207.

Mirman, D., McClelland, J. L., Holt, L. L., & Magnuson, J. S. (2008). Effects of attention on the strength of lexical influences on speech perception: Behavioral experiments and computational mechanisms. *Cognitive Science*, **32**, 398-417.

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 & Psychophysics*, **18**, 331-340.

Näätänen, R., Lehtokoski, A., Lennes, M., Cheour, M., Huotilainen, M., Iivonen, A., et al. (1997). Language-specific phoneme representations revealed by electric and magnetic brain responses. *Nature*, **385**, 432-434.

Näätänen, R., & Winkler, I. (1999). The concept of auditory stimulus representation in cognitive neuroscience. *Psychological Bulletin*, **125**, 826-859.

Nearey, T. M. (1990). The segment as a unit of speech perception. *Journal of Phonetics*, **18**, 347-373.

Nittrouer, S. (2004). The role of temporal and dynamic signal components in the perception of syllable-final stop voicing by children and adults. *Journal of the Acoustical Society of America*, **115**, 1777-1790.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral & Brain Sciences*, **23**, 299-370.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, **60**, 355-376.

Oden, G. C., & Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review*, **85**, 172-191.

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

Pallier, C., Sebastian-Gallés, N., Dupoux, E., Christophe, A., & Mehler, J. (1998). Perceptual adjustment to time-compressed speech: A cross-linguistic study. *Memory & Cognition*, **26**, 844-851.

Palmieri, T. J., & Gauthier, I. (2004). Visual object understanding. *Nature Reviews Neuroscience*, **5**, 291-303.

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 & Psychophysics*, **13**, 253-260.

Pisoni, D. B. (1977). Identification and discrimination of the relative onset time of two component tones: Implications for voicing perception in stops. *Journal of the Acoustical Society of America*, **61**, 1352-1361.

Pitt, M. A., & McQueen, J. M. (1998). Is compensation for coarticulation mediated by the lexicon? *Journal of Memory & Language*, **39**, 347-370.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science*, **212**, 947-950.

Repp, B. H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin*, **92**, 81-110.

Repp, B. H., & Liberman, A. M. (1987). Phonetic category boundaries are flexible. In S. Harnad (Ed.), *Categorical perception: The groundwork of cognition* (pp. 89-112). Cambridge: Cambridge University Press.

Rosen, S. M. (1979). Range and frequency effects in consonant categorization. *Journal of Phonetics*, **7**, 393-402.

Rosenblum, L. D. (2005). Primacy of multimodal speech perception. In D. B. Pisoni & R. E. Remez (Eds.), *The handbook of speech perception* (pp. 51-78). Oxford: Blackwell.

Saberi, K., & Perrott, D. R. (1999). Cognitive restoration of reversed speech. *Nature*, **398**, 760.

Samuel, A. G. (1977). The effect of discrimination training on speech perception: Noncategorical perception. *Perception & Psychophysics*, **22**, 321-330.

Samuel, A. G. (1982). Phonetic prototypes. *Perception & Psychophysics*, **31**, 307-314.

Samuel, A. G., & Pitt, M. A. (2003). Lexical activation (and other factors) can mediate compensation for coarticulation. *Journal of Memory & Language*, **48**, 416-434.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, **270**, 303-304.

Sharma, A., & Dorman, M. F. (2000). Neurophysiologic correlates of cross-language phonetic perception. *Journal of the Acoustical Society of America*, **107**, 2697-2703.

Slevc, L. R., & Miyake, A. (2006). Individual differences in second-language proficiency: Does musical ability matter? *Psychological Science*, **17**, 675-681.

Smith, E. C., & Lewicki, M. S. (2006). Efficient auditory coding. *Nature*, **439**, 978-982.

Spivey, M. (2007). *The continuity of mind*. New York: Oxford University Press.

Steinschneider, M., Volkov, I. O., Fishman, Y. I., Oya, H., Arezzo, J. C., & Howard, M. A., III (2005). Intracortical responses in human and monkey primary auditory cortex support a temporal processing mechanism for encoding of the voice onset time phonetic parameter. *Cerebral Cortex*, **15**, 170-186.

Stevens, K. N. (2000). *Acoustic phonetics.* Cambridge, MA: MIT Press.

Stewart, M. E., & Ota, M. (2008). Lexical effects on speech perception in individuals with "autistic" traits. *Cognition*, **109**, 157-162.

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

Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception & Performance*, **7**, 1074-1095.

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *Quarterly Journal of Experimental Psychology*, **43A**, 161-204.

Volaitis, L. E., & Miller, J. L. (1992). Phonetic prototypes: Influence of place of articulation and speaking rate on the internal structure of voicing categories. *Journal of the Acoustical Society of America*, **92**, 723-735.

Wade, T., & Holt, L. L. (2005a). Incidental categorization of spectrally complex non-invariant auditory stimuli in a computer game task. *Journal of the Acoustical Society of America*, **118**, 2618-2633.

Wade, T., & Holt, L. L. (2005b). Perceptual effects of preceding nonspeech rate on temporal properties of speech categories. *Perception & Psychophysics*, **67**, 939-950.

Walley, A. C. (2005). Speech perception in childhood. In D. B. Pisoni & R. E. Remez (Eds.), *The handbook of speech perception* (pp. 449-468). Oxford: Blackwell.

Werker, J. F. (1994). Cross-language speech perception: Development change does not involve loss. In J. C. Goodman & H. C. Nusbaum (Eds.), *The development of speech perception: The transition from speech sounds to spoken words* (pp. 93-120). Cambridge, MA: MIT Press.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior & Development*, **7**, 49-63.

Werker, J. F., & Tees, R. C. (1999). Influences on infant speech processing: Toward a new synthesis. *Annual Review of Psychology*, **50**, 509-535.

Wilson, B., & Dorman, M. F. (2007). The surprising performance of present-day cochlear implants. *IEEE Transactions on Biomedical Engineering*, **54**, 969-972.

Winkler, I., Kujala, T., Tiitinen, H., Sivonen, P., Alku, P., Lehtokoski, A., et al. (1999). Brain responses reveal the learning of foreign language phonemes. *Psychophysiology*, **36**, 638-642.

Wolfe, J. M., Kluender, K. R., Levi, D. M., Bartoshuk, L. M., Herz, R. S., Klatzky, R. L., et al. (2008). *Sensation and perception* (2nd ed.). Sunderland, MA: Sinauer Associates.

Zhang, Y., Kuhl, P. K., Imada, T., Kotani, M., & Pruitt, J. (2001). Brain plasticity in behavioral and neuromagnetic measures: A perceptual training study [Abstract]. *Journal of the Acoustical Society of America*, **110**, 2687.

NOTE

1. Speech is not conveyed solely by sound. SP research has studied the influence of other important sources of information, especially visual information from the face (for a review, see Colin & Radeau, 2003). Some have argued that SP is best considered amodal (Rosenblum, 2005), whereas others have fruitfully used speech as a means of investigating multimodal integration from separate sources of information (Massaro, 1998). Nonetheless, SP is possible when only acoustic information is present (e.g., over a telephone), and since the majority of SP research has focused on the acoustic mapping, we highlight it in this review.

(Manuscript received December 31, 2008;
revision accepted for publication March 12, 2010.)