CHAPTER

16

Speech Perception
The View from the Auditory System

Andrew J. Lotto[1] *and Lori L. Holt*[2]

[1]Speech, Language, & Hearing Sciences, University of Arizona, Tucson, AZ, USA; [2]Department of Psychology and the Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, PA, USA

16.1 INTRODUCTION

For much of the past 50 years, the main theoretical debate in the scientific study of speech perception has focused on whether the processing of speech sounds relies on neural mechanisms that are specific to speech and language or whether general perceptual/cognitive processes can account for all of the relevant phenomena. Starting with the first presentations of the Motor Theory of Speech Perception by Alvin Liberman and colleagues (Liberman, Cooper, Harris, MacNeilage, & Studdert-Kennedy, 1964; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Studdert-Kennedy, Liberman, Harris, & Cooper, 1970) and the critical reply from Harlan Lane (1965), many scientists defended "all-or-none" positions on the necessity of specialized speech processes, and much research was dedicated to demonstrations of phenomena that were purported to require general or speech-specific mechanisms (see Diehl, Lotto, & Holt, 2004 for a review of the theoretical commitments behind these positions). Whereas the "speech-is-special" debate continues to be relevant (Fowler, 2008; Lotto, Hickok, & Holt, 2009; Massaro & Chen, 2008; Trout, 2001), the focus of the field has moved toward more subtle distinctions concerning the relative roles of perceptual, cognitive, motor, and linguistic systems in speech perception and how each of these systems interacts in the processing of speech sounds. The result has been an opportunity to develop more plausible and complete models of speech perception/production (Guenther & Vladusich, 2012; Hickok, Houde, & Rong, 2011).

In line with this shift in focus, in this chapter we concentrate not on whether the general auditory system is sufficient for speech perception but rather on the ways that human speech communication appears to be constrained and structured on the basis of the operating characteristics of the auditory system. The basic premise is simple, with a long tradition in the scientific study of speech perception: the form of speech (at the level of phonetics and higher) takes advantage of what the auditory system does well, resulting in a robust and efficient communication system. We review here three aspects of auditory perception—discriminability, context interactions, and effects of experience—and discuss how the structure of speech appears to respect these general characteristics of the auditory system.

It should be noted that we include in our conception of the "auditory system" processes and constructs that are often considered to be "cognition," such as memory, learning, categorization, and attention (Holt & Lotto, 2010). This is in contrast to previous characterizations of "Auditorist" positions in speech perception that appeared to constrain explanations of speech phenomena to peculiarities of auditory encoding at the periphery. Most researchers who have advocated for general auditory accounts of speech perception actually propose explanations within a larger general auditory cognitive science framework (Holt & Lotto, 2008; Kluender & Kiefte, 2006). Recent findings in auditory neuroscience provide support for moving beyond simple dichotomies of perception versus cognition or top-down versus bottom-up or peripheral versus central. There have been demonstrations that manipulation of attention may affect the earliest stages of auditory encoding in the cochlea (Froehlich, Collet, Chanal, & Morgon, 1990; Garinis, Glattke, & Cone, 2011; Giard, Collet, Bouchet, & Pernier, 1994; Maison, Micheyl, & Collet, 2001) and experience with music and language

changes the neural representation of sound in the brain stem (Song, Skoe, Wong, & Kraus, 2008; Wong, Skoe, Russo, Dees, & Kraus, 2007). In line with these findings, we treat attention, categorization, and learning as intrinsic aspects of auditory processing.

16.2 EFFECTS OF AUDITORY DISTINCTIVENESS ON THE FORM OF SPEECH

At the most basic level, the characteristics of the auditory system must constrain the form of speech because the information-carrying aspects of the signal must be encoded by the system and must be able to be discriminated by listeners. Given the remarkable ability of normal-hearing listeners to discriminate spectral-temporal changes in simple sounds such as tones and noises, the resolution of the auditory system does not appear to provide much of a constraint on the possible sounds used for speech communication. The smallest discriminable frequency change for a tone of 1,000 Hz is just over 1 Hz (Wier, Jesteadt, & Green, 1977), and an increment in intensity of 1 dB for that tone will likely be detected by the listener (Jesteadt, Wier, & Green, 1977). However, it is a mistake to make direct inferences from discriminability of simple acoustic stimuli to the perception of complex sounds, such as speech. Speech perception is not a simple detection or discrimination task; it is more similar to a pattern recognition task in which the information is carried through changes in relative patterns across a complex multidimensional space. These patterns must be robustly encoded and perceptually discriminable for efficient speech communication.

To the extent that some patterns are more readily discriminable by the auditory system, they will presumably be more effective as vehicles for communication. Liljencrants and Lindblom (1972) demonstrated that one could predict the vowel inventories of languages relatively well by maximizing intervowel distances within a psychophysically scaled vowel space defined by the first two formant frequencies (in Mel scaling). For example, /i/, /a/, and /u/ are correctly predicted to be the most common set of vowels for a three-vowel language system based on the presumption that they would be most auditorily discriminable given that their formant patterns are maximally distinct in the vowel space. Vowel inventory predictions become even more accurate as one more precisely models the auditory representation of each vowel (Diehl, Lindblom, & Creeger, 2003; Lindblom, 1986). These demonstrations are in agreement with proposals that languages tend to use sounds that maximize auditory distinctiveness in balance with the value of reducing articulatory effort, such as Stevens' (1972, 1989) Quantal Theory, Lindblom's (1991) H&H Theory (which we return to below), and Ohala's (1993) models of sound change in historical linguistics.

The proposal that auditory distinctiveness is important for effective speech communication was pushed even further by the Auditory Enhancement Theory from Diehl and colleagues (Diehl & Kluender, 1987, 1989; Diehl, Kluender, Walsh, & Parker, 1991). According to Auditory Enhancement, speakers tend to combine articulations that result in acoustic changes that mutually enhance distinctiveness of the resulting sounds for the listener. For example, in English the voicing contrast between /b/ and /p/ when spoken between two vowels, such as *rabid* versus *rapid*, is signaled in part by the duration of a silent interval that corresponds to the lip closure duration, which is shorter for /b/. However, speakers also tend to lengthen the duration of the preceding vowel when producing a /b/. Kluender, Diehl, and Wright (1988) demonstrated that preceding a silent gap with a long-duration sound results in the perception of a shorter silent gap, even for nonspeech sounds; this can be considered a kind of durational contrast. Thus, when talkers co-vary short lip closure durations with longer preceding vowels and vice versa, they produce a clearer auditory distinction between /b/ and /p/. This is just one of numerous examples appearing to indicate that the need for auditory distinctiveness drives the phonetic structure of languages (Diehl, Kluender, & Walsh, 1990; Kingston & Diehl, 1995).

In addition to providing constraints on the global structure of spoken languages, there is good evidence that the *individual* behavior of speakers is influenced by the local needs of listeners for auditory distinctiveness. According to Lindblom's (1991) H(yper) & H(ypo) Theory of speech communication, speakers vary their productions from hyperarticulation to hypoarticulation depending on the contextual needs of the listener. Spoken utterances that are redundant with other sources of information or with prior knowledge may be spoken with reduced effort, resulting in reduced auditory distinctiveness. However, novel information or words that are likely to be misperceived by a listener are produced with greater clarity or hyperarticulation. In accordance with this theory, there have been many demonstrations that speakers modulate productions when speaking to listeners who may have perceptual challenges, such as hearing-impaired listeners or non-native language learners (Bradlow & Bent, 2002; Picheny, Durlach, & Braida, 1985, 1986).

Despite the continued success of the theories described, it remains a challenge to derive a valid metric of "auditory distinctiveness" for complex time-varying signals like speech (and equally difficult to

quantify "articulatory effort"). The classic psychophysical measures of frequency, intensity, and temporal resolution are simply not sufficient. The pioneering work of David Green regarding auditory profile analysis in which listeners discriminate amplitude pattern changes across a multitonal complex (Green, 1988; Green, Mason, & Kidd, 1984) was a step in the right direction because it could conceivably be applied to measuring the ability to discriminate steady-state vowel acoustics. However, vowel acoustics in real speech are much more complex and it is not clear that these measures scale up to predict intelligibility of speech at even the level of words. The future prospects of understanding how the operating characteristics of the auditory system constrain the acoustic elements used in speech communication are brighter given more recent approaches to psychoacoustic research that investigate the roles of context, attention, learning, and memory in general auditory processing (Kidd, Richards, Streeter, Mason, & Huang, 2011; Krishnan, Leech, Aydelott, & Dick, 2013; Ortiz & Wright, 2010; Snyder & Weintraub, 2013).

16.3 EFFECTS OF AUDITORY INTERACTION ON THE FORM OF SPEECH

The patterns of acoustic change that convey information in speech are notoriously complex. Speech sounds like /d/ and /g/ are not conveyed by a necessary or sufficient acoustic cue and there is no canonical acoustic template that definitively signals a linguistic message. Furthermore, variability is the norm. The detailed acoustic signature of a particular phoneme, syllable, or word varies a great deal across different contexts, utterances, and talkers. The inherent multidimensionality of the acoustic signatures that convey speech sounds and the variability along these dimensions presents a challenge for understanding how listeners readily map the continuous signal to discrete linguistic representations. This has been the central issue of speech perception research. Although some researchers have suggested that acoustic variability may serve useful functions in speech communication (Elman & McClelland, 1986; Liberman, 1996), the prevailing approach has been to explore how listeners accommodate or compensate for the messy physical acoustic signal to align it with native-language linguistic knowledge.

Although this framing of speech perception has dominated empirical research and theory, the focus on acoustic variability may lead us to pursue answers to the wrong questions. Like all perceptual systems, the auditory system transforms sensory input; it is not a linear system. It is possible that the nature of auditory perceptual transformations is such that the challenge of *acoustic* variability is mitigated when analyzed through the lens of *auditory* perception. Some of the more daunting mysteries about the ability of humans to accommodate acoustic variability in speech may arise from a lack of understanding of how the auditory system encodes complex sounds, generally.

Coarticulation is a case in point. As we talk, the mouth, jaw, and other articulators move very quickly, but not instantaneously, from target to target. Consequently, at any point in time the movement of the articulators is a function of the articulatory demands of previous and subsequent phonetic sequences as well as the "current" intended production. As a direct result, the acoustic signature of a speech sound is context-dependent. When /al/ precedes /ga/, for example, the tongue must quickly move from anterior to posterior occlusions to form the consonants. The effect of coarticulation is to draw /ga/ to a more anterior position (toward /al/). This context-sensitive shift in production impacts the resultant acoustic realization, making it more "da"-like because the place of tongue occlusion slides forward in the mouth toward the articulation typical of "da." Likewise, when /da/ is spoken after the more posteriorly articulated /ar/, the opposite pattern occurs; the acoustics of /da/ become more "ga"-like. This means that, due to coarticulation, the acoustic signature of the second syllables in "alga" and "arda" can be highly similar (Mann, 1980).

Viewed from the perspective of acoustic variability, this issue seems intractable. If the second consonant of "alga" and "arda" is signaled by highly similar acoustics, then how is it that we hear the distinct syllables "ga" and "da"? The answer lies in the incredible context dependence of speech perception; perception appears to compensate for coarticulation. This can be demonstrated by preceding a perceptually ambiguous syllable between /ga/ and /da/ with /al/ or /ar/. Whereas the acoustics of /ga/ produced after /al/ are more "da"-like, a preceding /al/ shifts perception of the ambiguous sound toward "ga." Similarly, /ar/ shifts perception of the same ambiguous sound toward "da." This pattern opposes the coarticulatory effects in speech production. In this example and many replications with other tasks and stimuli, coarticulation assimilates speech acoustics, but perception "compensates" in the opposing direction (Mann, 1980; Mann & Repp, 1980).

The traditional interpretation of these findings highlights that theoretical approaches have tended to discount what the auditory system can contribute to the challenges of speech perception. The flexibility of speech perception to make use of so many acoustic dimensions to signal a particular speech sound and the dependence of this mapping on context has

suggested to many that it is infeasible for these effects to arise from auditory processing. This challenge is part of what led to the proposal that motor representations might be better suited to serve as the basis of speech communication. But, by virtue of being sound, acoustic speech necessarily interfaces with early auditory perceptual operations. As noted, these operations are not linear; they do not simply convey raw acoustic input, they transform it. Thus, although acoustics are readily observable and provide a straightforward means of estimating input to the linguistic system, this representation is not equivalent to the *auditory* information available to the linguistic system. What might be gained by considering *auditory*—rather than *acoustic*—information?

Lotto and Kluender (1998) approached this question by examining whether perceptual compensation for coarticulation like that described for "alga" and "arda" really requires information about speech articulation, or whether the context sounds need to be speech at all. They did this by creating nonspeech sounds that had some of the acoustic energy that distinguishes /al/ from /ar/. These nonspeech signals do not carry information about articulation, talker identity, or any other speech-specific details. The result was two nonspeech tone sweeps, one with energy like /al/ and the other with energy mimicking /ar/. Lotto and Kluender found that when these nonspeech acoustic signals preceded /ga/ and /da/ sounds, the tone sweeps had the same influence as the /al/ and /ar/ sounds they modeled. So-called perceptual compensation for coarticulation is observed even for nonspeech contexts that convey no information about speech articulation.

This finding has been directly replicated (Fowler, 2006; Lotto, Sullivan, & Holt, 2003) and extended to other stimulus contexts (Coady, Kluender, & Rhode, 2003; Fowler, Brown, & Mann, 2000; Holt, 1999; Holt & Lotto, 2002) many times. Across these replications, the pattern of results reveals that a basic characteristic of auditory perception is to exaggerate contrast. Preceded by a high-frequency sound (whether speech or nonspeech), subsequent sounds are perceived to be lower-frequency. This is also true in the temporal domain; preceded by longer sounds or sounds presented at a slower rate, subsequent sounds are heard as shorter (Diehl & Walsh, 1989; Wade & Holt, 2005a, 2005b). Further emphasizing the generality of these effects, Japanese quail exhibit the pattern of speech context dependence that had been thought to be indicative of perceptual compensation for coarticulation (Lotto, Kluender, & Holt, 1997).

This example underscores the fact that *acoustic* and *auditory* are not one and the same. Whereas there is considerable variability in speech acoustics, some of this variability is accommodated by auditory perceptual processing. In this way, the form of speech can have coarticulation and still be an effective communication signal because the operating characteristics of the auditory system include exaggeration of spectral and temporal contrast. Lotto et al. (1997) argue that the symmetry of assimilated speech production and contrastive perception is not serendipitous, but rather is a consequence of organisms having evolved within natural environments in which sound sources are physically constrained in the sounds they can produce. Because of mass and inertia, natural sound sources tend to be assimilative, like speech articulators. Perceptual systems, audition included, tend to emphasize signs of change, perhaps because in comparison with physical systems' relative sluggishness rapid change is ecologically significant information. Having evolved like other perceptual systems to respect regularities of the natural environment, auditory processing transforms coarticulated acoustic signals to exaggerate contrast and, thus, eliminates some of the apparent challenges of coarticulation. We can communicate efficiently as our relatively sluggish articulators perform acrobatics across tens of milliseconds to produce speech, in part because our auditory system evolved to use acoustic signals from natural sound sources that face the same physical constraints.

These results also highlight the importance of considering higher-level auditory processing in constraining models of speech perception. Subsequent research has shown that the auditory system exhibits spectral and temporal contrast for more complex sound input (Holt, 2005, 2006a, 2006b; Laing, Liu, Lotto, & Holt, 2012). These studies indicate that the auditory system tracks the long-term average spectra (or rate; Wade & Holt, 2005a) of sounds, and that subsequent perception is relative to, and contrastive with, these distributional characteristics of preceding acoustic signals (Watkins, 1991; Watkins & Makin, 1994). These effects, described graphically in Figure 16.1, cannot be explained by low-level peripheral auditory processing; effects persist over more than a second of silence or intervening sound (Holt, 2005) and require the system to track distributional regularity across acoustic events (Holt, 2006a). These findings are significant for understanding talker and rate normalization, which refer to the challenges introduced to speech perception by acoustic variability arising from different speakers and different rates of speech. What is important is that the preceding context of sounds possess acoustic energy in the spectral (Laing et al., 2012) or temporal (Wade & Holt, 2005a, 2005b) region distinguishing the target phonemes, and not that the context carries articulatory or speech-specific information. Here, too, some of the challenges apparent from speech acoustics may be resolved in the transformation from acoustic to auditory.

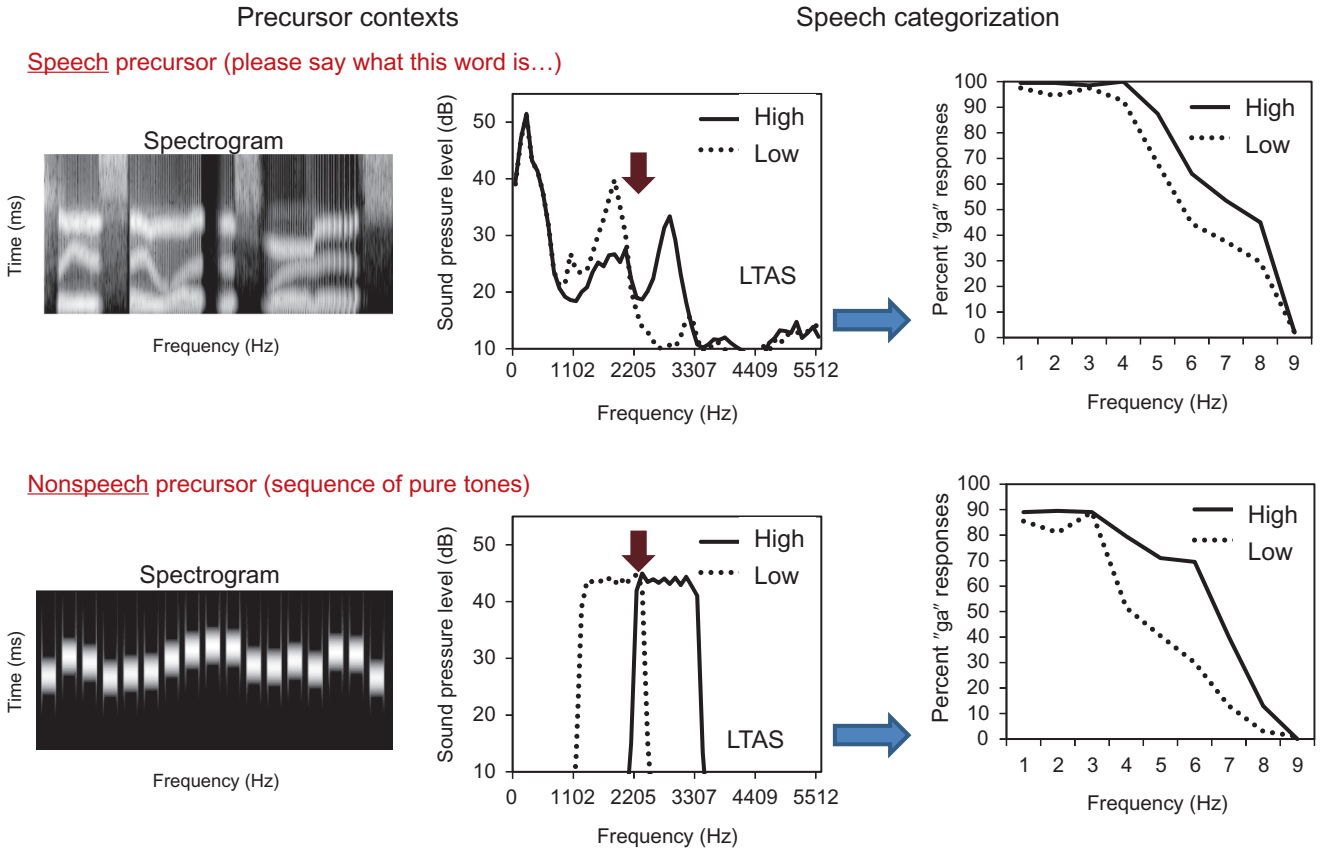

FIGURE 16.1 Precursor contexts and their effect on adult /ga/-/da/ categorization. Manipulation of the Long-Term Average Spectrum (LTAS) of both speech (top) and nonspeech (bottom) has a strong, contrastive influence on speech categorization. *From Laing, Liu, Lotto, and Holt (2012) with permission from the publishers.*

16.4 EFFECTS OF LEARNABILITY ON THE FORM OF SPEECH

Auditory representations are influenced greatly by both short-term and long-term experience. Categorical perception, the classic textbook example among speech perception phenomena, exemplifies this. When native-language speech varying gradually in its acoustics is presented to listeners, the patterns of identification change abruptly, not gradually, from one phoneme (or syllable or word) to another. Likewise, there is a corresponding discontinuity in discrimination such that pairs of speech sounds are more discriminable if they lie on opposite sides of the sharp identification boundary than if they lie on the same side of the identification curve's slope, even when they are matched in acoustic difference. Said another way, acoustically distinct speech sounds identified with the same label are difficult to discriminate, whereas those with different labels are readily discriminated. Despite the renown of categorical perception for speech, it is now understood that it is not specific to speech (Beale & Keil, 1995; Bimler & Kirkland, 2001; Krumhansl, 1991; Livingston, Andrews, & Harnad, 1998; Mirman, Holt, & McClelland, 2004), and that even speech is not entirely "categorical" (Eimas, 1963; Harnad, 1990; Liberman, Harris, Hoffman, & Griffith, 1957; Pisoni, 1973). Infants (Kuhl, 1991; McMurray & Aslin, 2005) and adults (Kluender, Lotto, Holt, & Bloedel, 1998; McMurray, Aslin, Tanenhaus, Spivey, & Subik, 2008) remain sensitive to within-category acoustic variation. Speech categories exhibit graded internal structure such that instances of a speech sound are treated as relatively better or worse exemplars of the category (Iverson & Kuhl, 1995; Iverson et al., 2003; Johnson, Flemming, & Wright, 1993; Miller & Volaitis, 1989).

We have argued that it may be more productive to consider speech perception as *categorization*, as opposed to *categorical* (Holt & Lotto, 2010). This may seem like a small difference in designation, but it has important consequences. Considering speech perception as an example of general auditory categorization provides a means of understanding how the system comes to exhibit relative perceptual constancy in the face of acoustic variability and does so in a native-language–specific manner. The reason for this is that although

there is a great deal of variability in speech acoustics, there also exist underlying regularities in the distributions of experienced native-language speech sounds. This is the computational challenge of categorization; discriminably different exemplars come to be treated as functionally equivalent. A system that can generalize across variability to discover underlying patterns and distributional regularities—a system that can *categorize*—may cope with the acoustic variability inherent in speech without need for invariance. Seeking invariance in the acoustic signatures of speech becomes less essential if we take a broader view that extends beyond pattern matching to consider active auditory processing that involves higher-order and multimodal perception, categorization, attention, and learning.

From this perspective, learning about how listeners acquire auditory categories can constrain behavioral and neurobiological models of speech perception. Whereas the acquisition of first and second language phonetic systems provides an opportunity to observe the development of complex auditory categories, our ability to model these categorization processes is limited because it is extremely difficult to control or even accurately measure a listener's history of experience with speech sounds. However, we are beginning to develop insights into auditory categorization from experiments using novel artificial nonspeech sound categories that inform our understanding about how speech perception and acquisition are constrained by general perceptual learning mechanisms (Desai, Liebenthal, Waldron, & Binder, 2008; Guenther, Husain, Cohen, & Shinn-Cunningham, 1999; Holt & Lotto, 2006; Holt, Lotto, & Diehl, 2004; Ley et al., 2012; Liebenthal et al., 2010).

One example of what this approach can reveal about how auditory learning constrains speech relates to a classic early example of the "lack of invariance" in speech acoustics. If one examines the formant frequencies corresponding most closely with /d/ as it precedes different vowels, then it is impossible to define a single acoustic dimension that uniquely distinguishes the sound as a /d/; the acoustics are greatly influenced by the following vowel (Liberman, Delattre, Cooper, & Gerstman, 1954). This kind of demonstration fueled theoretical commitments that speech perception is accomplished via the speech motor system in the hopes that this would provide a more invariant mapping than acoustics (Liberman et al., 1967). Viewed from the perspective of acoustics, perceptual constancy for /d/ seemed an intractable problem for auditory processing.

Wade and Holt (2005a, 2005b) modeled this perceptual challenge with acoustically complex nonspeech sound exemplars that formed categories signaled only by higher-order acoustic structure and not by any invariant acoustic cue (see Figure 16.2 for a representation of the stimulus set). Naïve participants experienced these sounds in the context of a videogame in which learning sound categories facilitated advancement in the game but was never explicitly required or rewarded. Within just a half-hour of game play, participants categorized the sounds and generalized their category learning to novel exemplars. This learning led to an exaggeration of between-category discriminability (of the sort traditionally attributed to categorical perception) as measured with electroencephalography (EEG; Liu & Holt, 2011). The seemingly intractable lack of acoustic invariance is, in fact, readily learnable even in an incidental task.

This is proof that the auditory system readily uses multimodal environmental information (modeled in the videogame as sound-object links, as in natural environments) to facilitate discovery of the distributional regularities that define the relations between category exemplars while generalizing across acoustic variability within categories. More than this, however, the approach can reveal details of auditory processing that constrain behavioral and neurobiological models of speech perception. Using the same nonspeech categories and training paradigm, Leech, Holt, Devlin, and Dick (2009) discovered that the extent to which participants learn to categorize nonspeech sounds is strongly correlated with the pretraining to post-training recruitment of left posterior temporal sulcus (pSTS) during presentation of the nonspeech sound category exemplars. This is unexpected because left pSTS has been described as selective for specific acoustic and informational properties of speech signals (Price, Thierry, & Griffiths, 2005). In recent work, Lim, Holt, and Fiez (2013) have found that left pSTS is recruited online in the videogame category training task in a manner that correlates with behavioral measures of learning. These results also demonstrate that recruitment of left pSTS by the nonspeech sound categories cannot be attributed to their superficial acoustic signal similarity to speech or to mere exposure. When highly similar nonspeech sounds are sampled such that category membership is random instead of structured, left pSTS activation is not related to behavioral performance.

As in the examples from the preceding sections, this series of studies demonstrates that there is danger in presuming that speech is fundamentally different from other sounds in either its acoustic structure or in the basic perceptual processes it requires. The selectivity of left pSTS for speech should not be understood to be selectivity for intrinsic properties of acoustic speech signals, such as the articulatory information that speech may carry. Instead, this region seems to meet the computational demands presented by learning to treat structured distributions of acoustically variable sounds as functionally equivalent.

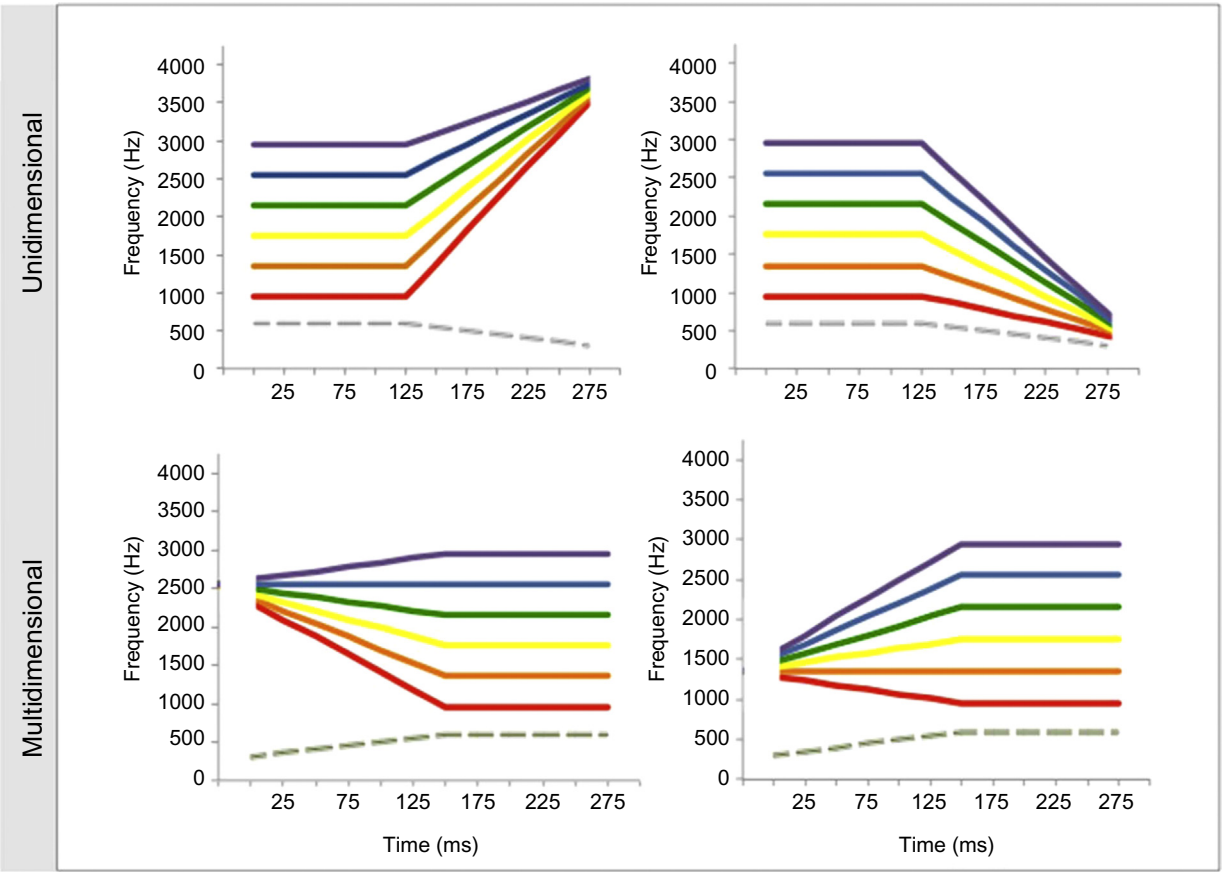

FIGURE 16.2 Schematic spectrograms showing the artificial nonspeech auditory categories across time and frequency. The dashed gray lines show the lower-frequency spectral peak, P1. The colored lines show the higher-frequency spectral peak, P2. The six exemplars of each category are composed of P1 and one of the colored P2 components pictured. Note that unidimensional categories are characterized by an offset glide that increases (top left) or decreases (top right) in frequency across all exemplars. No such unidimensional cue differentiates the multidimensional categories. *From Wade and Holt (2005a, 2005b) with permission from the publishers.*

Likewise, caution is warranted in presuming that the transformation from acoustic to auditory involves only a static mapping to stable, unchanging linguistic representations. The recruitment of putatively speech-selective left pSTS was driven by category learning in less than an hour (Lim et al., 2013). Thus, the behavioral relevance of the artificial, novel auditory categories drove reorganization of their transformations from acoustic to auditory. The examples we present here illustrate the facile manner by which auditory categories can be acquired. On an even shorter time scale, there is considerable evidence that the mapping of speech acoustics to linguistic representation is "tuned" by multiple information sources in an adaptive manner such as may be required to adapt to foreign accented speech or to speech in adverse, noisy environments (Kraljic, Brennan, & Samuel, 2008; Mehler et al., 1993; Vitela, Carbonell, & Lotto, 2012). The active, flexible nature of auditory processing puts learning in the spotlight and positions questions of speech perception in greater contact with other neurobiological approaches to understanding perception, cognition, and language.

16.5 MOVING FORWARD

The preceding sections provide a few brief examples of how general auditory processing may influence the perception of speech sounds as well as the structure of phonetic systems. These examples demonstrate that, at the very least, human speech communication appears to take advantage of the things that the auditory system does well—phonetic inventories tend to include sounds whose differences are well-encoded in the auditory system. The acoustic effects of coarticulation are just the types of interactions that the auditory system can accommodate, and the multidimensional structure of speech sounds form just the kinds of categories that are easily learned by the auditory system. Whether there are additional specialized processes required for speech perception, it is likely that the

auditory system constrains the way we talk to and perceive each other to a greater extent than has been acknowledged.

One of the beneficial outcomes of the fact that the auditory system plays a strong role in speech perception is that there is the opportunity for synergy between research of speech and of general auditory processing. Speech perception phenomena shine a light on auditory processes that have remained unilluminated by research of simpler acoustic stimuli. The theories regarding the auditory distinctiveness of speech sounds have inspired the search for better models of auditory encoding of complex stimuli and better functions for computing distinctiveness (Lotto et al., 2003). The existence of perceptual compensation for coarticulation and talker normalization provide evidence for spectral and temporal interactions in general auditory processing that are not evident when presenting stimuli in isolation (Holt, 2006a, 2006b; Holt & Lotto, 2002; Watkins & Makin, 1994). The complexity of speech categories along with the ease with which humans learn them is the starting point for most of the current work on auditory categorization (Goudbeek, Smits, Swingly, & Cutler, 2005; Lotto, 2000; Maddox, Molis, & Diehl, 2002; Smits, Sereno, & Jongman, 2006; Wade & Holt, 2005a, 2005b).

The vitality of auditory and speech cognitive neuroscience depends on continuing this trend of using speech and auditory phenomena to mutually inform and inspire each field.

References

Beale, J. M., & Keil, F. C. (1995). Categorical effects in the perception of faces. *Cognition*, *57*(3), 217–239.

Bimler, D., & Kirkland, J. (2001). Categorical perception of facial expressions of emotion: Evidence from multidimensional scaling. *Cognition and Emotion*, *15*(5), 633–658.

Bradlow, A., & Bent, T. (2002). The clear speech effect for non-native listeners. *Journal of the Acoustical Society of America*, *112*(1), 272–284.

Coady, J. A., Kluender, K. R., & Rhode, W. S. (2003). Effects of contrast between onsets of speech and other complex spectra. *Journal of the Acoustical Society of America*, *114*, 2225.

Desai, R., Liebenthal, E., Waldron, E., & Binder, J. R. (2008). Left posterior temporal regions are sensitive to auditory categorization. *Journal of Cognitive Neuroscience*, *20*(7), 1174–1188.

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

Diehl, R. L., & Kluender, K. R. (1987). On the categorization of speech sounds. In S. Harnad (Ed.), *Categorical perception: The groundwork of cognition* (pp. 226–253). London: Cambridge University Press.

Diehl, R. L., & Kluender, K. R. (1989). On the objects of speech perception. *Ecological Psychology*, *1*(2), 121–144.

Diehl, R. L., Kluender, K. R., & Walsh, M. A. (1990). Some auditory bases of speech perception and production. *Advances in Speech, Hearing and Language Processing*, *1*, 243–268.

Diehl, R. L., Kluender, K. R., Walsh, M. A., & Parker, E. M. (1991). Auditory enhancement in speech perception and phonology. In R. Hoffman, & D. Palermo (Eds.), *Cognition and the symbolic process: Analytical and ecological perspectives* (pp. 59–75). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Diehl, R. L., Lindblom, B., & Creeger, C. P. (2003). *Increasing realism of auditory representations yields further insights into vowel phonetics, Proceedings of the fifteenth international congress of phonetic sciences* (Vol. 2, pp. 1381–1384). Adelaide: Causal Publications.

Diehl, R. L., & Walsh, M. A. (1989). An auditory basis for the stimulus-length effect in the perception of stops and glides. *Journal of the Acoustical Society of America*, *85*(5), 2154–2164.

Eimas, P. D. (1963). The relationship between identification and discrimination along speech and non-speech continua. *Language and Speech*, *6*(4), 206–217.

Elman, J. L., & McClelland, J. L. (1986). Exploiting lawful variability in the speech wave. In J. S. Perkell, & D. H. Klatt (Eds.), *Invariance and variability of speech processes* (pp. 360–385). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Fowler, C. A. (2006). Compensation for coarticulation reflects gesture perception, not spectral contrast. *Perception and Psychophysics*, *68* (2), 161–177.

Fowler, C. A. (2008). The FLMP STMPed. *Psychonomic Bulletin and Review*, *15*(2), 458–462.

Fowler, C. A., Brown, J. M., & Mann, V. A. (2000). Contrast effects do not underlie effects of perceding liquids on stop-consonant identification by humans. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(3), 877–888.

Froehlich, P., Collet, L., Chanal, J.-M., & Morgon, A. (1990). Variability of the influence of a visual task on the active micromechanical properties of the cochlea. *Brain Research*, *508*(2), 286–288.

Garinis, A. C., Glattke, T., & Cone, B. K. (2011). The MOC reflex during active listening to speech. *Journal of Speech, Language, and Hearing Research*, *54*(5), 1464–1476.

Giard, M.-H., Collet, L., Bouchet, P., & Pernier, J. (1994). Auditory selective attention in the human cochlea. *Brain Research*, *633*(1), 353–356.

Goudbeek, M., Smits, R., Cutler, A., & Swingley, D. (2005). Acquiring auditory and phonetic categories. In H. Cohen, & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science* (pp. 497–513). Amsterdam: Elsevier.

Green, D. M. (1988). *Profile analysis: Auditory intensity discrimination*. New York: Oxford University Press.

Green, D. M., Mason, C. R., & Kidd, G. (1984). Profile analysis: Critical bands and duration. *The Journal of the Acoustical Society of America*, *75*(4), 1163–1167.

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*(5), 2900–2912.

Guenther, F. H., & Vladusich, T. (2012). A neural theory of speech acquisition and production. *Journal of Neurolinguistics*, *25*(5), 408–422.

Harnad, S. R. (1990). *Categorical perception: The groundwork of cognition*. New York: Cambridge University Press.

Hickok, G. S., Houde, J., & Rong, F. (2011). Sensorimotor integration in speech processing: Computational basis and neural organization. *Neuron*, *69*(3), 407–422.

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

Holt, L., & Lotto, A. (2010). Speech perception as categorization. *Attention, Perception, and Psychophysics*, *72*(5), 1218–1227.

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

Holt, L., & Lotto, A. J. (2002). Behavioral examinations of the level of auditory processing of speech context effects. *Hearing Research*, *167*, 156–169.

Holt, L. L. (1999). *Auditory constraints on speech perception: An examination of spectral contrast*. Ph.D. thesis, University of Wisconsin-Madison.

Holt, L. L. (2005). Temporally nonadjacent nonlinguistic sounds affect speech categorization. *Psychological Science*, *16*(4), 305–312.

Holt, L. L. (2006a). The mean matters: Effects of statistically defined nonspeech spectral distributions on speech categorization. *Journal of the Acoustical Society of America*, *120*, 2801–2817.

Holt, L. L. (2006b). Speech categorization in context: Joint effects of nonspeech and speech precursors. *Journal of the Acoustical Society of America*, *119*(6), 4016–4026.

Holt, L. L., & Lotto, A. J. (2008). Speech perception within an auditory cognitive neuroscience framework. *Current Directions in Psychological Science*, *17*(1), 42–46.

Iverson, P., Kuhl, P., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., et al. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition*, *87*(1), B47–B57.

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*(1), 553–562.

Jesteadt, W., Wier, C. C., & Green, D. M. (1977). Intensity discrimination as a function of frequency and sensation level. *Journal of the Acoustical Society of America*, *61*(1), 169–177.

Johnson, K., Flemming, E., & Wright, R. (1993). The hyperspace effect: Phonetic targets are hyperarticulated. *Language*, *69*(3), 505–528.

Kidd, G. J., Richards, V. M., Streeter, T., Mason, C. R., & Huang, R. (2011). Contextual effects in the identification of nonspeech auditory patterns. *Journal of the Acoustical Society of America*, *130*(6), 3926–3938.

Kingston, J., & Diehl, R. L. (1995). Intermediate properties in the perception of distinctive feature values. *Papers in Laboratory Phonology*, *4*, 7–27.

Kluender, K. R., Diehl, R. L., & Wright, B. A. (1988). Vowel length differences before voiced and voiceless consonants: An auditory explanation. *Journal of Phonetics*, *16*, 153–169.

Kluender, K. R., & Kiefte, M. (2006). Speech perception within a biologically realistic information-theoretic framework. In M. A. Gernsbacher, & M. Traxler (Eds.), *Handbook of psycholinguistics* (2nd ed., pp. 153–199). London: Elsevier.

Kluender, K. R., Lotto, A. J., Holt, L. L., & Bloedel, S. L. (1998). Role of experience for language-specific functional mappings of vowel sounds. *Journal of the Acoustical Society of America*, *104*(6), 3568–3582.

Kraljic, T., Brennan, S. E., & Samuel, A. G. (2008). Accommodating variation: Dialects, idiolects, and speech processing. *Cognition*, *107*(2), 54–81.

Krishnan, S., Leech, R., Aydelott, J., & Dick, F. (2013). School-age children's environmental object identification in natural auditory scenes: Effects of masking and contextual congruence. *Hearing Research*, *300*, 46–55.

Krumhansl, C. L. (1991). Music psychology: Tonal structures in perception and memory. *Annual Review of Psychology*, *42*(1), 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 and Psychophysics*, *50*(2), 93–107.

Laing, E. J. C., Liu, R., Lotto, A. J., & Holt, L. L. (2012). Tuned with a tune: Talker normalization via general auditory processes. *Frontiers in Psychology*, *3*, 203–227.

Lane, H. (1965). The motor theory of speech perception: A critical review. *Psychological Review*, *72*(4), 275–309.

Leech, R., Holt, L. L., Devlin, J. T., & Dick, F. (2009). Expertise with artificial nonspeech sounds recruits speech-sensitive cortical regions. *The Journal of Neuroscience*, *29*(16), 5234–5239.

Ley, A., Vroomen, J., Hausfeld, L., Valente, G., de Weerd, P., & Formisano, E. (2012). Learning of new sound categories shapes neural response patterns in human auditory cortex. *The Journal of Neuroscience*, *32*(38), 13273–13280.

Liberman, A., Harris, K., Hoffman, H., & Griffith, B. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, *54*, 358–368.

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

Liberman, A. M., Cooper, F. S., Harris, K. S., MacNeilage, P. F., & Studdert-Kennedy, M. (1964). *Some observations on a model for speech perception. Proceedings of the AFCRL symposium on models for the perception of speech and visual form*. Cambridge: MIT Press.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431–461.

Liberman, A. M., Delattre, P. C., Cooper, F. S., & Gerstman, L. J. (1954). The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs: General and Applied*, *68*(8), 1–13.

Liebenthal, E., Desai, R., Ellingson, M. M., Ramachandran, B., Desai, A., & Binder, J. R. (2010). Specialization along the left superior temporal sulcus for auditory categorization. *Cerebral Cortex*, *20*(12), 2958–2970.

Liljencrants, J., & Lindblom, B. (1972). Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language*, *48*(4), 839–862.

Lim, S., Holt, L.L., & Fiez, J.A. (2013). Context-dependent modulation of striatal systems during incidental auditory category learning. *Poster presentation at the 43rd Annual Conference of the Society for Neuroscience*. San Diego, CA.

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

Lindblom, B. (1991). The status of phonetic gestures. In I. G. Mattingly, & M. Studdert-Kennedy (Eds.), *Modularity and the motor theory of speech perception* (pp. 7–24). Hillsdale, NJ: Lawrence Erlbaum Associates.

Liu, R., & Holt, L. L. (2011). Neural changes associated with nonspeech auditory category learning parallel those of speech category acquisition. *Journal of Cognitive Neuroscience*, *23*(3), 683–698.

Livingston, K. R., Andrews, J. K., & Harnad, S. (1998). Categorical perception effects induced by category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(3), 732–753.

Lotto, A. J. (2000). Language acquisition as complex category formation. *Phonetica*, *57*, 189–196.

Lotto, A. J., Hickok, G. S., & Holt, L. L. (2009). Reflections on mirror neurons and speech perception. *Trends in Cognitive Sciences*, *13*(3), 110–114.

Lotto, A. J., & Kluender, K. R. (1998). General contrast effects in speech perception: Effect of preceding liquid on stop consonant identification. *Perception and Psychophysics*, *60*(4), 602–619.

Lotto, A. J., Kluender, K. R., & Holt, L. L. (1997). Perceptual compensation for coarticulation by Japanese quail (*Coturnix coturnix japonica*). *Journal of the Acoustical Society of America*, *102*, 1134–1140.

Lotto, A. J., Sullivan, S. C., & Holt, L. L. (2003). Central locus for nonspeech context effects on phonetic identification. *Journal of the Acoustical Society of America*, *113*(1), 53–56.

Maddox, W. T., Molis, M. R., & Diehl, R. L. (2002). Generalizing a neuropsychological model of visual categorization to auditory categorization of vowels. *Perception and Psychophysics*, *64*(4), 584–597.

Maison, S., Micheyl, C., & Collet, L. (2001). Influence of focused auditory attention on cochlear activity in humans. *Psychophysiology*, *38* (1), 35–40.

Mann, V., & Repp, B. (1980). Influence of vocalic context on perception of the [s]-[S] distinction. *Attention, Perception, and Psychophysics*, *28*(3), 213–228.

Mann, V. A. (1980). Influence of preceding liquid on stop-consonant perception. *Perception and Psychophysics*, *28*(5), 407–412.

Massaro, D. W., & Chen, T. H. (2008). The motor theory of speech perception revisited. *Psychonomic Bulletin and Review*, *15*(2), 453–457.

McMurray, B., & Aslin, R. N. (2005). Infants are sensitive to within-category variation in speech perception. *Cognition*, *95*(2), B15–B26.

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 and Performance*, *34*(6), 1609–1631.

Mehler, J., Sebastian, N., Altmann, G., Dupoux, E., Christophe, A., & Pallier, C. (1993). Understanding compressed sentences: The role of rhythm and meaning. *Annals of the New York Academy of Sciences*, *682*(1), 272–282.

Miller, J. L., & Volaitis, L. E. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception and Psychophysics*, *46*(6), 505–512.

Mirman, D., Holt, L., & McClelland, J. (2004). Categorization and discrimination of nonspeech sounds: Differences between steady-state and rapidly-changing acoustic cues. *Journal of the Acoustical Society of America*, *116*(2), 1198–1207.

Ohala, J. J. (1993). Sound change as nature's speech perception experiment. *Speech Communication*, *13*(1–2), 155–161.

Ortiz, J. A., & Wright, B. A. (2010). Differential rates of consolidation of conceptual and stimulus learning following training on an auditory skill. *Experimental Brain Research*, *201*(3), 441–451.

Picheny, M., Durlach, N., & Braida, L. (1985). Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. *Journal of Speech, Language, and Hearing Research*, *28*, 96–103.

Picheny, M. A., Durlach, N. I., & Braida, L. (1986). Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. *Journal of Speech, Language, and Hearing Research*, *29*, 434–446.

Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception and Psychophysics*, *13*, 253–260.

Price, C., Thierry, G., & Griffiths, T. (2005). Speech-specific auditory processing: Where is it? *Trends in Cognitive Sciences*, *9*(6), 271–276.

Smits, R., Sereno, J., & Jongman, A. (2006). Categorization of sounds. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(3), 733–754.

Snyder, J. S., & Weintraub, D. M. (2013). Loss and persistence of implicit memory for sound: Evidence from auditory stream segregation context effects. *Attention, Perception, and Psychophysics*, *75*, 1056–1074.

Song, J. H., Skoe, E., Wong, P., & Kraus, N. (2008). Plasticity in the adult human auditory brainstem following short-term linguistic training. *Journal of Cognitive Neuroscience*, *20*(10), 1892–1902.

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

Stevens, K. N. (1989). On the quantal nature of speech. *Journal of Phonetics*, *17*, 3–45.

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*(3), 234–249.

Trout, J. D. (2001). The biological basis of speech: What to infer from talking to the animals. *Psychological Review*, *108*(3), 523–549.

Vitela, A. D., Carbonell, K. M., & Lotto, A. J. (2012). Predicting the effects of carrier phrases in speech perception. *Poster presentation at the 53rd meeting of the Psychonomics Society*. Minneapolis, MN.

Wade, T., & Holt, L. (2005a). Incidental categorization of spectrally complex non-invariant auditory stimuli in a computer game task. *Journal of the Acoustical Society of America*, *118*(4), 2618–2633.

Wade, T., & Holt, L. L. (2005b). Perceptual effects of preceding nonspeech rate on temporal properties of speech categories. *Perception and Psychophysics*, *67*(6), 939–950.

Watkins, A. J. (1991). Central, auditory mechanisms of perceptual compensation for spectral-envelope distortion. *Journal of the Acoustical Society of America*, *90*(6), 2942–2955.

Watkins, A. J., & Makin, S. J. (1994). Perceptual compensation for speaker differences and for spectral-envelope distortion. *Journal of the Acoustical Society of America*, *96*(3), 1263–1282.

Wier, C. C., Jesteadt, W., & Green, D. M. (1977). Frequency discrimination as a function of frequency and sensation level. *Journal of the Acoustical Society of America*, *61*(1), 178–184.

Wong, P., Skoe, E., Russo, N. M., Dees, T., & Kraus, N. (2007). Musical experience shapes human brainstem encoding of linguistic pitch patterns. *Nature Neuroscience*, *10*(4), 420–422.