CHAPTER THREE

# Understanding Sound: Auditory Skill Acquisition

**Shannon L.M. Heald[1], Stephen C. Van Hedger and Howard C. Nusbaum**

The University of Chicago, Chicago, IL, United States
[1]Corresponding author: E-mail: smbowdre@uchicago.edu

## Contents

1. Introduction 54
2. Acoustic Challenges to Perception 57
3. Category Knowledge 59
4. Perceptual Learning 61
   4.1 Evidence of Plasticity 62
   4.2 Perceptual Stability and Attention 64
   4.3 Absolute Pitch as a Skill 67
   4.4 Training Absolute Pitch in Adults 69
5. Generalization and Transfer 72
   5.1 Recognizing Novel Acoustic Patterns 73
   5.2 Sleep and Perceptual Generalization 74
6. Context and Cognition 78
   6.1 Beyond Statistical Learning 81
7. Conclusions 82
Acknowledgment 84
References 84

## Abstract

Research on auditory perception often starts from an assumed theoretical framework of bottom-up acoustic signal decoding followed by pattern matching of signal information to memory. Some specific forms of auditory perception such as speech perception are often assumed to be mediated by specialized mechanisms, shaped by evolution to address the challenges of speech perception. However, neither of these broad approaches conceives of these systems as intrinsically adaptable and plastic; learning is typically considered as a separate process. This chapter addresses the question of auditory perceptual understanding and plasticity and argues that auditory perception, including speech perception, is an active cognitive process that incorporates learning. In contrast to a passive process in which patterns are mapped into memory in a direct comparative process, an active process forms hypotheses about pattern identity, and

*Psychology of Learning and Motivation*, Volume 67
ISSN 0079-7421
http://dx.doi.org/10.1016/bs.plm.2017.03.003

© 2017 Elsevier Inc.
All rights reserved.

tests these hypotheses to adaptively shift attention to aspects of signal information. The process of hypothesis testing for signal understanding comes into play in cases of signal ambiguity or uncertainty, and involves auditory working memory and the control of attention. This conceptualization of adaptive processing incorporates context-sensitive change into the perceptual system. We argue that perception should be viewed not as a fixed instrument of information reception, but as a dynamic, adaptive system that is changed by the act of perception. We present evidence for these claims, outline a framework for a theory of active auditory perception, and argue further against claims of critical periods as biological determinism for perceptual plasticity given that auditory perception is auditory learning.

## 1. INTRODUCTION

Expert radiologists can look at an X-ray and see vague gray scale shapes as three-dimensional organs and tumors (Lesgold et al., 1988). Chess experts look at a chess game in many different ways than chess novices (Chase & Simon, 1973; Krawczyk, Boggan, McClelland, & Bartlett, 2011). In both cases, deep expertise shapes visual perception, and seeing the world through expert eyes can be thought of as a perceptual skill. But while fingerprint-classifying experts and satellite photo interpreters can be said to have a visual skill, what is an auditory skill? One auditory ability that all hearing people share is fluent speech perception. We can recognize speech in spite of distortion and noise, at various rates of speech (and varying rates of talking), across differences in speakers, and even in speakers with accents or those eating hamburgers. Moreover as speech listeners, we can understand what is being said, something about where the speaker came from (accent/dialect), and something about how the speaker thinks and feels about what is being said. We understand this quickly and effectively and more effectively than any current speech recognition system. But is speech perception an ability or a skill?

The distinction between skill and ability in colloquial terms is usually focused on skills being learned but abilities being part of our individual nature or biological endowment (whatever that might mean). An ability might be developed, with age, or honed with experience, but it is presumably rooted in some kind of unlearned natively human biological endowment. An example might be athleticism as in, "she is a natural athlete!" On the other hand, a skill is something that is learned, that has to be practiced, possibly instructed, and not something we think of as part of a natural endowment. An example of this might be found in the domain

of medicine, when one says, "she is an incredible surgeon!" Of course both skills and abilities make use of certain endowments such as precision of fine motor control or perceptual acuity. And certainly both can be improved by practice and by instruction. But this raises a fundamental question that is sometimes controversial: in any particular skill or ability, there is both a starting place in mental and physical terms (usually already shaped experientially), and there is a developmental trajectory that is shaped by experiences of various kinds. This becomes controversial when claims are made that the starting point and trajectory are specifically biologically determined or entirely dependent on experience, assuming that these unspecified broad factors are really different or independent, which seems empirically questionable.

The typical human biological endowment anatomically includes hands, eyes, and ears, and a brain along with many other human parts. But some researchers would consider spoken language ability to be part of that endowment (e.g., Anderson & Lightfoot, 2002; Chomsky, 1975), given the pervasiveness of spoken language in humans (e.g., Lennenberg, 1967). Just as people with intact, typical anatomy all walk, see, and hear, they talk as well. Furthermore, walking, seeing, and hearing all develop as an interaction of experience with biological machinery, as does use of spoken language. But not all individuals walk and talk the same. We talk about someone having a good eye or a good ear, as in a good ear for accents. Thus, this anatomically linked set of abilities can be thought of as the foundation for developing certain skills through practice and perhaps training. Listeners should be able to improve their auditory abilities just as walkers can become skilled athletes, depending on the kinds of experiences they have, the motivation to pursue these experiences, and the effort they expend on them.

We may not think of listening as a skill, except when talking about some students who seem to lack it. However, given that auditory perception can be improved by specific kinds of experience and training, it seems plausible to consider this as a skill. For example, trained phoneticians can hear fine phonetic detail that typical listeners would never hear but can be seen in spectrograms (cf. Ladefoged, 2003) and this expertise has effects on brain structure (Golestani, Price, & Scott, 2011). Piano tuners spend tens of thousands of hours training their listening skills (Capleton, 2007), perform better on auditory tasks, and have associated neural changes (Teki et al., 2012). Expert musicians can learn tonal language pitch contours better than nonexperts and show changes in auditory neural coding (Wong, Skoe, Russo, Dees, & Kraus, 2007). And some blind individuals develop the ability to use echolocation to

navigate in the world without sight (Kolarik, Cirstea, Pardhan, & Moore, 2014). These are all cases of specific listening expertise developed from specific kinds of auditory experience with particular listening goals. While it is likely that not every listener can develop the same level of performance from the same amount and type of experience, it does seem that experience is a critical part of the manifestation of these perceptual skills. For example, the different experiences that lead to expert violinists versus expert actresses show different patterns of neural response, when presented with the same speeches and violin pieces, based on their respective skills (Dick, Lee, Nusbaum, & Price, 2011). These examples suggest that listening can be thought of as a skill, in that it can be honed by experience—practice and instruction—to develop in different ways, with different strengths.

The fact that auditory perception can be improved in different ways is basic information about the operation of the auditory perception system. Training or practice does not simply increase the sensitivity of the system to signal amplitude or improve signal-to-noise overall. Nor does it simply improve discrimination among signals generally across the spectrum. Instead, different aspects of listening ability can be improved by specific types of training or specific experiences.

In a system that processes sound patterns as a chain of neural transformations from the bottom-up, signals are only affected by local representations inherent to each neural mechanism (see Nusbaum & Schwab, 1986). Experience with specific sound patterns could in principle modify the sensitivities of these mechanisms (e.g., increased discriminability of signal differences) and possibly modify the transforms. But these changes resulting from experience would be expected to operate within a range of signal similarity over patterns. In other words, to the extent that a new pattern is similar to an old pattern within some criterion, prior experiences might affect the new pattern. But other information such as listener goals, task demands, and cross-categorization based on previous, dissimilar patterns, should not provide a basis for generalization in a purely passive bottom-up system.

The fact that perceptual training or practice can affect processing outside a simple criterion of signal similarity suggests a more complex system, one that is quite different from a bottom-up transformation of acoustic patterns to neural signals that are then matched against memories of past auditory experiences. This complexity is further supported by the evidence that suggests that auditory pattern perception is robust against noise, distortion, and signal variability.

## 2. ACOUSTIC CHALLENGES TO PERCEPTION

We generally perceive auditory patterns as corresponding to meaningful acoustic events, despite highly convolved auditory scenes that occur in the real world. For example, the soundstage on the street we typically experience contains a number of acoustic events ranging from cars on the road, to dogs barking. These acoustic patterns combine as pressure waves moving eardrums in complex patterns after distortion by the pinna (the outer ear), echoes (reflections and reverberations), effects of sound source motion, and various other contributions of the physical environment. In other words, there are sound sources, some stationary and some moving, and there are physical objects that do not generate sounds but affect transmission of sounds, and there is the acoustic transformation of sound by the pinna that aids in localization of sounds. Each eardrum feeds the cochlea a time-varying signal that is the complex convolution of all the sounds and effects together. As a result of our biological endowment together with perceptual experience, listeners can separate the sound objects on the soundstage spatially and identify them (e.g., Bregman, 1990; Cherry, 1953). The ability to separate signals into acoustic events or acoustic forms, a parsing of the soundstage, is one area of auditory perception that could in principle be enhanced as an auditory skill.

But beyond the problem of segmenting the soundstage into auditory forms there is the problem of recognizing those forms as corresponding to specific acoustic events such as a clock ticking or a dog barking. Once the signals from different sources are separated, reducing interference among the signals, there is the problem of recognizing an acoustic pattern as arising from a particular acoustic event. We generally have the sense that this is something we do with great effectiveness and efficiency. However even for the same acoustic event, acoustic patterns that reach our ears are seldom identical. Acoustic patterns change by a host of distortions such as reflection, reverberation, distance, even humidity in the air. The acoustic pattern produced on a stage is not the same as the pattern that reaches the middle seats of an auditorium. Thus, there is wide variation in the acoustic patterns that we can interpret as the same acoustic event.

This problem is a classic one in speech perception identified as a lack of invariance between acoustic patterns and the intended linguistic meaning. The sound spectrograph, in displaying the acoustic properties of speech as

three-dimensional time × frequency × amplitude plots revealed the spectrotemporal patterns of speech (Potter, Kopp, & Green, 1947). Between talkers, there is variation in vocal tract size and shape that translates into differences in the acoustic realization of phonemes (Fant, 1960; Stevens, 1998). However, even local changes over time in linguistic experience (Cooper, 1974; Evans & Iverson, 2007), affective state (Barrett & Paus, 2002), speaking rate (Gay, 1978; Miller & Baer, 1983), and fatigue (Lindblom, 1963; Moon & Lindblom, 1994) can alter the acoustic realization of a given phoneme. Beyond this, there is clear evidence that idiosyncratic articulatory differences in how individuals produce phonemes result in acoustic differences (Liberman, Cooper, Harris, MacNeilage, & Studdert-Kennedy, 1967). Similar sources of variability hold for higher levels of linguistic representation, such as syllabic, lexical, prosodic, and sentential levels of analysis (cf. Heald & Nusbaum, 2014). Moreover, a highly variable acoustic signal is by no means unique to speech. In music, individuals have a perception of melodic stability or preservation of a melodic "gestalt" despite changes in tempo (Handel, 1993; Monahan, 1993), pitch height or chroma (Handel, 1989), and instrumental timbre (Zhu, Chen, Galvin, & Fu, 2011). In fact, perhaps with a few contrived exceptions (such as listening to the same audio recording with the same speakers in the same room with the same background noise from the same physical location), we are not exposed to the same acoustic pattern of a particular auditory object twice.

If more than one acoustic pattern corresponds to a particular interpretation, this does not pose any particular computational problem. There are a number of different ways to approach this from simply enumerating exemplars of the interpretation to more complex ways of generalizing recognition depending on the disparities among the patterns. This kind of mapping can be solved by a variety of deterministic computational models. However, when a particular pattern has more than one interpretation, it presents a substantial computational problem. A pattern that maps to many alternative interpretations is a nondeterministic computational problem (see Nusbaum & Magnuson, 1997). The many-to-many mapping problem, often termed the lack of invariance problem, has limited the performance of speech recognition computer systems. Even as performance improves in these systems, the improvements are gained by computational brute force and still do not equal human performance, especially in cases of environmentally challenging listening situations involving noise and distortion. Nusbaum and Magnuson (1997)

suggested that theoretical proposals such as motor theory (e.g., Mattingly & Liberman, 1985) may be insufficient.

One possibility is that perceptual stability arises from the ability to form and use categories or classes of functional equivalence. It is a longstanding assertion in cognitive psychology that categorization serves to reduce psychologically irrelevant variability, carving the world up into meaningful parts (Bruner, Goodnow, & Austin, 1956). In audition, some have argued that the categorical nature of speech perception originates in the architecture of the perceptual system (Elman & McClelland, 1986; Holt & Lotto, 2010). More recent theories have suggested that speech categories arise out of sensitivity to the statistical distribution of occurrences of speech tokens (for a review, see Feldman, Griffiths, Goldwater, & Morgan, 2013). Indeed, it has been proposed that the ability to extract statistical regularities in one's environment, which could occur by an unsupervised or implicit process, shapes our perceptual categories in both speech (cf. Maye & Gerken, 2000; Maye, Werker, & Gerken, 2002; Strange & Jenkins, 1978; Werker & Polka, 1993; Werker & Tees, 1984) and music (cf. Lynch & Eilers, 1991, 1992; Lynch, Eilers, Oller, & Urbano, 1990; Soley & Hannon, 2010; Van Hedger, Heald, Huang, Rutstein, & Nusbaum, 2016). An often-cited example in speech research is that an infant's ability to discriminate sounds in their native language increases with linguistic exposure, while the ability to discriminate sounds that are not linguistically functional in their native language decreases (Werker & Tees, 1983). Further work in speech development by Nittrouer and colleagues has shown that the shaping of perceptual sensitivities and acoustic to phonetic mappings by one's native language experience occurs throughout adolescence, indicating that individuals remain sensitive to the statistical regularities of acoustic cues and how they covary with sound meaning distinctions throughout their development (e.g., Nittrouer & Lowenstein, 2007; Nittrouer & Miller, 1997). Therefore, it seems that given enough listening experience, individuals are able to learn how multiple acoustic cues work in concert to denote a particular meaning, even when no single cue is necessary or sufficient. As with any skill, practice is critical to improving performance.

## 3. CATEGORY KNOWLEDGE

Beyond learning the statistical structure of sound patterns, however, there is a higher level of knowledge. Developing a skill includes cognitive

structures that can organize and relate patterns—category structure. Individuals are not only sensitive to the statistical regularities of items that give rise to functional classes or categories, but also to the systematic regularities *among* the resulting categories themselves. This hierarchical source of information, which goes beyond the information within any specific individual category, could aid in disambiguating a physical signal that has multiple meanings. Consider the examples of speech and music: Knowledge of the organization and structure among the constituent categories of each system provides a map of category relationships that can constrain perception. Listeners may possess category knowledge that works collectively as a long-term, experientially defined system to orchestrate a cohesive perceptual world (see Billman & Knutson, 1996; Bruner, 1973; Goldstone, Kersten, & Cavalho, 2012).

In music, the implied key of a musical piece organizes the interrelations among pitch classes in a hierarchical structure (Krumhansl & Kessler, 1982; Krumhansl & Shepard, 1979). Importantly, these hierarchical relations become strengthened as a function of listening experience, suggesting that experience with tonal areas or keys shapes how individuals organize pitch classes (cf. Krumhansl & Keil, 1982). Hierarchical relationships are also seen in speech among various phonemic classes, initially described as a featural system (e.g., Chomsky & Halle, 1968) with distributional constraints on phonemes and phonotactics. For a given talker, vowel categories are often discussed as occupying a vowel space that roughly corresponds to the speaker's articulatory space (Ladefoged & Broadbent, 1957). Some authors have posited that point vowels, which represent the extremes of the acoustic and articulatory space, may be used to calibrate changes in the space across individuals, as they systematically bound the rest of the vowel inventory (Gerstman, 1968; Joos, 1948; Lieberman, Crelin, & Klatt, 1972). Due to the concomitant experience of visual information and acoustic information (rooted in the physical process of speech sound production), there are also systematic relations that extend between modalities. For example, an auditory /ba/ paired with a visual /ga/ often yields the perceptual experience of /da/ due to the systematic relationship of place of articulation among those functional classes (McGurk & MacDonald, 1976).

Given these examples, it is clear that within both speech and music, perceptual categories are not isolated entities. Rather, listening experience over time confers systematicity that can be meaningful. Such relationships may be additionally important to ensure stability in a system that is heavily influenced by recent perceptual experience, as stability may exist through

interconnections within the category system. Long-term learning mechanisms may override short-term changes that are inconsistent with the system, while in other cases, allow for such changes to generalize to the rest of the system to achieve consistency.

## 4. PERCEPTUAL LEARNING

Given that listeners form systematic categorical knowledge that organizes perceptual processing of sound patterns, it is important to understand the relationship of this learning to the perceptual process. While there is clear evidence that listeners rapidly learn the statistical distributions of their acoustic environments, both for the formation of perceptual categories and the relationships that exist among them, auditory recognition models do not typically incorporate learning into the process of recognition. The general approach seems to be that "learning" is a developmental process that antedates full competence in perception and is governed by separable learning systems that are independent of the process of perception.

Older speech perception models such as feature-detector theories (e.g., Blumstein & Stevens, 1981), ecological theories (Fowler & Galantucci, 2005), motor theories (e.g., Liberman & Mattingly, 1985), and interactive theories (TRACE, e.g., McClelland & Elman, 1986; C-CuRe: McMurray & Jongman, 2011) do not describe any mechanism to establish or update perceptual representations or categorical knowledge. This means that these types of models implicitly assume that the representations that guide the perceptual process are more stable than plastic. While C-CuRE (McMurray & Jongman, 2011) might be thought of as highly adaptive by allowing different levels of abstraction to interact during perception, this model does not make explicit claims about how the representations that guide perception are established either in terms of the formation of auditory objects or the features that comprise them. For example, the identification of a given vowel depends on the first (F1) and second (F2) formant values, but some of these values will be ambiguous depending on the linguistic context and talker. According to C-CuRE, once the talker's vocal characteristics are known, a listener can make use of these formant values. The listener can compare the formant values of the given signal against the talker's average F1 and F2, helping to select the likely identification of the vowel. Importantly, for the C-CuRE model, feature meanings are part of the recognition system. While there is some suggestion that this

knowledge could be derived from linguistic input and may be amended, the model itself has remained agnostic as to how and when this information is obtained and updated by the listener. A similar issue arises in other interactive models of speech perception (e.g., TRACE: McClelland & Elman, 1986; Hebb-Trace: Mirman, McClelland, & Holt, 2006) and models of pitch perception (e.g., Anantharaman, Krishnamurthy, & Feth, 1993; Gockel, Moore, & Carlyon, 2001).

Of course, there are exceptions in which some models explicitly incorporate sensitivity to context and dynamic processing (see Case, Tuller, Ding, & Kelso, 1995; Kleinschmidt & Jaeger, 2015; Lancia & Winter, 2013; Mirman et al., 2006; Tuller, Case, Ding, & Kelso, 1994). However, these exceptions are framed as computational—mathematical models of processing that are directed at explaining patterns of behavioral results and not typically related to specific cognitive or neural mechanisms. Indeed, it is not at all clear how these can be related to the neurobiology of auditory perception, which often seems to come at the problem of explaining auditory perception from an entirely different direction.

## 4.1 Evidence of Plasticity

There is clear empirical evidence for neural changes due to learning that have been incorporated into some neurobiological models of general auditory perception (McLachlan & Wilson, 2010; Shamma & Fritz, 2014; Weinberger, 2004, 2015). But this is less true for neurobiological models of speech perception, which traditionally limit consideration to perisylvian language areas (Fitch, Miller, & Tallal, 1997; Friederici, 2012; Hickok & Poeppel, 2007; Rauschecker & Scott, 2009) and ignore brain regions that have been implicated in category learning, such as the striatum, the thalamus, and prefrontal attention-working memory regions (Ashby & Maddox, 2005; McClelland, McNaughton, & O'Reilly, 1995). Further, the restriction of speech models to perisylvian language areas marks an extreme cortical myopia (cf. Parvizi, 2009) of the auditory system, as it ignores the corticofugal pathways that exist between cortical and subcortical regions such as the medial geniculate nucleus in the thalamus, the inferior colliculus in the midbrain, the superior olive and cochlear nucleus in the pons, all the way down to the inner ear. Research has shown that higher-level cognitive functions can reorganize subcortical structures as low as the cochlea.

For example, selective attention or discrimination training has been demonstrated to enhance the spectral peaks of evoked otoacoustic emissions

produced in the inner ear (de Boer & Thornton, 2008; Giard, Collet, Bouchet, & Pernier, 1994; Maison, Micheyl, & Collet, 2001). Inclusion of the corticofugal system in neurobiological models of speech would allow the system, through feedback and top-down control, to adapt to ambiguity or change in the speech signal by selectively enhancing the most diagnostic spectral cues for a given talker or expected circumstance, even before it reaches perisylvian language areas. Including the corticofugal system can thus drastically change how extant models, which are entirely cortical, explain top-down effects in speech and music. While the omission of corticofugal pathways and brain regions associated with category learning is likely not an intentional omission but a simplification for the sake of experimental tractability, it is clear that such an omission has large-scale consequences for modeling auditory perception, speech, or otherwise. Indeed, the inclusion of learning areas and adaptive corticofugal connections on auditory processing requires a vastly different view of perception, in that even the earliest moments of auditory processing are guided by higher cognitive processing via expectations and listening goals. In this sense, it is unlikely that learning and adaptability can be simply grafted on top of current cortical models of perception. The very notion that learning and adaptive connections could be omitted, however (even for the sake of simplicity) is in essence, a tacit statement that the representations that guide recognition are more stable than plastic.

The theoretic assumption that perceptual representations are more stable than plastic may be influenced by our subjective experience of the world as perceptually stable. In music, relative perceptual constancy can be found for a given melody despite changes in key, tempo, or instrument. Similarly, in speech, a given speech sound can be recognized despite changes in phonetic environment and talker. Listeners are certainly not insensitive to acoustic differences other than pitch or acoustic-phonetics, but different listening goals can arguably shape the way listeners attend to acoustic variability that is not relevant to those goals. Listening goals organize perceptual attention to select properties that are relevant to those goals and direct attention away from cues that are not (cf. Goldstone & Hendrickson, 2010). Change deafness demonstrates that simply changing listening goals significantly alters detection of a change in talker in a phone conversation (Fenn et al., 2011). Participants did not detect an unheralded change in talker during a phone conversation, but could detect the change if told to explicitly monitor for it. This suggests that listening goals modulate how we understand acoustic patterns, and this in turn shapes the way attention is directed towards acoustic variability.

## 4.2 Perceptual Stability and Attention

Perceptual classification or categorization here should not be confused with categorical perception (cf. Holt & Lotto, 2010). But categorical perception is a clear demonstration of the way that cognitive categories stabilize perceptual experience reducing the perceptual impact of putatively irrelevant acoustic variability—at least irrelevant to the immediate goals of perception. *Categorical perception*, as it has been defined in speech research, refers to perception of speech sounds based on their phonetic category rather than their auditory character and thus identification and discrimination changes abruptly from one category to another (e.g., Liberman, Harris, Hoffman, & Griffith, 1957). Categorical perception suggests that despite changes in listening goals, perceptual discrimination of any two speech sounds (although only true for some consonants) depends on the probability of classifying these stimuli as belonging to different categories (e.g., Studdert-Kennedy, Liberman, Harris, & Cooper, 1970). Categorization more generally refers a classification process in which cues that indicate a between-category difference are attended more than cues that differentiate tokens within a category (Harnad, 1987). Indeed, even within the earliest examples of categorical perception (a phenomenon that, in theory, completely attenuates within-category variability), there appears to be some retention of within-category discriminability (e.g., see Liberman et al., 1957). Native English listeners can reliably rate or respond to some acoustic realizations of phonetic categories (e.g., "ba") as better versions than others (e.g., Carney, Widin, & Viemeister, 1977; Iverson & Kuhl, 1995; Pisoni & Lazarus, 1974; Pisoni & Tash, 1974). Other studies have shown that within-category phonetic variability affects subsequent lexical processing (Dahan, Magnuson, Tanenhaus, & Hogan, 2001; Gow, McMurray, & Tanenhaus, 2003; McMurray, Tanenhaus, & Aslin, 2002). There is also evidence that within-category discrimination can exceed what would be predicted from category identification responses (Halpern & Zatorre, 1979). Indeed, Holt and Lotto (2010) have suggested that the task structure typically employed in categorical perception tasks may be what is driving the manifestation of within category homogeneity that is characteristic of categorical perception. This means that the experiential stability of categorical perception in speech does not eliminate the availability of acoustic variability, but it may reflect an attentional focus on phonetically relevant perceptual classification guided by knowledge, experience, and context.

However, this kind of perceptual stability is not specific to speech alone. In music, the perception of pitch chroma categories among absolute pitch (AP) possessors is categorical in the sense that AP possessors show sharp identification boundaries between note categories (e.g., Ward & Burns, 1982). However, AP possessors also show reliable within-category differentiation when providing goodness judgments within a note category (e.g., Levitin & Rogers, 2005). Graded evaluations within a category are further seen in musical intervals, where sharp category boundaries indicative of categorical perception are also generally observed at least for musicians (Siegel & Siegel, 1977). Another way of stating this is that listening goals defined by the task structure modulate the way attention is directed towards acoustic variance. Given that music often requires the use of between category pitch information (e.g., in slurs), the goals in music perception and production are somewhat different from the goals in speech perception and production and could potentially account for any differences in perceptual processing.

While there is clear evidence that individuals possess the ability to attend to acoustic variability, even within perceptual categories, it is still unclear from available research whether listeners are influenced by acoustic variability that is attenuated due to listening goals. More specifically, it is unclear whether the representations that guide perception are influenced by subtle, within-category acoustic variability, even if this variability appears to be functionally irrelevant for current listening goals. Putatively irrelevant acoustic variability, even if not consciously experienced, may still affect subsequent perception. For example, Gureckis and Goldstone (2008) have argued that the preservation of variability (in our case, the acoustic trace independent of the way in which the acoustics relate to an established category structure due to a current listening goal) allows for perceptual plasticity within a system, as adaptability can only be achieved if individuals are sensitive (consciously or unconsciously) to potentially behavioral relevant changes in within-category structure. In this sense, without the preservation of variability listeners would fail to adapt to situations in which the identity of perceptual objects rapidly change. Indeed, there is a growing body of evidence supporting the view that the preservation of acoustic variability can be used in service of flexibly adapting a previously established category or instantiating a novel category. In speech, adult listeners are able to learn perceptual categories not present in their native language, even when the acoustic cues needed to learn the novel category structure are in direct conflict with a preexisting category structure. Adult native Japanese listeners,

who presumably become insensitive to the acoustic differences between /r/ and /l/ categories through accrued experience listening to Japanese, are nevertheless able to learn this nonnative discrimination through explicit perceptual training (Bradlow, Pisoni, Akahane-Yamada, & Tohkura, 1997; Ingvalson, Holt, & McClelland, 2012; Lively, Pisoni, Yamada, Tokhura, & Yamada, 1994), rapid incidental perceptual learning (Lim & Holt, 2011), as well as through the accrual of time residing in English-speaking countries (Ingvalson, McClelland, & Holt, 2011). Further, adult English speakers are able to learn the nonnative Thai prevoicing contrast, which functionally splits their native /b/ category (Pisoni, Aslin, Perey, & Hennessy, 1982) and to distinguish between different Zulu clicks, which make use of completely novel acoustic cues (Best, McRoberts, & Sithole, 1988).

Beyond retaining an ability to form non-native perceptual categories in adulthood, there is also clear evidence that individuals are able to update and amend the representations that guide their processing of native speech. Clarke and Luce (2005) showed that within moments of listening to a new speaker, listeners modify their classification of stop consonants to reflect the new speaker's productions, suggesting that linguistic representations are extremely plastic in that they can be adjusted online to optimize perception. This finding has been replicated in a study that further showed that participants' lexical decisions reflect recently heard acoustic probability distributions (Clayards, Tanenhaus, Aslin, & Jacobs, 2008).

Perceptual flexibility also can be demonstrated at a higher level, presumably due to discernible higher-order structure. Work in our lab has demonstrated that individuals are able to rapidly learn synthetic speech produced by rule that is defined by poor and often misleading acoustic cues. In this research, no words ever repeat during testing or training, so that the learning of a particular synthesizer is thought to entail the redirection of attention to the most diagnostic and behaviorally relevant acoustic cues across multiple phonemic categories in concert (see Fenn, Nusbaum, & Margoliash, 2003; Francis & Nusbaum, 2009; Francis, Nusbaum, & Fenn, 2007; Nusbaum & Schwab, 1986) in much the same way as in learning new phonetic categories (Francis & Nusbaum, 2002). Given these studies, it appears that the process of categorization in pursuit of current listening goals does not completely attenuate acoustic variability, suggesting that auditory information that is not used initially in perceptual decisions may be retained in long-term memory to affect other processing.

## 4.3 Absolute Pitch as a Skill

Beyond speech, the representations that guide music perception also appear to be remarkably flexible. Wong, Roy, and Margulis (2009) have demonstrated that individuals are able to learn multiple musical systems through passive listening exposure. This "bimusicality" is not merely the storage of two, modular systems of music (Wong, Chan, Roy, & Margulis, 2011); though it is unclear whether early exposure (i.e., within a putative critical period) is necessary to develop this knowledge. In support of the notion that even adult listeners can come to understand a novel musical system that may parse pitch space in a conflicting way compared to Western music, Loui and Wessel (2008) have demonstrated that adult listeners of Western music are able to learn a novel artificial musical grammar. In their paradigm, individuals heard melodies composed using the Bohlen–Pierce scale—a musical system that is strikingly different from Western music, as it consists of 13 equally spaced notes within a three-octave range as opposed to 12 equally spaced notes within a two-octave range. Nevertheless, after mere minutes of listening to 15 Bohlen–Pierce melodies that conformed to a finite-state grammar, listeners were able to recognize these previously heard melodies as well as generalize the rules of the finite-state grammar to novel melodies.

Even within the Western musical system, adults display plasticity for learning categories thought to be unlearnable in adulthood. A particularly salient example of adult plasticity within Western music learning comes from the phenomenon of AP—the ability to name or produce any musical note without the aid of a reference note (see Deutsch, 2013 for a review). AP has been conceptualized as a rare ability, manifesting in as few as 1 in every 10,000 individuals in Western cultures (Bachem, 1955), though the mechanisms of AP acquisition are still debated. While there is some research arguing for a genetic predisposition underlying AP (e.g., Baharloo, Johnston, Service, Gitschier, & Freimer, 1998; Theusch, Basu, & Gitschier, 2009), with even some accounts claiming that AP requires little or no environmental shaping (Ross, Olson, & Gore, 2003), most theories of AP acquisition adhere to an early learning framework (e.g. Crozier, 1997). This framework predicts that only individuals with early note naming experience would be candidates for developing AP categories. As such, previously naive adults should not be able to learn AP. This early learning argument of AP has been further explained as a "loss" of AP processing without early interventions, either from music or language (i.e., tonal languages), in which AP is emphasized (cf. Deutsch, Henthorn, & Dolson, 2004; Sergeant & Roche, 1973).

In support of this explanation, infants appear to process pitch both absolutely and relatively, though they switch to relative pitch cues when AP cues become unreliable (Saffran, Reeck, Niebuhr, & Wilson, 2005).

The flexibility of auditory classification and sound understanding appears to be a general principle of auditory perception. In other words, although perception of known categorical systems (phonetics, music) is very stable, in spite of the stability the category organizations and categorization process bring to recognition, this perception seems to remain mutable even into adulthood. The notion that in early childhood, by exposure, use or instruction, we acquire a system of knowledge that guides and shapes perception and experience of sound, and once acquired, this is a frozen, fixed system, appears to be incorrect. But to retain this flexibility, to revise understanding, there needs to be a cognitive substrate capable of supporting this.

As indicated for speech perception, although categorical knowledge can quickly determine perception, putatively irrelevant acoustic variability (or irrelevant acoustic signal properties) is not lost as a result of processing. Depending on task goals or demands of perception, listeners appear to be able to access auditory information that is not typically used in recognition. Similarly, for music, there is mounting evidence that most listeners (regardless of possessing AP) can perceive and remember AP information, albeit not at the performance level of someone with AP. If AP is a rare ability, dependent on a particular unique genetic substrate, acquired by training within a childhood critical period, listeners without AP should be naive about absolute pitch information in music, beyond the ability to discriminate those pitches, subject to typical psychoacoustic limitations.

But research has demonstrated clearly that listeners without AP can hear when familiar music recordings have been subtly shifted in pitch (e.g., Schellenberg & Trehub, 2003; Terhardt & Seewan, 1983). As these individuals cannot explicitly name the musical notes involved (at least without a reference note), this ability has been termed "implicit AP." Moreover, Van Hedger, Heald, and Nusbaum (2016) have shown that implicit AP exists for a familiar nonmusical sound (the tone used by the FCC to censor broadcasts) even though it does not align in frequency with a musical note. Furthermore, Van Hedger, Heald, Huang, et al. (2016) have shown that adult listeners without AP can also judge, above chance, whether an isolated musical note conforms to conventional Western intonation (in which the "A" above "middle C" is equal to 440 Hz)—an ability thought to be unique to AP possessors. These results demonstrate long-term auditory sensory memory in adult listeners without AP that can undergird the ability of AP listeners.

In fact the studies show that AP listeners are even better than non-AP listeners at these tasks, demonstrating the effect of skill acquisition as opposed to simple ability.

Moreover, if listeners without AP can detect pitch changes, perception and memory for the AP of a prior sound experience does not depend on a special ability. Moreover, this ability in non-AP listeners does not depend on any musical training. Although it is typically assumed that AP depends on early musical training, and thus on knowledge of a musical category system, at least "implicit AP" does not use knowledge of explicit musical note category labels. The fact that mistuning within a note category can also be detected by non-AP listeners strengthens that conclusion.

However, although these studies focus on the ability to perceive the correct or incorrect tuning of a sound, it is probably most appropriate to characterize this as involving both perception and memory for AP. Non-AP listeners in these studies can only determine the correctness of the tuning of a familiar musical passage, not a previously unknown piece of music. This means that listeners must have durable auditory memories for the sound of the notes used in familiar music, albeit not the note names of course. It further means, that when presented with a musical passage, these auditory memories of familiar music provide the basis for judging tuning. This suggests that there is a ubiquitous long-term auditory sensory memory, which may provide the foundation for learning true AP. This further suggests that AP need not depend either on a special genetic endowment or on musical training within a critical period. Rather, acquisition of AP may be based on the use of specific cognitive mechanisms to learn the relationships between pitch information held in memory and associated note categories. This could be tested by finding out if it is possible to train adults to have AP!

## 4.4 Training Absolute Pitch in Adults

It is interesting to note that a recent study reported by Gervain et al. (2013) argued that learning AP is governed by a critical period within which musical training (pitch-label associative learning) must occur. They claimed that the drug valproate (also called Depakote) would open the critical period for learning and showed (in one condition, but questionably not another) that with the drug administered to adults, there were significant improvements in AP performance compared to a no-drug condition. Unfortunately, there is little evidence to support the claim that valproate "opens a critical period" for learning, although it does affect a number of neurotransmitters known to be involved in working memory and neural

plasticity. However, the study does show that some adults, under some conditions, can improve above chance in learning to assign notes to pitches.

It is important to note that in the past, a few studies reported improvements in absolute note identification without drug administration (Brady, 1970; Rush, 1989) contra Gervain et al. (2013). A more recent study by Van Hedger, Heald, Koch, and Nusbaum (2015) rejects completely the argument of Gervain et al. (2013). According to Gervain et al. (2013), learning AP without drug administration should be impossible for adults past the putative critical period. However, Van Hedger, Heald, Koch, et al. (2015) demonstrated significant improvements in AP in a single training session of adult listeners with no drug administration. Further, the results showed that auditory working memory capacity significantly predicted the amount of AP learning shown by the adults. This empirically rejects the claim that (1) a drug is needed to reopen a critical period for AP, (2) that there is any critical period for AP, (3) that there is any specific-to-AP biological endowment necessary for learning AP. Furthermore, follow up studies demonstrated that the learning demonstrated by these adults is retained for 5–7 months after the end of training. The results demonstrate that a general auditory cognitive mechanism, auditory working memory, is important for learning AP.

Despite evidence that it is possible to train AP in adults past a putative critical period, one might still argue that the adult learning of AP categories represents a fundamentally different phenomenon than that of early acquired AP. First, the current evidence of adult plasticity in acquiring AP does not yet reach the level of performance as that of AP listeners who acquired note categories early in life. Although learning does last for months without subsequent practice (Van Hedger, Heald, Koch, et al., 2015), the accuracy level is below that of "early learners." Further, we know from studies of "implicit AP" such as Schellenberg and Trehub (e.g., 2003) and our own research (Van Hedger, Heald, Huang, et al., 2016) that even listeners without AP have some ability to detect pitch mismatches with remembered sounds. While this ability establishes clear evidence of long-term auditory sensory memory in listeners that can provide the foundation for this learning (Van Hedger, Heald, Huang, et al., 2016), it could be that without early training "true" AP cannot be achieved in adult listeners.

The critical period view of AP shares much in common with views of bird song development. For example, zebra finches acquire their songs through exposure to a tutor song during a critical period of development, and the song is basically crystallized in that form after acquisition

(Doupe & Kuhl, 2008). From this model system, one would imagine that AP that is learned early in life establishes pitch-note relationships systematically (cf. Krumhansl & Keil, 1982), and that these relationships are then crystallized in memory to form absolute references for note classification later in life. Note categories within an early acquired AP population are thought to be highly stable once established (Ward & Burns, 1982), only being alterable in very limited circumstances, such as through physiological changes to the auditory system as a result of aging (cf. Athos et al., 2007) or pharmaceutical interventions (e.g., Kobayashi, Nisijma, Ehara, Otsuka, & Kato, 2001).

However, recent empirical evidence has suggested that even within this early acquired AP population, there exists a great deal of plasticity in note category representations that is tied to particular listening experiences. Wilson, Lusher, Martin, Rayner, and McLachlan (2012) reported reductions in AP ability as a function of whether an individual plays a "movable *do*" instrument (i.e., an instrument in which a notated "C" actually belongs to a different pitch chroma category, such as "F"), suggesting that nascent AP abilities might be undone through inconsistent sound-to-category mappings. Dohn, Garza-Villarreal, Ribe, Wallentin, and Vuust (2014) reported differences in note identification accuracy among AP possessors that could be explained by whether one was actively playing a musical instrument, suggesting that AP ability might be "tuned up" by recent musical experience.

Both of these studies indicate that particular regularities in acoustic experience may affect overall note category accuracy within an AP population, though they do not speak to whether the *structure* of the note categories can be altered through experience once they are acquired. Indeed, one of the hallmarks of AP is not only being able to accurately label a given pitch with its note category (e.g., C#), but also provide a goodness rating of how well that pitch conforms to the category (e.g., flat, in-tune, or sharp). Presumably, this ability to label some category members as better than others stems from either a fixed note—frequency association established early in life, or through the consistent environmental exposure of listening to music that is tuned to a very specific standard (e.g., in which the "A" above "middle C" is tuned to 440 Hz).

Adopting the first explanation, plasticity of AP category structure should not be possible. Adopting the second explanation, AP category structure should be modifiable and tied to the statistical regularities of hearing particular tunings in the environment. Our previous work has clearly demonstrated evidence in support of this second explanation—that is, the

structure of note categories for AP possessors is plastic and dependent on how music is tuned in the current listening environment (Hedger, Heald, & Nusbaum, 2013). In our paradigm, AP possessors assigned goodness ratings to isolated musical notes. Not surprisingly, in-tune notes (according to an A = 440 Hz standard) were rated as more "in-tune" than notes that deviated from this standard by one-third of a note category. However, after listening to a symphony that was slowly flattened by one-third of a note category, the same participants began rating similarly flattened versions of isolated notes as more "in-tune" than the notes that were in-tune based off of the A = 440 Hz standard. These findings suggest that AP note categories are held in place by the recent listening environment, not by a fixed and immutable note–frequency association that is established early in life.

Research demonstrates that it is possible to train AP in adult listeners, thereby rejecting theories of a special AP genetic endowment and a critical period for acquisition. The tuning of the mental note categories in AP listeners is also mutable. Thus, musical note representations do not appear to be crystallized after a critical period of development, as all listeners (AP and non-AP) appear to have flexible auditory systems that can develop new sound representations as well as modify existing sound representations. This flexibility appears to be grounded in a long-term auditory sensory memory, bolstering the argument that AP is a skill, not a special ability. AP appears to be a skill that makes use of general cognitive mechanisms, including auditory working memory, long-term auditory sensory memory, and involves learning associations between the formal systematic knowledge structures of music and the sound patterns used in music. As a skill AP is developed by experience, and reinforced culturally by musical convention. In many respects, this description shares much in common with speech perception, which suggests that there is good reason to view speech perception as a skill in much the same way.

## 5. GENERALIZATION AND TRANSFER

The argument that auditory perception is trainable and plastic in response to training and experience might not be sufficient to talk about a skill. This is an argument about the nature of perception and not about skill per se. But to develop a skill, that is, to reach levels of specialized performance by enhancing particular aspects of ability through practice or training,

this is a necessary foundation. That should be obvious. Equally obvious is the fact that something is not really a skill if it is restricted to the training experiences that improved performance. For a skill to be useful, that is, to be truly considered as a skill it must provide performance enhancement in contexts and situations that go beyond the experiences that shaped development. It is critical for a skill to generalize from training to novel experiences. AP is an example of one such skill of auditory perception, and speech perception is arguably another.

## 5.1 Recognizing Novel Acoustic Patterns

We have already discussed how speech perception is robust to acoustic challenges to recognition including the lack of invariance problem. Listeners recognize speech in spite of speaking rate variability, changes in talker, acoustic pattern variability across phonetic context, as well as distortion and noise of various kinds. This is broadly how listeners can apply their knowledge of spoken language to recognize speech over a broad range of novel changes in acoustic patterns. On the one hand, if speech perception operated only on the basis of similarity of pattern structure, we might not consider that a robust skill. For any set of patterns, by application of a Minkowski distance metric, there are sets of patterns that bear some similarity to those patterns with dissimilarity increasing along dimensions of acoustic properties. If the first set of patterns is learned, and attention is directed to the acoustic properties of those patterns for purpose of classifying them, it is possible to make a simple, first-principled statement about classification of patterns that were not learned. If we measure the ability to discriminate acoustic properties along the Minkowski distance metric for learners, we can predict simply which novel patterns can be classified appropriately simply because of a failure to discriminate them from the learned patterns. This will look like generalization of learning, or it might be taken as transfer from one set of learned patterns to an unlearned set. But this is not particularly interesting. Skilled performance depends on generalization to novel patterns, not simply acoustically different patterns, as transfer to physically different patterns could occur simply based on a failure of discrimination. As such, generalization to novel patterns that can easily be discriminated from a learned set provides a test of abstract learning.

Listeners are able to direct their attention towards the acoustic cues that are the most phonetically diagnostic for a given source. This process can be distinct from simple rote memorization, reflecting instead a kind of abstract knowledge for how to direct one's attention to a given talker's speech. More

specifically, when presented with computer-generated synthetic speech that is hard to understand, listeners can adapt to this speech. After eight 1-h training sessions, listeners show huge improvements in recognition performance (Schwab, Nusbaum, & Pisoni, 1985) gaining 50 percentage points from a base level pretest performance of around 25% correct word recognition. Moreover, this improvement comes from testing on novel words that were not part of training. Further, these improvements lasted for at least 6 months following the end of training. Greenspan, Nusbaum, and Pisoni (1988) demonstrated that training on isolated words generalizes to sentence stimuli, and vice versa. Even training on a very small set of spoken words in a single 1-h session produces some generalization, as it improves recognition performance for previously unheard words.

The fact that adult listeners can learn to improve significantly in recognition of speech produced by a very flawed and defective computer speech synthesizer demonstrates plasticity. But that this performance is accomplished in recognition of words that were not part of the training, that possess novel phonetic contexts, is a clear demonstration of generalization. The fact that this generalized learning is robust over 6 months indicates that this is more skill-like than a biological endowment fixed in early childhood. Indeed, we can ask what leads to this robust and enduring generalization.

## 5.2 Sleep and Perceptual Generalization

Fenn et al. (2003) addressed this question by testing the hypothesis that sleep plays an important role in stabilizing learning to last durably over time. Using the training paradigm for synthetic speech, listeners were given one session of training with a pretest and a posttest. For some of the listeners, the posttest was delayed by a waking 12-h period. For some, the delay of 12-h included a period of normal sleep. The results showed that listeners' recognition performance dropped by 10 percentage points over a waking retention interval but did not drop over a retention interval with sleep. In a within-subjects condition, it was possible to see the reduction in performance by evening with performance restored after sleep. After eliminating circadian explanations, the results indicated that sleep plays a significant role in ameliorating the effects of performance loss and stabilizing memory for long-term retrieval. Two additional groups, trained either in the morning or the evening and then tested 24 h later demonstrated that sleep prevented subsequent loss in addition to restoring any previous losses. For generalized learning this research demonstrated that sleep plays an important role in

retaining the effects of training for speech perception. In a more recent study, Fenn, Margoliash, and Nusbaum (2013) demonstrated that sleep's consolidation effects on learning to recognize synthetic speech is specific to generalized learning. Rote learning, or learning from a small number of repeated words, did not show consolidation benefits from sleep, despite showing a loss over a waking interval. What was important about this was the fact that even for the rote learning condition, there was a small amount of generalization of learning to the novel words. Even though there was no consolidation of the rote learning for a repeated small set of words, the generalization effect of learning these words did show consolidation. Thus there is a clear difference in the memory processes that retain specific experiences from learning and those that generalize beyond these experiences, in that sleep appears to only benefit the latter.

It is important to note that the consolidation of generalized learning during sleep is not unique to learning to perceive speech. Brawn, Fenn, Nusbaum, and Margoliash (2008) demonstrated that sleep consolidates generalized sensorimotor learning for complex video games much in the same way that it does for speech perception. For sensorimotor learning (Brawn et al., 2008) and motor learning (Brawn, Fenn, Nusbaum, & Margoliash, 2010), sleep restores what is lost over a waking retention period and protects against subsequent interference. It is also important to note that this pattern of consolidation is not unique to human learning. Brawn, Nusbaum, and Margoliash (2010) demonstrated that starlings show the same pattern of sleep consolidation in learning to discriminate among novel bird songs. Demonstrating that starlings show the same kind of sleep consolidation for perceptual learning as humans allowed a test of the basis for forgetting during the waking retention interval. Brawn, Nusbaum, and Margoliash (2013) trained starlings on one song discrimination task (task A) and then trained them on a second, interfering song discrimination task (task B). Interference from task B adversely affected task A performance (retroactive interference) and task A also interfered with task B (proactive interference). Sleep eliminated the effect of interference and restored performance. Even more interesting, however, was the demonstration that sleep did not "erase" the interfering task learning but instead separately consolidated both tasks, as if separating their memory representations.

Given that speech perception training demonstrates clear generalization as does sensorimotor training, it seems plausible that other auditory perceptual training should show generalization as well. As noted above, short-term exposure to detuned music shifts the tuning of the mental representation of

notes for listeners with AP, even for notes that were not in the original detuned music exposure (Hedger et al., 2013). Moreover, in training adult listeners to have AP, it is also the case that they show generalization of learning, even from a single session of training to notes that were not heard during training (Van Hedger, Heald, Koch, et al., 2015). As with perceptual learning of speech, learning musical note categories with the appropriate training appears to develop as a perceptual skill, which depends on general cognitive mechanisms such as long-term sensory storage and sleep consolidation.

Indeed, the evidence of plasticity and generalization in auditory perception for both speech and music suggests that both systems may be subserved by common perceptual learning mechanisms. Recent work exploring the relationship between speech and music processing has found mounting evidence that musical training improves several aspects of speech processing, though it is debated whether these transfer effects are due to general enhancements in auditory processing (e.g., pitch perception) versus an enhanced representation of phonological categories. Hypotheses such as OPERA (Patel, 2011) posit that musical training may enhance aspects of speech processing (1) when there is anatomical *overlap* between networks that process the acoustic features shared between music and speech, (2) when the perceptual *precision* required of musical training exceed that of general speech processing, (3) when the training of music elicits positive *emotions*, (4) when musical training is *repetitive*, and (5) when the musical training engages *attention*. Indeed, the OPERA hypothesis provides a framework for understanding many of the empirical findings within the music-to-speech transfer literature. Musical training helps individuals to detect speech in noise (Parbery-Clark, Skoe, Lam, & Kraus, 2009), presumably through strengthened auditory working memory, which requires directed attention. Musicians are also better able to use nonnative tonal contrasts to distinguish word meanings (Wong & Perrachione, 2007), presumably because musical training has made pitch processing more precise. This explanation can further be applied to the empirical findings that show that musicians are better able to subcortically track the pitch of emotional speech (Strait, Kraus, Skoe, & Ashley, 2009).

Recent work has further demonstrated that musical training can also influence the categorical perception of speech. Bidelman, Weiss, Moreno, and Alain (2014) found that musicians showed steeper identification functions of vowels that varied along a categorical speech continuum, and moreover these results could be modeled by changes at multiple levels of the auditory

pathway (both subcortical and cortical). In a similar study, Wu et al. (2015) found that Chinese musicians were better able to discriminate within-category lexical tone exemplars in a categorical perception task compared to nonmusicians, though, unlike Bidelman et al. (2014), the between-category differentiation between musicians and nonmusicians was comparable. Wu and colleagues interpret the within-category improvement among musicians in an OPERA framework, arguing that musicians have more precise representations of pitch that allow for fine-grained distinctions within a linguistic category.

Finally, there is emerging evidence that certain kinds of speech expertise may enhance musical processing, demonstrating a proof-of-concept of the bidirectionality of music—speech transfer effects. Specifically, nonmusician speakers of a tonal language (Cantonese) showed auditory processing advantages in pitch acuity and music perception that nonmusician speakers of English did not show (Bidelman, Hutka, & Moreno, 2013). While there is less evidence supporting this direction of transfer, this is perhaps not surprising, as speech expertise is ubiquitous in a way music expertise is not. Thus, transfer effects from speech to music processing are more constrained, as one has to design a study in which there (1) exists substantial differences in speech expertise, and (2) this difference in expertise must theoretically relate to some aspect of music processing (e.g., pitch perception).

How can these transfer effects between speech and music be interpreted in the larger context of auditory object plasticity? Given the evidence across speech and music that recent auditory events profoundly influence the perception of auditory objects within each system, it stands to reason that recent auditory experience from one system of knowledge (e.g., music) may influence subsequent auditory perception in the other system (e.g., speech), assuming there is overlap among particular acoustic features of both systems. Indeed, there is some empirical evidence to conceptually support this idea. An accumulating body of work has demonstrated that the perception of speech sounds is influenced by the long-term average spectrum (LTAS) of a preceding sound, even if that preceding sound is nonlinguistic in nature (e.g., Holt & Lotto, 2002; Holt, Lotto, & Kluender, 2000). This influence of nonlinguistic sounds on speech perception appears to reflect a general sensitivity to spectro-temporal distributional information, as the nonlinguistic preceding context can influence speech categorization even when it is not immediately preceding the to-be-categorized speech sound (Holt, 2005). While these results do not directly demonstrate that recent experience in music can influence the way in which a speech sound

is categorized, it is reasonable to predict that certain kinds of experiences in music or speech (e.g., a melody played in a particular frequency range) may alter the way in which subsequent speech sounds are perceived. This suggests that while skills may develop particular aspects of ability, the underlying cognitive substrates remain available to serve as the basis for transfer.

## 6. CONTEXT AND COGNITION

The empirical evidence demonstrates that specific training experiences can develop listening skills and that such changes persist after learning. The focus of training on specific stimuli with specific listening goals and contextual experience develops a base of perceptual knowledge in the form of an organized system of categories. However, transfer between skill areas suggests that while there is development of domain-specific knowledge, there is development of more general cognitive mechanisms such as auditory working memory and control of auditory selective attention. In order for listeners to be able to transfer experience between different domains of listening, or for listeners to be able to shift their listening goals within a single domain (e.g., shifting one's attention to the message of speech instead of the source of the speech), a listener must maintain acoustic pattern information even if it is not relevant for a particular listening goal or skill, as it may be relevant for subsequent goals or skills. For this reason, even non-relevant acoustic pattern information (in terms of either current goal or skill) may still be preserved and incorporated, if lawful, into the representations that guide perception. Indeed, listeners are faced with continual changes in how phonetic categories are acoustically realized over time at both a community level (Labov, 2001; Watson, Maclagan, & Harrington, 2000) and at an idiosyncratic level (Bauer, 1985; Evans & Iverson, 2007). As such, neural representations must preserve aspects of variability outside of processes that produce forms of perceptual constancy. Another way to say this, skills must be dynamic and sufficiently plastic to accommodate signal changes diachronically as well as to meet unexpected acoustic challenges synchronically.

Tuller and colleagues (Case et al., 1995; Tuller et al., 1994) have proposed a nonlinear dynamic model of speech perception that is highly context dependent. Robust perceptual recognition (i.e., perceptual constancy) occurs by attraction to "perceptual magnets" that are modified nonlinearly through experience. Crucial to their model, listeners remain sensitive to the fine-grain acoustic properties of auditory input as recent experience can

induce a shift in perception. A similar functionality has been the goal of a model proposed by Kleinschmidt and Jaeger (2015) in which perceptual stability in speech is achieved through recognition "strategies" that vary depending on the degree to which a signal is familiar, based on past experience. This flexible strategic approach based on prior familiarity is critical for successful perception, as a system that is rigidly fixed in acoustic-to-meaning mappings would fail to recognize (perhaps by misclassification) perceptual information that was distinct from past experience, whereas a system that is too flexible might require a listener to continually start from scratch. Robust recognition is not achieved through the activation of a fixed set of features, but through listening expectations based on the statistics of prior experience. In this way, perceptual constancy arising from such a system could be thought of as an emergent property that results from the comparison of prior experience to bottom-up information from (1) the signal and (2) recent listening experience (i.e., context).

Within a window of recent experience, what kinds of cues convey to a listener that a deviation from expectations has occurred? Listeners must flexibly shift between different situations that may have different underlying statistical distributions (Qian, Jaeger, & Aslin, 2012), using contextual cues that signal a change in an underlying statistical structure (Gebhart, Aslin, & Newport, 2009). One particularly clear and ecologically relevant contextual cue comes from a change in source information—that is, a change in talker for speech, or instrument for music. For example, when participants learn novel words from distributional probabilities of items across two unrelated artificial languages (i.e., that mark words using different distributional probabilities), they only show reliable transfer of learning across both languages when the differences between languages are contextually cued through different talkers (Weiss, Gerfen, & Mitchel, 2009). This is presumably because without a contextual cue to index the specific language, listeners must rely on the *overall* accrued statistics of their past experience in relation to the sample of language drawn from the current experience, which may be too noisy to adequately learn or deploy. More recent work has demonstrated that the kind of cueing necessary to parse incoming distributional information into multiple representations can come from temporal cues as well. Gonzales, Gerken, and Gómez (2015) found that infants could reliably differentiate statistical input from two accents if temporally separated. This suggests that even in the absence of a salient perceptual distinction between two sources of information (e.g., speaker), listeners can nevertheless use other kinds of cues to meaningfully use variable input to

form expectations that can constrain recognition. To be clear, these results suggest that experience with the different statistics of pattern sets, given a context cue that appropriately identifies the different sets, may subsequently shape the way listeners direct attention to stimulus properties highlighting a possible way in which top-down interactions (via cortical or corticofugal means) may reorganize perception.

Work by Magnuson and Nusbaum (2007) has shown that attention and expectations alone may influence the way listeners tune their perception to context, specifically through demonstrating that the performance costs typical of adjusting to talker variability, were modulated solely by expectations of hearing one or two talkers. In their study, listeners expecting to hear a single talker did not show performance costs in word recognition found when listeners were expecting to hear two talkers, even though the acoustic tokens were identical (for a similar finding in music, see Van Hedger, Heald, & Nusbaum, 2015). Related work by Magnuson, Yamada, and Nusbaum (1995) showed that this performance cost is still observed when shifting between two familiar talkers. This example of contextual tuning illustrates that top-down expectations, which occur outside of statistical learning, can fundamentally change how talker variability is accommodated in word recognition. This finding is conceptually similar to research by Niedzielski (1999), who demonstrated that vowel classification differed depending on whether listeners thought the vowels were produced by a speaker from Windsor, Ontario or Detroit, Michigan—cities that have different speech patterns but are close in distance. Similarly Johnson, Strand, and D'Imperio (1999) showed that the perception of "androgynous" speech was altered when presented with a male versus female face. Linking the domains of speech and music, recent work has demonstrated that the pitch of an identical acoustic signal is processed differently depending on whether the signal is interpreted as spoken or sung (Vanden Bosch der Nederlanden, Hannon, & Snyder, 2015).

Kleinschmidt and Jaeger (2015) has offered a computational approach on how expectations may influence the understanding of a signal. Specifically, they posit that until a listener has enough direct experience with a talker, a listener must supplement their observed input with their prior beliefs, which are brought online via expectations. However, this suggests that prior expectations are only necessary until enough direct experience is accrued. Another possibility, suggested by Magnuson and Nusbaum (2007), is that prior expectations are able to shape the interpretation of an acoustic pattern, regardless of accrued experience, as most acoustic patterns are nondeterministic (ambiguous). More specifically, Magnuson and Nusbaum show that

when a many-to-many mapping between acoustic cues and their meanings occurs this requires more cognitive, active processes, such as a change in expectation that may then direct attention to resolve the recognition uncertainty (cf. Heald & Nusbaum, 2014). Given these observations, auditory perception cannot be a purely passive, bottom-up process, as expectations about the interpretation of a signal can change the nature of how that signal is processed.

## 6.1 Beyond Statistical Learning

If recent experience and expectations shape perception, it follows that the ability to learn signal and pattern statistics is not solely sufficient to explain the empirical accounts of rapid perceptual plasticity within auditory object recognition. Changes in expectations appear to alter the priors the observer uses and may do so by violating the local statistics (prior context), such as when a talker changes. Further, there must be some processing in which one may resolve the inherent ambiguity or uncertainty that arises from the fact that the environment can be represented by multiple associations among cues. Listeners must determine the relevant associations weighing the given context under a given listening goal to direct attention appropriately (cf. Heald & Nusbaum, 2014). We argue that the uncertainty in weighing potential interpretations puts a particular emphasis on recent experience, as temporally local changes in contextual cues or changes in the variance of the input can signal to a listener that the underlying statistics have changed, altering how attention is distributed among the available cues to appropriately interpret a given signal. Importantly, this window of recent experience may also help solidify or alter listener expectations. In this way, recent experience may act as a buffer or an anchor against which the current signal and current representations are compared to previous experience. This would allow for rapid adaptability across a wide range of putatively stable representations, such as note category representations for AP possessors (Hedger et al., 2013), conceptualizations of auditory pitch (Dolscheid, Shayan, Majid, & Casasanto, 2013), and phonetic category representations (Evans & Iverson, 2004; Huang & Holt, 2012; Ladefoged & Broadbent, 1957; Liberman, Delattre, Gerstman, & Cooper, 1956; Mann, 1986).

It is important to consider exactly how plasticity engendered by a short-term window relates to a putatively stable, long-term representation of an auditory object. Given the behavioral and neural evidence previously discussed, it does not appear to be the case that auditory representations are static entities once established. Instead, auditory representations appear to

be heavily influenced by recent perceptual context and further, these changes persist in time after learning has concluded. However, this does not imply that there is no inherent stability built into the perceptual system. As previously discussed, perceptual categories in speech and music are not freestanding entities, but rather are a part of a constellation of categories that possess meaningful relationships with one another. Stability may exist through interconnections that exist in the category systems. Long-term neural mechanisms may work to remove rapid cortical changes that are inconsistent with the system, while in other cases, allow such changes to generalize to the rest of the system to achieve consistency.

## 7. CONCLUSIONS

People talk about the "skill of listening" but in the vernacular it means paying attention to what is said. But even in the scientific use we can become more specific along these lines. In essence, developing a particular auditory perceptual skill is the shaping of attention to the acoustic signal for purposes of recognizing the meaning of that signal. The process of shaping perceptual attention is under the control of the listener's goals, expectations, and task demands, on the one hand, but also influenced by the immediate context, long-term experience, and explicit knowledge on the other. These are not independent factors, but all have been demonstrated to have substantial effects on auditory skill development and continue to operate during the use of an auditory skill.

To have a particular auditory perceptual skill, such as AP is to have the capacity to perform an auditory task above the level of the average listener. This capacity depends on learning a knowledge structure of categories and relationships that can serve to organize perception, separating the acoustic chaff from the wheat in service of recognition. Further, an auditory skill depends on the knowledge about the way context can be used to constrain interpretations and experience in practicing this use. However, an auditory skill also depends on a set of more general cognitive mechanisms such as auditory working memory, long-term auditory sensory memory, and the control of sensory attention. These cognitive mechanisms likely differ substantially between individuals. Whether these individual differences reflect some biological endowment or whether they are themselves developed through experience, such as training and practice, is currently an open scientific question.

This cognitive substrate of memory and attention licenses the development of different auditory skills such as AP. But it reflects what must be the fundamental architecture of auditory perception. The lack of invariance problem, which entails the uncertainty of one acoustic pattern having multiple possible interpretations, suggests an active processing system may be necessary to consider and test among alternative possible interpretations of the signal (as hypotheses). Such processing necessarily involves (1) a working memory system during testing, (2) a long-term memory system for generating specific hypotheses, and (3) control of attention to specific auditory properties for testing the hypotheses (see Nusbaum & Schwab, 1986). Rather than assuming separate and independent learning and perceptual processing systems for auditory perception, we suggest that learning and plasticity are part of this active architecture that allows listeners to rapidly adjust to new contexts using attention and working memory.

For decades, speech research has treated speech perception as an innate, biologically programmed "faculty". This was conceptualized as a module of mind (Fodor, 1983) that is essentially encoded in genes to be expressed as a processing system or mental "organ." Much of the early research on the development of speech perception conformed to this framework in assuming that only during a critical period of development, a biological sensitive period, could any change in speech processing occur. However, research on the plasticity of speech perception throughout adulthood suggests that there is more of an effect of an early acquired "critical mass" (of experience) than any evidence for a critical period (Nusbaum & Goodman, 1994; Nusbaum & Lee, 1992). A critical mass of experience may be difficult to change with small doses of intervention experiences and, without a clear theory, difficult to test as a hypothesis.

Researchers still operate as though speech perception is a faculty rather than a skill, but we argue that it is important to systematically consider this question rather than assume an answer. There is substantial evidence for plasticity in adults, evidence for the operation of learning and consolidation mechanisms, and evidence of generalized effects of learning on performance that are robust over time. In this way, speech perception seems to operate in much the same way as AP although with clearly a higher level of practice than any other such listening capacity. By incorporating speech perception into a skill-oriented framework for general auditory processing, it is possible to move beyond extant dogmas and productively test new hypotheses that result from speech perception as depending on a common cognitive architecture that can be specialized through specific experiences

such as training and practice. While a number of researchers treat speech perception in a general auditory framework, this is typically not an auditory skill framework. The typical approach within a general auditory processing framework assumes a developed and fixed perceptual mechanism operating on a variable set of knowledge acquired through experience. A skills framework would integrate the adaptive process into the recognition processing, offering a way of accounting for a broader set of data and new insights into understanding perception and therapeutic approaches to remediating hearing loss, stroke, and other medical conditions that impair speech understanding.

## ACKNOWLEDGMENT

This work was supported by the Multidisciplinary University Research Initiatives (MURI) Program of the Office of Naval Research through grant, DOD/ONR N00014-13-1-0205.

## REFERENCES

Anantharaman, J. N., Krishnamurthy, A. K., & Feth, L. L. (1993). Intensity-weighted average of instantaneous frequency as a model for frequency discrimination. *The Journal of the Acoustical Society of America, 94*(2), 723–729.

Anderson, S. R., & Lightfoot, D. W. (2002). *The language organ: Linguistics as cognitive physiology*. Cambridge: Cambridge University Press.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology, 56*, 149–178.

Athos, E. A., Levinson, B., Kistler, A., Zemansky, J., Bostrom, A., Freimer, N., & Gitschier, J. (2007). Dichotomy and perceptual distortions in absolute pitch ability. *Proceedings of the National Academy of Sciences of the United States of America, 104*(37), 14795–14800.

Bachem, A. (1955). Absolute pitch. *The Journal of the Acoustical Society of America, 27*(6), 1180–1185.

Baharloo, S., Johnston, P. A., Service, S. K., Gitschier, J., & Freimer, N. B. (1998). Absolute pitch: An approach for identification of genetic and nongenetic components. *The American Journal of Human Genetics, 62*(2), 224–231.

Barrett, J., & Paus, T. (2002). Affect-induced changes in speech production. *Experimental Brain Research, 146*(4), 531–537.

Bauer, L. (1985). Tracing phonetic change in the received pronunciation of British English. *Journal of Phonetics, 13*, 61–81.

Best, C. T., McRoberts, G. W., & Sithole, N. M. (1988). Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. *Journal of Experimental Psychology, 14*, 345–360.

Bidelman, G. M., Hutka, S., & Moreno, S. (2013). Tone language speakers and musicians share enhanced perceptual and cognitive abilities for musical pitch: Evidence for bidirectionality between the domains of language and music. *PLoS One, 8*(4). http://dx.doi.org/10.1371/journal.pone.0060676.

Bidelman, G. M., Weiss, M. W., Moreno, S., & Alain, C. (2014). Coordinated plasticity in brainstem and auditory cortex contributes to enhanced categorical speech perception in musicians. *European Journal of Neuroscience, 40*(4), 2662–2673.

Billman, D., & Knutson, J. F. (1996). Unsupervised concept learning and value systematicity: A complex whole aids learning the parts. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 458–475.
Blumstein, S. E., & Stevens, K. N. (1981). Phonetic features and acoustic invariance in speech. *Cognition, 10*(1), 25–32.
de Boer, J., & Thornton, A. R. D. (2008). Neural correlates of perceptual learning in the auditory brainstem: Efferent activity predicts and reflects improvement at a speech-in-noise discrimination task. *Journal of Neuroscience, 28*(19), 4929–4937.
Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. I. (1997). Training Japanese listeners to identify English/r/and/l: IV. Some effects of perceptual learning on speech production. *The Journal of the Acoustical Society of America, 101*(4), 2299–2310.
Brady, P. T. (1970). Fixed-scale mechanism of absolute pitch. *The Journal of the Acoustical Society of America, 48*(4B), 883–887.
Brawn, T. P., Fenn, K. M., Nusbaum, H. C., & Margoliash, D. (2008). Consolidation of sensorimotor learning during sleep. *Learning & Memory, 15*(11), 815–819.
Brawn, T. P., Fenn, K. M., Nusbaum, H. C., & Margoliash, D. (2010). Consolidating the effects of waking and sleep on motor-sequence learning. *The Journal of Neuroscience, 30*, 13977–13982.
Brawn, T. P., Nusbaum, H. C., & Margoliash, D. (2010). Sleep-dependent consolidation of auditory discrimination learning in adult starlings. *The Journal of Neuroscience, 30*(2), 609–613.
Brawn, T., Nusbaum, H. C., & Margoliash, D. (2013). Sleep consolidation of interfering auditory memories in starlings. *Psychological Science*. http://dx.doi.org/10.1177/0956797612457391 (Published online 22 February 2013).
Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound*. MIT Press.
Bruner, J. S. (1973). *Beyond the information given: Studies in the psychology of knowing*. New York, NY: Norton.
Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking*. New York: Wiley.
Capleton, B. (2007). *Theory and practice of piano tuning*. Malvern, UK: Amarilli Books.
Carney, A. E., Widin, G. P., & Viemeister, N. F. (1977). Noncategorical perception of stop consonants differing in VOT. *The Journal of the Acoustical Society of America, 62*(4), 961–970.
Case, P., Tuller, B., Ding, M., & Kelso, J. A. (1995). Evaluation of a dynamical model of speech perception. *Perception & Psychophysics, 57*(7), 977–988.
Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4*, 55–81.
Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America, 25*(5), 975–979.
Chomsky, N. (1975). *Reflections on language*. New York: Pantheon.
Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York: Harper & Row.
Clarke, C., & Luce, P. (2005). Perceptual adaptation to speaker characteristics: VOT boundaries in stop voicing categorization. In *ISCA workshop on plasticity in speech perception*.
Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition, 108*(3), 804–809.
Cooper, W. (1974). Adaptation of phonetic feature analyzers for place of articulation. *The Journal of the Acoustical Society of America, 56*, 617–627.
Crozier, J. B. (1997). Absolute pitch: Practice makes perfect, the earlier the better. *Psychology of Music, 25*(2), 110–119.
Dahan, D., Magnuson, J. S., Tanenhaus, M. K., & Hogan, E. (2001). Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. *Language and Cognitive Processes, 16*, 507–534.
Deutsch, D. (2013). Absolute pitch. In D. Deutsch (Ed.), *The psychology of music* (3rd ed., pp. 141–182). San Diego: Elsevier.

Deutsch, D., Henthorn, T., & Dolson, M. (2004). Absolute pitch, speech, and tone language: Some experiments and a proposed framework. *Music Perception: An Interdisciplinary Journal, 21*(3), 339–356.

Dick, F., Lee, H. L., Nusbaum, H., & Price, C. (2011). Auditory-motor expertise alters "speech selectivity" in professional musicians and actors. *Cerebral Cortex, 21*, 938–948.

Dohn, A., Garza-Villarreal, E. A., Ribe, L. R., Wallentin, M., & Vuust, P. (2014). Musical activity tunes up absolute pitch ability. *Music Perception: An Interdisciplinary Journal, 31*(4), 359–371.

Dolscheid, S., Shayan, S., Majid, A., & Casasanto, D. (2013). The thickness of musical pitch: Psychophysical evidence for linguistic relativity. *Psychological Science, 24*(5), 613–621.

Doupe, A. J., & Kuhl, P. K. (2008). Birdsong and human speech: Common themes and mechanisms. In H. P. Zeigler, & P. Marler (Eds.), *Neuroscience of birdsong* (pp. 5–31). Cambridge, England: Cambridge University Press.

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 in speech processes* (pp. 360–385). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Evans, B. G., & Iverson, P. (2004). Vowel normalization for accent: An investigation of best exemplar locations in northern and southern British English sentences. *The Journal of the Acoustical Society of America, 115*(1), 352–361.

Evans, B. G., & Iverson, P. (2007). Plasticity in vowel perception and production: A study of accent change in young adults. *The Journal of the Acoustical Society of America, 121*(6), 3814–3826.

Fant, G. (1960). *Acoustic theory of speech production* (2nd ed.). The Hague, Netherlands: Mouton.

Feldman, N. H., Griffiths, T. L., Goldwater, S., & Morgan, J. L. (2013). A role for the developing lexicon in phonetic category acquisition. *Psychological Review, 120*(4), 751.

Fenn, K. M., Margoliash, D., & Nusbaum, H. C. (2013). Sleep restores loss of generalized but not rote learning of synthetic speech. *Cognition, 128*, 280–286.

Fenn, K. M., Nusbaum, H. C., & Margoliash, D. (2003). Consolidation during sleep of perceptual learning of spoken language. *Nature, 425*(6958), 614–616.

Fenn, K. M., Shintel, H., Atkins, A. S., Skipper, J. I., Bond, V. C., & Nusbaum, H. C. (2011). When less is heard than meets the ear: Change deafness in a telephone conversation. *The Quarterly Journal of Experimental Psychology, 64*(7), 1442–1456.

Fitch, R. H., Miller, S., & Tallal, P. (1997). Neurobiology of speech perception. *Annual Review of Neuroscience, 20*(1), 331–353.

Fodor, J. A. (1983). *Modularity of mind: An essay on faculty psychology*. MIT Press.

Fowler, C. A., & Galantucci, B. (2005). The relation of speech perception and speech production. In *The handbook of speech perception* (pp. 632–652).

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

Francis, A., & Nusbaum, H. C. (2009). Effects of intelligibility on working memory demand for speech perception. *Attention, Perception, & Psychophysics, 71*, 1360–1374.

Francis, A. L., Nusbaum, H. C., & Fenn, K. (2007). Effects of training on the acoustic–phonetic representation of synthetic speech. *Journal of Speech, Language, and Hearing Research, 50*(6), 1445–1465.

Friederici, A. D. (2012). The cortical language circuit: From auditory perception to sentence comprehension. *Trends in Cognitive Sciences, 16*(5), 262–268.

Gay, T. (1978). Effect of speaking rate on vowel formant movements. *The Journal of the Acoustical Society of America, 63*(1), 223–230.

Gebhart, A. L., Aslin, R. N., & Newport, E. L. (2009). Changing structures in midstream: Learning along the statistical garden path. *Cognitive Science, 33*(6), 1087–1116.
Gerstman, L. (1968). Classification of self-normalized vowels. *IEEE Transactions on Audio and Electroacoustics, 16*, 78–80.
Gervain, J., Vines, B. W., Chen, L. M., Seo, R. J., Hensch, T. K., & Werker, J. F. (2013). Valproate reopens critical-period learning of absolute pitch. *Frontiers in Systems Neuroscience, 7*, 102.
Giard, M., Collet, L., Bouchet, P., & Pernier, J. (1994). Auditory selective attention in the human cochlea. *Brain Research, 633*(1), 353–356.
Gockel, H., Moore, B. C., & Carlyon, R. P. (2001). Influence of rate of change of frequency on the overall pitch of frequency-modulated tones. *The Journal of the Acoustical Society of America, 109*(2), 701–712.
Goldstone, R. L., & Hendrickson, A. T. (2010). Categorical perception. *Wiley Interdisciplinary Reviews: Cognitive Science, 1*(1), 69–78.
Goldstone, R. L., Kersten, A., & Cavalho, P. F. (2012). Concepts and categorization. In A. F. Healy, & R. W. Proctor (Eds.), *Experimental psychology: Vol. 4. Comprehensive handbook of psychology* (pp. 607–630). New Jersey: Wiley.
Golestani, N., Price, C., & Scott, S. K. (2011). Born with an ear for dialects? Structural plasticity in the 'expert' phonetician brain. *Journal of Neuroscience, 31*(11), 4213–4220.
Gonzales, K., Gerken, L., & Gómez, R. L. (2015). Does hearing two dialects at different times help infants learn dialect-specific rules? *Cognition, 140*, 60–71.
Gow, D., McMurray, B., & Tanenhaus, M. K. (November 2003). Eye movements reveal the time course of multiple context effects in the perception of assimilated speech. In *Poster presented at the 44th annual meeting of the psychonomics society, Vancouver, Canada.*
Greenspan, S. L., Nusbaum, H. C., & Pisoni, D. B. (1988). Perceptual learning of synthetic speech produced by rule. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 421–433.
Gureckis, T. M., & Goldstone, R. L. (2008). The effect of the internal structure of categories on perception. In *Proceedings of the 30th annual conference of the cognitive science society* (pp. 1876–1881). Austin, TX: Cognitive Science Society.
Halpern, A. R., & Zatorre, R. J. (1979). Identification, discrimination, and selective adaptation of simultaneous musical intervals. *The Journal of the Acoustical Society of America, 65*(S1), S40.
Handel, S. (1989). *Listening: An introduction to the perception of auditory events*. Cambridge, MA: MIT Press.
Handel, S. (1993). The effect of tempo and tone duration on rhythm discrimination. *Perception & Psychophysics, 54*(3), 370–382.
Harnad, S. (1987). Psychophysical and cognitive aspects of categorical perception: A critical overview. In *Categorical perception: The groundwork of cognition* (pp. 1–52). Cambridge University Press.
Heald, S. L., & Nusbaum, H. C. (2014). Speech perception as an active cognitive process. *Frontiers in Systems Neuroscience, 8*(35), U12–U87.
Hedger, S. C., Heald, S. L., & Nusbaum, H. C. (2013). Absolute pitch may not be so absolute. *Psychological Science, 24*(8), 1496–1502.
Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience, 8*(5), 393–402.
Holt, L. L. (2005). Temporally nonadjacent nonlinguistic sounds affect speech categorization. *Psychological Science, 16*(4), 305–312.
Holt, L. L., Lotto, A. J., & Kluender, K. R. (2000). Neighboring spectral content influences vowel identification. *The Journal of the Acoustical Society of America, 108*(2), 710–722.

Holt, L. L., & Lotto, A. J. (2002). Behavioral examinations of the level of auditory processing of speech context effects. *Hearing Research, 167*(1–2), 156–169. http://dx.doi.org/10.1016/S0378-5955(02)00383-0.

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

Huang, J., & Holt, L. L. (2012). Listening for the norm: Adaptive coding in speech categorization. *Frontiers in Psychology, 3*, 1–6.

Ingvalson, E. M., Holt, L. L., & McClelland, J. L. (2012). Can native Japanese listeners learn to differentiate/r–l/on the basis of F3 onset frequency? *Bilingualism: Language and Cognition, 15*(02), 255–274.

Ingvalson, E. M., McClelland, J. L., & Holt, L. L. (2011). Predicting native English-like performance by native Japanese speakers. *Journal of Phonetics, 39*(4), 571–584.

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.

Johnson, K., Strand, E. A., & D'Imperio, M. (1999). Auditory–visual integration of talker gender in vowel perception. *Journal of Phonetics, 2727*(4), 359–384. http://dx.doi.org/10.1006/jpho.1999.0100.

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

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review, 122*(2), 148.

Kobayashi, T., Nisijma, K., Ehara, Y., Otsuka, K., & Kato, S. (2001). Pitch perception shift: A rare-side effect of carbamazepine. *Psychiatry and Clinical Neurosciences, 55*(4), 415–417.

Kolarik, A. J., Cirstea, S., Pardhan, S., & Moore, B. C. J. (April 1, 2014). A summary of research investigating echolocation abilities of blind and sighted humans. *Hearing Research, 310*, 60–68.

Krawczyk, D. C., Boggan, A. L., McClelland, M. M., & Bartlett, J. C. (2011). Brain organization of perception in chess experts. *Neuroscience Letters, 499*, 64–69.

Krumhansl, C. L., & Keil, F. C. (1982). Acquisition of the hierarchy of tonal functions in music. *Memory & Cognition, 10*(3), 243–251.

Krumhansl, C. L., & Kessler, E. J. (1982). Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys. *Psychological Review, 89*(4), 334.

Krumhansl, C. L., & Shepard, R. N. (1979). Quantification of the hierarchy of tonal functions within a diatonic context. *Journal of Experimental Psychology: Human Perception and Performance, 5*(4), 579.

Labov, W. (2001). *Principles of linguistic change- volume 2: Social factors*. Oxford: Blackwell.

Ladefoged, P. (2003). *Phonetic data analysis. An introduction to fieldwork and instrumental techniques*. Oxford: Blackwell.

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *The Journal of the Acoustical Society of America, 29*(1), 98–104.

Lancia, L., & Winter, B. (2013). The interaction between competition, learning, and habituation dynamics in speech perception. *Laboratory Phonology, 4*, 221–257.

Lennenberg, E. H. (1967). *Biological foundations of language*. New York: John Wiley & Sons.

Lesgold, A., Rubinson, H., Feltovich, P., Glaser, R., Klopfer, D., & Wang, Y. C. (1988). Expertise in a complex skill: Diagnosing x-ray pictures. In T. H. Michelene, R. Glaser, & M. J. Farr (Eds.), *The nature of expertise* (pp. 311–342). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc, xxxvi, 434 pp.

Levitin, D. J., & Rogers, S. E. (2005). Absolute pitch: Perception, coding, and controversies. *Trends in Cognitive Sciences, 9*(1), 26–33.

Liberman, A. M., Cooper, F. S., Harris, K. S., MacNeilage, P. F., & Studdert-Kennedy, M. (1967). Some observations on a model for speech perception. In W. Wathen-Dunn (Ed.), *Models for the perception of speech and visual form*. Cambridge: Mass: MIT Press.

Liberman, A. M., Delattre, P. C., Gerstman, L. J., & Cooper, F. S. (1956). Tempo of frequency change as a cue for distinguishing classes of speech sounds. *Journal of Experimental Psychology, 52*(2), 127.

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*(5), 358.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21*(1), 1–36. http://dx.doi.org/10.1016/0010-0277(85)90021-6.

Lieberman, P., Crelin, E. S., & Klatt, D. H. (1972). Phonetic ability and related anatomy of the newborn and adult human, Neanderthal man, and the chimpanzee. *American Anthropologist, 74*(3), 287–307.

Lim, S. J., & Holt, L. L. (2011). Learning foreign sounds in an alien world: Videogame training improves non-native speech categorization. *Cognitive Science, 35*(7), 1390–1405.

Lindblom, B. (1963). Spectrographic study of vowel reduction. *The Journal of the Acoustical Society of America, 35*, 1773–1781.

Lively, S. E., Pisoni, D. B., Yamada, R. A., Tokhura, Y., & Yamada, T. (1994). Training Japanese listeners to identify English/r/and/l/. Long-term retention of new phonetic categories. *The Journal of the Acoustical Society of America, 96*, 2076–2087.

Loui, P., & Wessel, D. (2008). Learning and liking an artificial musical system: Effects of set size and repeated exposure. *Musicae Scientiae, 12*(2), 207–230.

Lynch, M. P., & Eilers, R. E. (1991). Children's perception of native and nonnative musical scales. *Music Perception, 9*, 121–132.

Lynch, M. P., & Eilers, R. E. (1992). A study of perceptual development for musical tuning. *Perception & Psychophysics, 52*, 599–608.

Lynch, M. P., Eilers, R. E., Oller, D. K., & Urbano, R. C. (1990). Innateness, experience, and music perception. *Psychological Science, 1*, 272–276.

Magnuson, J. S., & Nusbaum, H. C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology: Human Perception and Performance, 33*(2), 391–409.

Magnuson, J. S., Yamada, R. A., & Nusbaum, H. C. (1995). The effects of talker variability and familiarity on mora perception and talker identification. In *ATR Human Information Processing Research Laboratories Technical Report TR-H-158*.

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

Mann, V. A. (1986). Distinguishing universal and language-dependent levels of speech perception: Evidence from Japanese listeners' perception of English "l" and "r". *Cognition, 24*(3), 169–196.

Mattingly, I. G., & Liberman, A. M. (1985). The motor theory of speech perception revised. *Cognition, 21*, 1–36. http://dx.doi.org/10.1016/0010-0277(85)90021-6.

Maye, J., & Gerken, L. (March 2000). Learning phonemes without minimal pairs. In *Proceedings of the 24th Annual Boston University Conference on Language Development* (Vol. 2, pp. 522–533).

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82*(3), B101–B111.

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

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review, 102*(3), 419.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746–748.

McLachlan, N., & Wilson, S. (2010). The central role of recognition in auditory perception: A neurobiological model. *Psychological Review, 117*(1), 175–196. http://dx.doi.org/10.1037/a0018063. pii:2009-25263-003.

McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review, 118*, 219–246.

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

Miller, J. L., & Baer, T. (1983). Some effects of speaking rate on the production of/b/and/w. *The Journal of the Acoustical Society of America, 73*(5), 1751–1755.

Mirman, D., McClelland, J. L., & Holt, L. L. (2006). An interactive Hebbian account of lexically guided tuning of speech perception. *Psychonomic Bulletin & Review, 13*(6), 958–965.

Monahan, C. B. (1993). Parallels between pitch and time and how they go together. In T. J. Tighe, & W. J. Dowling (Eds.), *Psychology and music: The understanding of melody and rhythm*. Hillsdale, NJ: Erlbaum.

Moon, S. J., & Lindblom, B. (1994). Interaction between duration, context, and speaking style in English stressed vowels. *The Journal of the Acoustical Society of America, 96*(1), 40–55.

Niedzielski, N. (1999). The effect of social information on the perception of sociolinguistic variables. *Journal of Language and Social Psychology, 18*(1), 62–85. http://dx.doi.org/10.1177/0261927X99018001005.

Nittrouer, S., & Lowenstein, J. H. (2007). Children's weighting strategies for word-final stop voicing are not explained by auditory sensitivities. *Journal of Speech, Language, and Hearing Research, 50*(1), 58–73.

Nittrouer, S., & Miller, M. E. (1997). Predicting developmental shifts in perceptual weighting schemes. *The Journal of the Acoustical Society of America, 101*(4), 2253–2266.

Nusbaum, H. C., & Goodman, J. C. (1994). Learning to hear speech as spoken language. In J. C. Goodman, & H. C. Nusbaum (Eds.), *The development of speech perception: The transition from speech sounds to spoken words*. Cambridge: MIT Press.

Nusbaum, H. C., & Lee, L. (1992). Learning to hear phonetic information. In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (Eds.), *Speech perception, production, and linguistic structure* (pp. 265–274). Tokyo: OHM Publishing Company.

Nusbaum, H. C., & Magnuson, J. S. (1997). Talker normalization: Phonetic constancy as a cognitive process. In K. A. Johnson, & J. W. Mullennix (Eds.), *Talker variability in speech processing*. New York, NY: Academic Press.

Nusbaum, H. C., & Schwab, E. C. (1986). The role of attention and active processing in speech perception. *Pattern Recognition by Humans and Machines, 1*, 113–157.

Parbery-Clark, A., Skoe, E., Lam, C., & Kraus, N. (2009). Musician enhancement for speech-in-noise. *Ear and Hearing, 30*(6), 653–661. http://dx.doi.org/10.1097/AUD.0b013e3181b412e9.

Parvizi, J. (2009). Corticocentric myopia: Old bias in new cognitive sciences. *Trends in Cognitive Sciences, 13*(8), 354–359.

Patel, A. D. (June 2011). Why would musical training benefit the neural encoding of speech? The OPERA hypothesis. *Frontiers in Psychology, 2*, 1–14. http://dx.doi.org/10.3389/fpsyg.2011.00142.

Pisoni, D. B., Aslin, R. N., Perey, A. J., & Hennessy, B. L. (1982). Some effects of laboratory training on identification and discrimination of voicing contrasts in stop consonants. *Journal of Experimental Psychology: Human Perception and Performance, 8*(2), 297.

Pisoni, D. B., & Lazarus, J. H. (1974). Categorical and noncategorical modes of speech perception along the voicing continuum. *The Journal of the Acoustical Society of America, 55*(2), 328–333.

Pisoni, D. B., & Tash, J. (1974). Reaction times to comparisons within and across phonetic categories. *Perception & Psychophysics, 15*(2), 285–290.

Potter, R. K., Kopp, G. A., & Green, H. C. (1947). *Visible speech*. New York: D. Van Nostrand Co.

Qian, T., Jaeger, T. F., & Aslin, R. N. (2012). Learning to represent a multi-context environment: More than detecting changes. *Frontiers in Psychology, 3*, 228.

Rauschecker, J. P., & Scott, S. K. (2009). Maps and streams in the auditory cortex: Nonhuman primates illuminate human speech processing. *Nature Neuroscience, 12*(6), 718–724.

Ross, D. A., Olson, I. R., & Gore, J. C. (2003). Absolute pitch does not depend on early musical training. *Annals of the New York Academy of Sciences, 999*(1), 522–526.

Rush, M. A. (1989). *An experimental investigation of the effectiveness of training on absolute pitch in adult musicians* (Doctoral dissertation). The Ohio State University.

Saffran, J. R., Reeck, K., Niebuhr, A., & Wilson, D. (2005). Changing the tune: The structure of the input affects infants' use of absolute and relative pitch. *Developmental Science, 8*(1), 1–7.

Schellenberg, E. G., & Trehub, S. E. (2003). Good pitch memory is widespread. *Psychological Science, 14*(3), 262–266.

Schwab, E. C., Nusbaum, H. C., & Pisoni, D. B. (1985). Effects of training on the perception of synthetic speech. *Human Factors, 27*, 395–408.

Sergeant, D. C., & Roche, S. (1973). Perceptual shifts in the auditory information processing of young children. *Psychology of Music, 1*, 39–48.

Shamma, S., & Fritz, J. (2014). Adaptive auditory computations. *Current Opinion in Neurobiology, 25*, 164–168. http://dx.doi.org/10.1016/j.conb.2014.01.011.

Siegel, J. A., & Siegel, W. (1977). Categorical perception of tonal intervais: Musicians can't tellsharp fromflat. *Perception & Psychophysics, 21*(5), 399–407.

Soley, G., & Hannon, E. E. (2010). Infants prefer the musical meter of their own culture: A cross-cultural comparison. *Developmental Psychology, 46*(1), 286.

Stevens, K. (1998). *Acoustic phonetics*. Cambridge, MA: MIT Press.

Strait, D. L., Kraus, N., Skoe, E., & Ashley, R. (2009). Musical experience promotes subcortical efficiency in processing emotional vocal sounds. *Annals of the New York Academy of Sciences, 1169*, 209–213. http://dx.doi.org/10.1111/j.1749-6632.2009.04864.x.

Strange, W., & Jenkins, J. J. (1978). Role of linguistic experience in the perception of speech. In *Perception and experience* (pp. 125–169). Springer US.

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.

Teki, S., Kumar, S., von Kriegstein, K., Stewart, L., Lyness, C. R., Moore, B. C., … Griffiths, T. D. (2012). Navigating the auditory scene: An expert role for the hippocampus. *Journal of Neuroscience, 32*(35), 12251–12257.

Terhardt, E. S., & Seewan, M. M. (1983). Aural key identification and its relationship to absolute pitch. *Music Perception, 1*, 63–83.

Theusch, E., Basu, A., & Gitschier, J. (2009). Genome-wide study of families with absolute pitch reveals linkage to 8q24. 21 and locus heterogeneity. *The American Journal of Human Genetics, 85*(1), 112–119.

Tuller, B., Case, P., Ding, M., & Kelso, J. A. S. (1994). The nonlinear dynamics of speech categorization. *Journal of Experimental Psychology: Human Perception and Performance, 20*(1), 3–16.

Vanden Bosch der Nederlanden, C. M., Hannon, E. E., & Snyder, J. S. (2015). Finding the music of speech: Musical knowledge influences pitch processing in speech. *Cognition, 143*, 135–140.

Van Hedger, S. C., Heald, S. L., Huang, A., Rutstein, B., & Nusbaum, H. C. (2016). Telling in-tune from out-of-tune: Widespread evidence for implicit absolute intonation. *Psychonomic Bulletin & Review*. http://dx.doi.org/10.3758/s13423-016-1099-1.

Van Hedger, S. C., Heald, S. L., Koch, R., & Nusbaum, H. C. (2015). Auditory working memory predicts individual differences in absolute pitch learning. *Cognition, 140*, 95–110.

Van Hedger, S. C., Heald, S. L. M., & Nusbaum, H. C. (2015). The effects of acoustic variability on absolute pitch categorization: Evidence for contextual tuning. *Journal of the Acoustical Society of America, 138*, 436–446.

Van Hedger, S. C., Heald, S. L., & Nusbaum, H. C. (2016). What the [bleep?]: Enhanced absolute pitch memory for a 1000 Hz sine tone. *Cognition, 154*, 139–150.

Ward, W. D., & Burns, E. M. (1982). Absolute pitch. In D. Deutsch (Ed.), *The psychology of music*. Academic Press.

Watson, C. I., Maclagan, M., & Harrington, J. (2000). Acoustic evidence for vowel change in New Zealand English. *Language Variation and Change, 12*(1), 51–68.

Weinberger, N. M. (2004). Specific long-term memory traces in primary auditory cortex. *Nature Reviews Neuroscience, 5*(4), 279–290.

Weinberger, N. M. (2015). New perspectives on the auditory cortex: Learning and memory. *Handbook of Clinical Neurology, 129*, 117–147.

Weiss, D. J., Gerfen, C., & Mitchel, A. D. (2009). Speech segmentation in a simulated bilingual environment: A challenge for statistical learning? *Language Learning and Development, 5*(1), 30–49.

Werker, J. F., & Polka, L. (1993). The ontogeny and developmental significance of language-specific phonetic perception. In B. de Boysson-Bardies, S. de Schonen, P. Jusczyk, P. MacNeilage, & J. Morton (Eds.), *Developmental neurocognition: Speech and face processing in the first year of life*. The Netherlands: Kluwer Academic Publishers B.V.

Werker, J. F., & Tees, R. C. (1983). Developmental changes across childhood in the perception of non-native speech sounds. *Canadian Journal of Psychology, 37*(2), 278.

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

Wilson, S. J., Lusher, D., Martin, C. L., Rayner, G., & McLachlan, N. (2012). *Music Perception: An Interdisciplinary Journal, 29*(3), 285–296.

Wong, P. C., Chan, A. H., Roy, A., & Margulis, E. H. (2011). The bimusical brain is not two monomusical brains in one: Evidence from musical affective processing. *Journal of Cognitive Neuroscience, 23*(12), 4082–4093.

Wong, P. C. M., & Perrachione, T. K. (2007). Learning pitch patterns in lexical identification by native English-speaking adults. *Applied Psycholinguistics, 28*(4), 565–585. http://dx.doi.org/10.1017/S0142716407070312.

Wong, P. C., Roy, A. K., & Margulis, E. H. (2009). Bimusicalism: The implicit dual enculturation of cognitive and affective systems. *Music Perception: An Interdisciplinary Journal, 27*(2), 81–88.

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

Wu, H., Ma, X., Zhang, L., Liu, Y., Zhang, Y., & Shu, H. (2015). Musical experience modulates categorical perception of lexical tones by native Chinese speakers. *Frontiers in Psychology, 6*(MAR), 436. http://dx.doi.org/10.3389/fpsyg.2015.00436.

Zhu, M., Chen, B., Galvin, J. J., & Fu, Q. J. (2011). Influence of pitch, timbre and timing cues on melodic contour identification with a competing masker (L). *The Journal of the Acoustical Society of America, 130*(6), 3562–3565.