Dimension-selective attention as a possible driver of dynamic, context-dependent re-weighting in speech processing

Lori L. Holt [1,2]
Adam T. Tierney [3,4]
Giada Guerra [3,4]
Aeron Laffere [3]
Frederic Dick [3,4,5]

[1] Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213
[2] Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, PA 15213
[3] Department of Psychological Sciences, Birkbeck College, University of London, London, WC1E 7HX
[4] Centre for Brain and Cognitive Development, Birkbeck College, London, WC1E 7HX
[5] Department of Experimental Psychology, University College London, London, WC1H 0AP

Corresponding Author:

Lori L. Holt
Professor, Department of Psychology
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
loriholt@cmu.edu

Highlights:

- Speech processing requires continuous reweighting across many acoustic dimensions
- This dynamic mapping may reflect the dynamics of auditory attentional mechanisms
- Animal neurobiological models can help to determine the putative role for attention
- We present results from a new attentional paradigm that ties together human and non-human research

Keywords:
speech perception
auditory selective attention
auditory learning
auditory plasticity
perceptual weight

Abstract

The contribution of acoustic dimensions to an auditory percept is dynamically adjusted and reweighted based on prior experience about how informative these dimensions are across the long-term and short-term environment. This is especially evident in speech perception, where listeners differentially weight information across multiple acoustic dimensions, and use this information selectively to update expectations about future sounds. The dynamic and selective adjustment of how acoustic input dimensions contribute to perception has made it tempting to conceive of this as a form of non-spatial auditory selective attention. Here, we review several human speech perception phenomena that might be consistent with auditory selective attention although, as of yet, the literature does not definitively support a mechanistic tie. We relate these human perceptual phenomena to illustrative nonhuman animal neurobiological findings that offer informative guideposts in how to test mechanistic connections. We next present a novel empirical approach that can serve as a methodological bridge from human research to animal neurobiological studies. Finally, we describe four preliminary results that demonstrate its utility in advancing understanding of human non-spatial dimension-based auditory selective attention.

1.1 Introduction

Understanding a friend's speech in a crowded cafe, tracking the quality of a sick child's breathing through a nursery monitor, and following the melody of a violin within an orchestra all require extracting the most informative dimensions for the task at hand from a complex mix of acoustic signals. Each of these scenarios can be conceived of as a variant of the classic 'cocktail party effect' (Cherry, 1953), whereby selective and sustained endogenous attention is directed to a particular sound source as it evolves in time. Experimental paradigms modeling these 'cocktail party' scenarios have often examined the contribution of acoustic dimensions conveying sound sources' spatial position in disambiguating target signals from irrelevant background scenes. This is appropriate given the importance of detecting and orienting to acoustic events in space. Yet, listeners must employ selective auditory attention even when spatial cues are unavailable (such as over the telephone or in listening to an orchestral recording over earbuds) or unreliable (as in listening within reverberant environments).

Indeed, even in the absence of spatial cues, listeners appear to *dynamically adjust and selectively weight* the contribution of multiple acoustic dimensions to an auditory percept based on prior experience about how informative these dimensions are - individually and in concert – to behavior. Speech perception provides an excellent case-in-point because individual speech sounds, phonemes like /b/ and /p/, are defined across multiple acoustic dimensions. Typically, no one acoustic dimension is necessary or sufficient to unambiguously signal a phoneme. Even more, factors like long-term and short-term acoustic distributional regularities and the adjacent sound context can impact the effectiveness of specific acoustic dimensions in signaling speech sounds. Even for a well-learned auditory skill like speech perception, the mapping of acoustics to percept remains flexible. The *dynamic* and *selective* adjustment of how robustly different acoustic dimensions contribute to speech recognition has made it tempting to conceive of this as a form of non-spatial auditory *selective attention*.

Our aim in this review is to explore this possibility. We first review some general background in the mapping of acoustics to speech. We next describe several speech perception phenomena to illustrate the highly dynamic nature of the mapping from acoustics to phoneme. We discuss how each of these phenomena resonates with a colloquial understanding of selective attention. But, we caution that it is important to recognize that *attention* may be best thought of as a cognitive placeholder that does not, in and of itself, point to a specific neurobiological mechanism (e.g., Cohen, Romero, Servan-Schreiber, & Farah, 1994).

We demonstrate this point by relating the human speech phenomena to illustrative neurobiological findings from nonhuman animal models. The neurobiological work offers informative guideposts in how we might make mechanistic connections back to human speech recognition. More specifically, it suggests that it would be unwise to be wholly satisfied with characterization of these speech phenomena as selective attention. There remains more explanatory work to be done, as a constellation of candidate neurobiological mechanisms exist that may support the dynamic nature of mapping acoustic input to behaviorally-relevant sounds like speech.

But, how might we make progress in advancing dimension-based selective attention from a cognitive placeholder to a real mechanistic understanding of human auditory behavior, including speech perception? After all, there remains a substantial distance between speech perception and approaches from nonhuman animal neurobiology. In the final section of the paper, we outline a novel empirical approach to human dimension-based selective attention that may serve as a methodological bridge between human and nonhuman animal literatures. By more closely aligning human experimental approaches with those that have been successful in nonhuman animal neurobiology, it may be possible to draw from the vital interpretive frameworks provided by neurobiological research. We briefly describe four empirical results to demonstrate the utility of this approach in advancing understanding of human dimension-based auditory selective attention with the ultimate aim of achieving a more nuanced model of the multiple mechanisms potentially at play in phenomena for which we use dimension-based selective attention as a cognitive placeholder.

1.2 Examples from Speech Processing

To situate our examples, it is useful to begin with some common background in speech acoustics. Consider the simple act of deciding whether your conversation partner has uttered /b/ or /p/, as in *beach* versus *peach*. If you know of one acoustic dimension related to speech communication, there is a very good chance it is voice onset time (VOT). The superstar of acoustic speech dimensions, VOT is defined in articulatory terms as the length of time between the release of a stop-consonant like /b/ or /p/ and the onset of voicing, the vibration of the vocal folds (Stevens, 2000). If you hold your fingers to your larynx while uttering *beach* and *peach* you will notice that the delay from when your lips release the consonant and your vocal folds being to vibrate is a bit longer for the 'voiceless' consonant /p/ than the 'voiced' consonant /b/. This has multiple acoustic outcomes (Lisker, 1986).

Chief among them, there is a greater temporal lag from the acoustic release burst associated with opening the mouth and the onset of a periodic acoustic signal originating from vibration of the vocal folds. Accordingly, it is rather easy to morph from voiced to voiceless consonants by parametrically lengthening this delay to create a series of speech sounds varying across VOT. At least in part as a result of this ease, the significance of VOT as an acoustic dimension in signaling voicing category distinctions like /b/-/p/, /d/-/t/, and /g/-/k/ has been studied across 100s, perhaps 1000s, of experiments spanning many languages (Abramson & Whalen, 2017).

Recent neurobiological research has very elegantly demonstrated that it is possible to recover a voicing code in human superior temporal cortex (Mesgarani, Cheung, Johnson, & Chang, 2014). These very exciting results can give the impression that we have discovered the neural code that supports categorization of an utterance as *beach* or *peach.* And, according to the classic 'textbook' understanding of the mapping from acoustics to phonetic categories, this would be true. But, contemporary research on the mapping of complex speech acoustics to phonetic categories makes clear that this textbook understanding is in need of an update. The situation is, in fact, more complex.

1.2.1 The Textbook Understanding of Speech Processing, With Contemporary Updates

To situate the examples that follow below, it is important to recognize that theory and research directed at human speech processing have long grappled with the issue of how the complex acoustic dimensions that vary across speech signals relate to phonemes, the linguistically distinct units of sound that differentiate meaning in a language such as /b/ versus /p/ in *beach* versus *peach*. Chances are very good that in opening an introductory perception or cognition textbook you will find a figure characterizing the *categorical perception* of speech (e.g., Wolfe, Kluender, Levi, Bartoshuk, & Herz, 2015). Perhaps the best-known phenomenon of speech perception (often demonstrated across the superstar dimension, VOT), categorical perception refers to the observation that listeners' identification of speech sounds does not vary gradually across incremental changes in an acoustic speech dimension. Instead, there is an abrupt shift across a restricted range of acoustic change. Endpoint stimuli are identified as one phoneme with near-ceiling performance that transitions sharply to near-ceiling identification of another phoneme. This categorical response appeared to be consistent with a mapping of speech acoustics to discrete, symbolic phonemic representations (Liberman, Harris, Hoffman, & Griffith, 1957). By this view, the subtle details of acoustic dimensions are unavailable once they are mapped discretely to a phoneme. Additionally, this view emphasized the mapping of individual dimensions to phonemes, as in VOT to /b/-/p/, and led to a long (and ultimately somewhat fruitless, (Blumstein & Stevens, 1985; Lisker, 1985) search for invariant acoustic cues that map to phonemes.

Contemporary research suggests that it is more productive to characterize speech as *categorized* rather than *categorical* (Holt & Lotto, 2010). The mapping looks much less discrete when speech perception is studied using more continuous methods. Listeners consistently rate some speech instances as 'better' exemplars of a speech category than others (e.g., Iverson & Kuhl, 1995; Utman, 1998; Utman, Blumstein, & Sullivan, 2001). Eyetracking and graded electroencephalographic (EEG) responses further reveal that fine-grained acoustic details of an utterance affect its categorization (e.g., Aydelott & Bates, 2004; Mcmurray, Aslin, Tanenhaus, Spivey, & Subik, 2008; Mcmurray, Tanenhaus, & Aslin, 2002; Utman, 1998; Utman et al., 2001; Utman, Blumstein, & Burton, 2000) and memory (e.g., Bradlow, Nygaard, & Pisoni, 1999; Goldinger, 1996; Nygaard, Sommers, & Pisoni, 1995). When we move away from binary responses typical of categorical perception tasks (did you hear *beach* or *peach*?), behavior suggests a rich internal structure in the representation of phonemes. Today, it is much more common to conceptualize the mapping from acoustics to *perceptual phonetic categories* that are neither discrete nor symbolic and instead possess rich internal structure that reflects the distributional characteristics of the experience that drove category learning (Holt & Lotto, 2010; Holt, Lotto, & Kluender, 2000).

By this more contemporary perspective, there is no need to search for an invariant acoustic cue uniquely differentiating a particular phonemic contrast. Instead, phonetic categories can be considered to reside in a highly multidimensional perceptual space that maps the acoustic complexity of speech across multiple dimensions. Correspondingly, there is increasing appreciation that it is critical to consider *auditory* rather than *acoustic* dimensions (like the manipulation leading to a step-wise VOT stimulus series), in appreciation of the important transformations in early auditory processing that warp the perceptual space conveyed by acoustic dimensions. (It is a somewhat ironic aside that some of the best evidence for nonlinearities in the mapping of acoustic speech dimensions to auditory dimensions comes from the superstar of acoustic dimensions driving so much research, VOT; Holt, Lotto, & Diehl, 2004).

Finally, and most critically for the present review, contemporary research is rich with examples that even these auditory dimensions do not stably map to phonetic categories. Instead, the mapping is a much more dynamic process, indicating that the textbook understanding of an invariant, or even consistent, mapping from acoustics to speech is in need of an update, and that we should take care in concluding that neural decoding in human cortex conveys the complete mechanistic basis of human speech recognition. In the next sections, we consider some specific examples, and how they might be related to short-term plasticity and attentional modulation as well as longer-term learning about the informational environment.

1.2.1 Perceptual Weight in Speech Categorization

As central (and well-studied) as VOT is in signaling voicing categories in speech, there is in fact a constellation of as many as 16 acoustic dimensions that co-vary with English /b/-/p/ category membership (Lisker, 1986). For example, in addition to VOT, the fundamental frequency (F0, associated with voice pitch) of the following vowel co-varies with /b/-/p/ category membership. When we utter *peach*, the following vowel tends to have a somewhat higher F0 than when we utter *beach*. Correspondingly, listeners rely upon both dimensions in phonetic categorization. When VOT is acoustically ambiguous, for example, utterances with higher F0 frequencies are categorized as /p/ whereas those with lower F0 frequencies are categorized as /b/. Critically, listeners do not rely upon these dimensions in equal measure. Rather, behavioral (Francis, Kaganovich, & Driscoll-Huber, 2008; Holt & Lotto, 2006; Iverson & Kuhl, 1995), neural (Scharinger, Herrmann, Nierhaus, & Obleser, 2014), and developmental (Nittrouer, Lowenstein, & Packer, 2009; Wellmann, Holzgrefe, Truckenbrodt, Wartenburger, & Höhle, 2012) evidence indicates that listeners *perceptually weight* acoustic dimensions, with some dimensions contributing more robustly to perception than others.

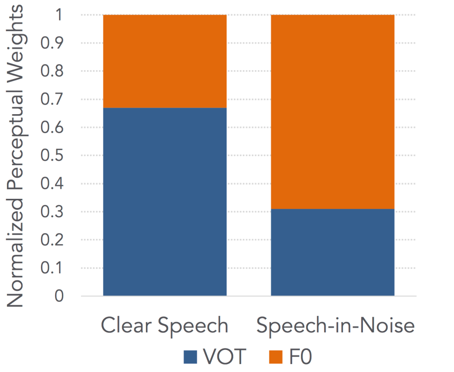

Figure 1. Listeners categorized a grid of speech sounds varying across VOT and F0 as /b/ or /p/. The relationship of each acoustic dimension to categorization responses was calculated using regression, with normalized regression weights providing a measure of perceptual weight. In Clear Speech, VOT is most diagnostic of /b/-/p/ categorization. But, in white noise, F0 dominates categorization of the same speech sounds.

As we will discuss in more detail below, prior research demonstrates that perceptual weights are a function of the long-term statistics of the input (Francis, Ciocca, Wong, Leung, & Chu, 2006; Holt & Lotto, 2006; Toscano & Mcmurray, 2010), they are specific to one's native language (Iverson et al., 2003; Kondaurova & Francis, 2008; 2010), and they emerge over a rather protracted developmental timeline extending at least into late childhood (Idemaru & Holt, 2013). For present purposes, the point is simply that although multiple auditory dimensions signal phonetic category identity, their contributions are not equivalent. Some dimensions carry greater *perceptual weight* than others. Figure 1 illustrates this for /b/ versus /p/ categorization, where perceptual weight is calculated as the normalized regression coefficient related to /b/-/p/ categorization by native-English listeners across a grid of speech syllables varying parametrically in VOT and F0. For clear speech (Figure 1, Clear Speech), both acoustic dimensions inform /b/-/p/ categorization, but VOT carries greater perceptual weight. It better predicts how listeners will categorize a sound than F0.

Many investigators have noted the potential for selective attention to play a role in perceptual weighting of acoustic dimensions in speech processing, in the sense that selective attention appears to be consistent with the demand to direct processing to diagnostic dimensions in the presence of the rich acoustic information available across multiple input dimensions (Francis & Nusbaum, 2002; Gordon, Eberhardt, & Rueckl, 1993; Heald & Nusbaum, 2014). This conceptualization suggests a potentially dynamic process, one not rigidly wired at the culmination of development. Figure 1 provides a simple example that underscores this point. When the same /b/-/p/ stimuli used to calculate perceptual weights across VOT and F0 in clear speech (Figure 1, Clear Speech) are presented in modest levels of white noise, *categorization is more dependent on F0 and less dependent on VOT* (Figure 1, Speech-in-Noise). We can speculate that the F0 dimension may be more robust to noise and therefore a more valuable indicator of phonetic category identity under noisy conditions. Whether accomplished by processes consistent with 'attention' or through other means, the shift in perceptual weights apparent in Figure 1 makes it clear that listeners rely on different acoustic dimensions in speech categorization in adverse versus clear listening environments. Perceptual weights are labile. VOT is the star dimension focused upon in textbook examples, but it only shines under the right circumstances.

This compelling example is not a mere parlor trick of perception. It informs us that discovering a neural code for VOT, or any other acoustic dimension that informs speech perception, takes us only part of the way to understanding how the auditory system maps complex acoustics to objects for recognition. A complete account will require a deeper understanding of how acoustic dimensions are weighted in auditory recognition because the very dimensions that inform auditory object recognition are not fixed. Rather, listeners flexibly shift reliance on acoustic dimensions according to the demands of the listening environment.

1.2.2 Perceptual Learning Over the Long-term

Although speech category learning gets underway even before an infant's first birthday (Conboy & Kuhl, 2011; Kuhl, 2004) there is a long developmental tail that extends into at least early adolescence in establishing the perceptual weights of acoustic dimensions (Zevin, 2012). For example, the onset frequency of the third formant (F3) is the acoustic dimension that best predicts English /r/-/l/ category membership in the acoustics of native talkers' speech (Iverson et al., 2003), although the onset frequency of the second formant (F2) is also diagnostic to a lesser degree. Among mature listeners, these distributional regularities of English speech input are reflected in /r/-/l/ perceptual categorization. Adult listeners rely more on F3 onset frequency, giving it greater perceptual weight, than F2 onset frequency. But, although typically-developing native-English-learning children ages 4.5, 5.5, and even 8.5 years use the dominant, F3, input dimension to accurately categorize English /r/-/l/, they fail to rely upon F2 as a secondary diagnostic dimension like adults (Idemaru & Holt, 2013). This indicates a much longer developmental course for phonetic category development than is typically appreciated (Zevin, 2012).

Moreover, this pattern of development underscores the fact that perceptual weighting arises, at least in part, from dimensions' informativeness in signaling category identity (Holt & Lotto, 2006; McMurray & Jongman, 2011). The distributional regularities of speech input shape perceptual weight of input dimensions. Efficient categorizers ultimately learn to perceptually weight the multiple dimensions that define speech categories in relation to the dimensions' reliability, or informativeness, in signaling a category (Holt & Lotto, 2006). Additionally, perceptual weight is likely to be impacted additionally by basic auditory representation (some dimensions are more robustly encoded by the auditory system than others) and even task (dimensions heavily weighted for phonetic categorization may be much less relied upon in identifying a talker). Either of these latter factors may play a role, as well, in the perceptual weight shifts evident in Figure 1. Accordingly, some accounts have emphasized learning to attend selectively to diagnostic dimensions as an important component of phonetic category learning (Heald & Nusbaum, 2014; Kondaurova & Francis, 2010). (Attention-based approaches to category learning and warping have long been used in vision research, e.g., Kruschke, Kappenman, & Hetrick, 2005; Nosofsky, 1986).

If efficient speech comprehension heavily relies on the process of learning and maintaining representations of higher-dimensional auditory categories, then one might expect that localized patterns of neural activation related to speech processing (or, indeed, seemingly *selectively* related to speech) might also be associated with the emergence of new nonspeech auditory categories. As a test of this hypothesis, Leech, Holt, Devlin, and Dick (2009) asked whether video game play that drives implicit nonspeech auditory categorization (Wade & Holt, 2005) would change responses to the trained nonspeech sounds in canonical 'speech-selective' cortex. They found that subjects' ability to categorize these novel sounds after training was significantly correlated with pre-to-post training change in fMRI activation in a part of the left posterior superior temporal sulcus that has been implicated in speech processing and phonemic categorization (Dehaene-Lambertz et al., 2005; Desai, Liebenthal, Waldron, & Binder, 2008).

Studying how adult listeners learn artificial, nonspeech auditory categories has informed thinking because it is difficult to gain an experimental foothold in understanding how learning operates over long-term speech category development since direct manipulation of children's speech input is infeasible. As adults learn novel, artificial auditory categories they must learn to pull together auditory dimensions according to training-related task demands and feedback to form new representations. Learning new auditory categories that generalize to novel instances changes the partitioning of auditory representational space (Liu & Holt, 2011) in a manner that can be described as 'warping' or exaggeration of the mapping of input to emphasize categorization-relevant acoustic dimensions, or alternatively as plasticity that directs selective attention to these dimensions. Indeed, provided with appropriate training, listeners can learn to attend selectively to acoustic dimensions that do not typically contribute to native-language speech perception (Kondaurova & Francis, 2010), and this impacts electrophysiological response to speech (Ylinen et al., 2010). The observations potentially argue for construing perceptual learning of auditory (including speech) categories over the long-term as involving allocation of selective attention to the most diagnostic acoustic dimensions.

1.2.3 Perceptual Learning Across the Short-term

The challenge for human communication is even greater because we often encounter talkers with foreign accents, dialects, or speech idiosyncrasies. In these cases, the speech input is 'warped' relative to the pattern of experience that established the long-term perceptual weights, with the potential for acoustic input dimensions to relate differently to phonetic categories.

This challenge is met by a highly flexible perceptual system capable of tracking short-term input regularities and dynamically adapting reliance upon specific acoustic dimensions. Recall, from above, that both VOT and F0 contribute to English /b/-/p/ categorization, with VOT more diagnostic than F0 in clear speech. These dimensions are also correlated in English speech productions. Stimuli with longer VOT, typical of /p/, also tend to have higher F0 frequencies whereas those with shorter VOT, typical of /b/, are associated with lower F0 frequencies. Mature listeners are sensitive to this relationship. When VOT is acoustically ambiguous and insufficient to reliably signal /b/ versus /p/, listeners label higher-F0 stimuli as /p/ and lower-F0 stimuli as /b/.

It is possible to model real-world encounters with foreign-accented speech by manipulating the short-term distribution of speech experience across an experiment. For example, Idemaru and Holt (2011) had listeners categorize speech sounds as *beer* or *pier* with a button press. The majority of trials were 'exposure' trials in which the speech exemplars were unambiguously signaled by the dominant perceptual dimension (VOT) and the secondary dimension (F0) was correlated in the canonical manner (Figure 2a). This conveyed a short-term distribution of speech experience that aligned with the long-term regularities of English. Without a change in task or other overt cues, Idemaru and Holt introduced a subtle 'artificial accent' by shifting the distribution statistics between VOT and F0 acoustic dimensions. In the Reverse Block shown in Figure 2a, VOT continued to unambiguously signal category membership across the exposure trials. But the secondary, F0, dimension was now associated with the VOT dimension in manner counter to long-term English experience. In the Reverse block, shorter VOTs were associated with higher F0s and longer VOTs were associated with lower F0s. This produced an artificial accent that changed the short-term input regularities in a manner akin to some natural foreign-language accents (Kim & Lotto, 2002).

Idemaru and Holt (2011; 2014) assessed the impact of this shift in short-term regularities across speech input dimensions by observing overt categorization decisions across infrequent 'test' trials intermixed with the exposure trials (orange diamond and purple square symbols, Figure 2). For these two stimuli, the dominant dimension, VOT, was acoustically ambiguous and therefore provided poor information about phonetic category identity. As a result, the two test stimuli differed only in the secondary, F0, dimension. As such, categorization of the test stimuli provided a metric of the perceptual weight of F0 – how diagnostic the F0 dimension is in signaling /b/-/p/ categorization. If listeners rely exclusively on VOT, categorization of the two test stimuli will not differ. But, to the extent that F0 informs category membership, categorization of the two test stimuli will differ. The magnitude of this difference provides a measure of the perceptual weight of F0 in /b/-/p/ categorization as a function of the short-term speech regularities manipulated across blocks via the exposure stimuli.

It is important to point out that there was no explicit training or feedback. Listeners were not informed about the shift in input from the Canonical to the Reverse block, the talker remained constant, the test trials were not differentiated from the exposure trials, and the task was always simply to identify the word. The range of dimension variability experienced across blocks fell within that experienced for the talker, and it went largely unnoticed by participants. Moreover, the range of values experienced across dimensions was constant across the experiment (only the relationship changed), so variability across a dimension was not a factor.

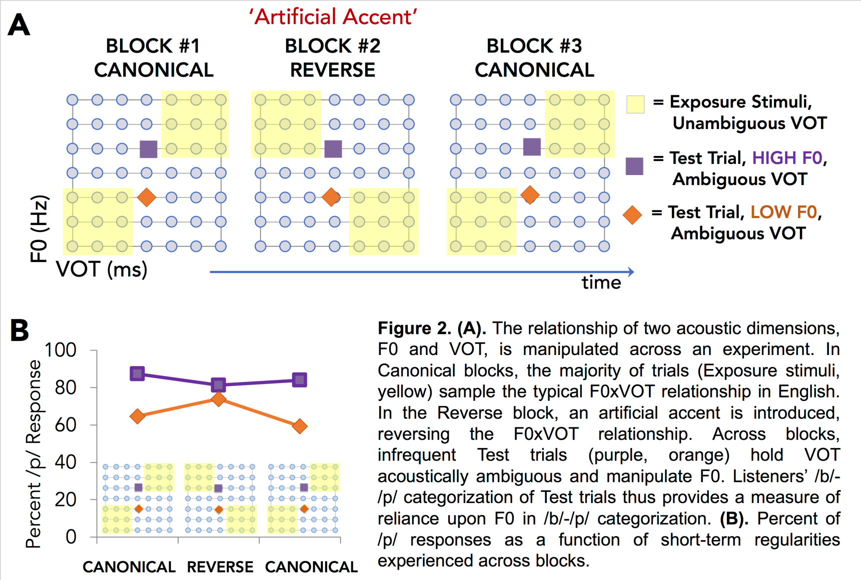

Figure 2. (A). The relationship of two acoustic dimensions, F0 and VOT, is manipulated across an experiment. In Canonical blocks, the majority of trials (Exposure stimuli, yellow) sample the typical F0xVOT relationship in English. In the Reverse block, an artificial accent is introduced, reversing the F0xVOT relationship. Across blocks, infrequent Test trials (purple, orange) hold VOT acoustically ambiguous and manipulate F0. Listeners' /b/-/p/ categorization of Test trials thus provides a measure of reliance upon F0 in /b/-/p/ categorization. **(B).** Percent of /p/ responses as a function of short-term regularities experienced across blocks.

Figure 2b illustrates the impact of short-term regularities across speech input dimensions on listeners' reliance on F0 to signal /b/-/p/ categorization. When short-term input aligns with native-language experience (Canonical Blocks), listeners relied upon the secondary, F0, dimension to make category decisions. It provided information across which to differentially categorize the test stimuli as /b/ (Low F0) and /p/ (High F0). This is simply a reflection of the fact that secondary dimensions informed categorization, albeit less robustly than the dominant dimensions.

However, upon introduction of the artificial accent in the Reverse block – a short-term change in input regularities – reliance upon F0 to inform /b/-/p/ categorization was rapidly down-weighted (Figure 2b). When the short-term input shifted such that F0 mapped to VOT in a manner inconsistent with long-term speech input regularities, the F0 dimension was no longer as informative to /b/-/p/ categorization. Note that the down-weighting of F0 in informing speech categorization does not appear to reflect a wholesale shift in attention away from the secondary, F0, dimension; listeners rapidly resumed reliance F0 in a final Canonical Block, indicating that they continued to track F0 in the input. Rather, the data suggest a continuous, dynamic modulation of input dimensions' contributions to phonetic categorization, adjusted to accommodate short-term input regularities.

On the face of it, this dynamic adjustment in the weighting functions with which auditory dimensions map to phonetic categories could be described as consistent with rapid adjustments in selective attention to auditory dimensions. However, our currently incomplete understanding of human auditory selective attention makes it difficult to determine definitively whether this is a viable model. The conundrum for advancing a mechanistic understanding of whether selective attention plays a role is that we do not yet have a rich body of evidence regarding the boundaries and constraints of dimension-based auditory selective attention to definitively determine whether it is playing a role. Even so, these behavioral results highlight the inherently dynamic nature of the mapping from acoustic speech input to behaviorally-relevant categories like phonemes and words.

1.2.4 The Impact of Context in Speech Categorization

Even quite subtle changes in distributions of sound experienced across a single input dimension can influence how an acoustic input dimension factors into phonetic categorization. To illustrate, consider categorization of speech syllables that vary perceptually from /ga/ to /da/. In English speech productions, these syllables are best differentiated by the third formant onset frequency (F3). Accordingly, F3 onset frequency carries a strong perceptual weight in /ga/-/da/ categorization. As is typical in identification responses across a series of stop consonants like /g/ and /d/, there is a rather steep slope in the transition from identifying lower F3 onsets as /ga/ to identifying higher F3 onsets as /da/. This steep identification function (consistent with what is traditionally interpreted as categorical perception) invites the inference that a specific range of lower-frequency range of F3 onset frequencies map to /g/ and another specific, higher-frequency, range of F3 onset frequencies map to /d/(e.g., Lotto & Kluender, 1998).

However, perception of isolated syllables only tells part of the story. Consider what happens when a simple sentence precedes the /ga/-/da/ syllables. As first demonstrated long ago (Ladefoged & Broadbent, 1957), preceding context can have a substantial influence on speech categorization. A contemporary example demonstrates this for /ga/-/da/ (Laing, Liu, Lotto, & Holt, 2012). In this study, a precursor phrase (*Please say what this word is*) preceded the /ga/-/da/ syllables varying in F3 onset frequency and listeners simply categorized the final /ga/-/da/ syllable. In one block of trials, the precursor phrase was manipulated to subtly emphasize somewhat higher frequencies in a frequency band in the range of /ga/-/da/ F3 onset frequencies. In another block of trials, the same precursor phrase emphasized lower frequencies in the same band. Said another way, the phrases subtly differed in the long-term average spectrum of the preceding speech. On each trial, listeners simply reported whether they heard *ga* or *da* in the context of one of the two precursor phrases.

The results demonstrate that phonetic perception is influenced by the long-term average spectrum of precursor sounds. In the context of a precursor sentence with exaggerated higher F3-band frequencies, the mapping of F3 onset frequency shifts to result in more /ga/ categorizations. In the context of exaggerated lower F3-band frequencies in the precursor phrase, the same speech target syllables are more often categorized as /da/. Thus, a precursor can have a substantial effect on how the F3 onset frequency input dimension maps to phonetic categories. This may provide a means by which the system accomplishes talker normalization (Assgari & Stilp, 2015; Huang & Holt, 2012; Ladefoged & Broadbent, 1957).

Perhaps more surprising, precursor contexts across which spectrally-biased long-term average spectra emerge need not be speech to impact phonetic categorization (Holt, 2005; 2006b). When a series of pure tones sampling the higher versus lower F3-band frequencies precedes /ga/-/da/ syllables, phonetic categorization is also shifted (Holt, 2005; 2006b; Laing et al., 2012). In the context of a sequence of higher-frequency tones, categorization shifts to /ga/. The same speech syllables are more often reported as /da/ when preceded by a lower-frequency sequences of tones. Here, as in the case of speech precursor sentences, the direction of the influence of context is *spectrally contrastive*. Higher-frequency precursors lead subsequent acoustic information to be more often mapped to the category characterized by lower F3 onset frequencies, /ga/, and vice versa. This pattern of spectral contrast has been observed across many speech categories (Lotto & Holt, 2006), evoked by precursor sentences (Assgari & Stilp, 2015; Huang & Holt, 2012; Laing et al., 2012), single syllables (Huang & Holt, 2012; Lotto & Kluender, 1998), and across nonspeech contexts varying from tones to notched noise (Holt, 2005; 2006a; 2006b; Lotto & Kluender, 1998). Across these studies, the findings are consistent in revealing that the mapping of an input dimension to an auditory representation, here a phonetic category, is not fixed. Rather, the auditory system appears to track the distribution of spectral energy evolving across the long-term average spectrum of incoming speech and the mapping of subsequent acoustic information is relative to, and contrastively with, the distribution of acoustic information experienced in prior context.

A rather spectacular non-speech demonstration of such acoustic context effects was recently reported by Chambers et al. (2017). The authors took advantage of a classic auditory stimulus, a Shepherd tone, made of octave-separated pure tones distributed across all audible frequencies. If one sequentially presents two Shepherd tones separated by a base frequency of 6 semitones (a musical 'tritone'), the average listener is equally likely to hear a pitch shift going up or going down (although individual listeners can have quite strong and stable bias for hearing an up or down shift). However, when such a Shepherd tone pair is preceded by an acoustic context, subjects' perception of the direction of this ambiguous pitch shift could be quasi-deterministically manipulated, whereby the contiguity of the separate frequency elements of the context tones with the two test tones decides the percept. This result shows that a basic auditory perceptual decision - the direction of a local pitch contour - is strongly driven by active integration with prior acoustic information.

Although speculative, these demonstrations from human behavior may be consistent with accounts of auditory selective attention that emphasize optimization of auditory cortical filters for task performance and for enhancing selectivity to task-relevant information via contrast enhancement (e.g., Fritz et al., 2007; Jääskeläinen et al., 2007; 2011).

1.2.5 Summary

We began with something simple: how might the auditory system recognize a spoken word *beach* from *peach.* The textbook answer to this question is straightforward and has influenced our approach to evaluating neurobiological evidence for speech recognition. The traditional understanding is that the system recognizes a diagnostic auditory cue, like VOT, which maps to a phonetic category. By this view, it is quite natural to conceive of the mapping from input to auditory object, in the cases above phonetic categories, as examples of sensory ‘encoding’ to relatively stable features or dimensions. Thus, when we see patterns of activation in the brain that correspond closely with acoustic dimensions we know to be significant in signaling a particular phonetic category (Mesgarani et al., 2014) it is tempting to conclude that we have cracked the speech code.

The phenomena reviewed above collectively illustrate the dynamic nature of the mapping of auditory dimensions to behaviorally-relevant representations and actions. They reveal the need for a less static perspective on how input is mapped to behaviorally-relevant auditory representations and highlight that the dividing lines between perception, attention and learning are likely to be quite blurry -- if they exist at all. The very mapping of acoustics to auditory dimensions and objects is dependent upon an organism’s prior history of experience, the short-term experience evolving in the local input, and statistical relationships relating the present sound exemplar to those experienced previously.

These effects are well illustrated by perception of speech, but they are not exclusive to speech. Humans and other mammals are very sensitive to changes in the salience, task-relevance, and composition of the acoustic dimensions of complex and ecologically important sounds (Holt & Lotto, 2006; Leech et al., 2009b; Leech, Gygi, Aydelott, & Dick, 2009a; Shamma & Fritz, 2014). Indeed, listeners appear to be able to shift attention across multiple simultaneously-present acoustic dimensions to home in on the ones that are diagnostic in guiding behavior (Henry, Herrmann, & Obleser, 2015; Herrmann, Henry, & Obleser, 2013a; Herrmann, Henry, Scharinger, & Obleser, 2013b; Herrmann, Schlichting, & Obleser, 2013c; Idemaru & Holt, 2011). As we noted above, this non-spatial *dimension-based auditory attention* has received rather little empirical study in human auditory cognitive neuroscience. Thus, although there are suggestive connections of the phenomena reviewed above with attention, and although selective attention has been evoked as a potential contributor to the highly dynamic mapping of input in speech perception, it remains the case that explanatory power is compromised without a more solid mechanistic understanding of non-spatial dimension-based attention in auditory processing.

To illustrate this point, we next briefly review several illustrative nonhuman animal studies of auditory processing that provide potentially useful guideposts in making headway on this issue. Collectively, they illustrate that although 'attention' is useful as a placeholder, the phenomena to which it is directed are unlikely to arise from a single mechanism, or across a constant level of representation or timescale. These illustrative examples also offer direction in considering how to build new human paradigms that can connect better with open questions about whether auditory selective attention – and plasticity associated with it – play a substantive role in the dynamic mapping of acoustic dimensions to speech reviewed above.

1.3 Nonhuman Animal Neurobiological Studies

The neural mechanisms of active listening (in contrast to passive 'hearing') have been increasingly the focus of research in understanding the hierarchy of cortical areas identified in the mammalian auditory system (Hackett, 2011; Morillon, Hackett, Kajikawa, & Schroeder, 2015). Nonhuman mammal studies have shown that behavioral manipulation of attentional systems can modulate, and even alter, the topography of tonotopic maps (Bieszczad & Weinberger, 2010; Weinberger, 2007), and that this modulation is important for learning. Moreover, recent studies demonstrate that neuronal receptive fields in regions along the cortical hierarchy are modulated in response to the behavioral relevance of auditory dimensions (Atiani et al., 2014; David, Fritz, & Shamma, 2012; Shamma & Fritz, 2014; Winkowski, Bandyopadhyay, Shamma, & Kanold, 2013; Yin, Fritz, & Shamma, 2014). We briefly (and selectively) review a few illustrative examples that may be useful in connecting animal neurobiological frameworks with phenomena we reviewed above.

Perceptual Weighting. Multiple species exhibit sensitivity to combinations of acoustic input dimensions (e.g., Atencio, Sharpee, & Schreiner, 2008), making it tempting connect these literatures with the multidimensional nature of speech categories and the dynamic nature by which input dimensions are mapped to behaviorally-relevant categories. Indeed, a recent study demonstrates that plasticity in adult rodents impacts auditory sensitivity to combinations of acoustic input dimensions (Shepard, Lin, Zhao, Chong, & Liu, 2015). Using single-unit recordings and electrophysiological mapping in an adult mouse model, Shepard et al. demonstrate that auditory core cortical activity differentiates species-specific vocal categories. Moreover, a distinct set of core auditory cortical (putative pyramidal) neurons develop increased sensitivity to specific *combinations* of auditory dimensions in newly-acquired vocalization categories. At a population level, this plasticity reflects the differential weighting across acoustic input dimensions associated with behaviorally-relevant vocalization categories. Inasmuch as the auditory representation of behaviorally-relevant acquired categories comes to reflect the combinations of acoustic dimensions signaling the categories with differential perceptual weights, this model may provide a productive framework for discovering neurobiological bases of perceptual weighting in the auditory system, how these weightings emerge with experience, and how they might be dynamically re-weighted by short-term regularities in the input, as observed for speech (Idemaru & Holt, 2011).

Dimension-based Attention to Acoustic Frequency. In both human and non-human animals, auditory attention is often studied by comparing neuronal responses when the animal is engaged in a demanding behavioral task (Tsunada, Liu, Gold, & Cohen, 2015) or specific readiness state (Carcea, Insanally, & Froemke, 2017), versus passive listening or less constrained activity. This makes it difficult to disambiguate effects of task, overall arousal, motor activity, and cross-modal attentional allocation from the effects of attention *within* a given dimension -- for instance, attending to a higher or lower frequency band. Recently, Schwartz and David (2017) created a novel rodent experimental paradigm to direct attention to one of two frequency bands. Ferrets were simultaneously presented with 2 streams of dynamically filtered narrowband noise, with each band presented at a different spatial location to enable behavior. Distributed over multiple trials, one band contained embedded higher-SNR 'cue' tones at the band's center frequency (serving to draw the ferret's attention to that band), with both bands containing embedded 'probe' tones at lower SNRs, which served as target and foil stimuli. With training, ferrets very accurately detected target and ignored foil tones. But, in contrast to what might have been expected from work in vision (where attending to one part of retinotopic space increased firing for neurons preferring that location) as well as in recent auditory mapping work (Da Costa, Van Der Zwaag, Miller, Clarke, & Saenz, 2013), Schwartz and David (2017) found that most primary auditory cortex neurons' responses (spike rate) to the narrowband noise around the attended tone frequency *decreased* compared to when the same noise was ignored. By contrast, spiking to the probe tone did not change significantly depending on whether its frequency band was attended or ignored. The authors suggest that this pattern may reflect very narrowly tuned adaptive suppression of *non-informative* noise around the cued frequency. This possibility will be interesting to test in future work and that also harkens back to the human studies discussed above showing adaptive reweighting of auditory cues based on their utility for extracting information from the speech stream.

The Impact of Context. The directionality of the context-dependent speech phenomena we reviewed above, and others like it in the literature, is *contrastive*. The pattern of results is such that perception is shifted *away* from the acoustic input dimensions of the preceding context, consistent with neural systems that emphasize change. Whether speech or nonspeech, precursors sampling a higher-frequency band shift speech categorization toward categories characterized by lower-frequency spectral energy. However, the alignment of the dimension or feature distinguishing the speech categories -- for example the third formant frequency band in /ga/ versus /da/ -- with the dimension or feature manipulated across the precursor context appears to be critical. Recall that manipulating the long-term average spectrum of a precursor in the third formant (F3) frequency band shifts /ga/-/da/ speech categorization. However, manipulation of the long-term average spectrum of preceding speech or nonspeech in the first formant (F1) frequency band has no effect on /ga/-/da/ categorization even though manipulations to the F1 frequency band do produce contrastive context effects on vowel categorization across vowels distinguished by their F1 frequencies (Huang & Holt, 2012).

Animal neurobiological studies suggest that stimulus-specific adaptation (SSA) exhibits an intriguingly similar profile in both its dimension- or feature-selectivity and response characteristics (Ulanovsky, Las, & Nelken, 2003) In SSA, neural responses to a particular stimulus are reduced in amplitude and delayed in latency when a stimulus with similar acoustics precedes it compared to the neural response to the same stimulus presented in isolation. However, SSA is not evident when the precursor stimulus is distinct enough that it fails to activate overlapping stimulus-specific neural populations (Jääskeläinen et al., 2007; May et al., 1999). In line with proposals made by Ulanovsky and colleagues (2003) the depression of neural response to regularity and the corresponding exaggeration of change with enhanced neural response may provide a means by which the system responds to regularity present across input dimensions. (See Hermann, Henry, & Obleser, 2013a for an example in human listeners).

Further, in a series of human behavioral studies of speech categorization Holt (2006) observed that the mean frequency of a distribution of tones (whether the distribution varied across 1000 Hz or included only tones repeated at the mean frequency) was the best predictor of its influence of categorization of subsequent speech. This resonates with findings from animal neurobiology. Ulanovsky et al. (2003) examined the response of primary auditory cortex neurons to equally probable, equal-amplitude tones with 20 different frequencies. The responses of the primary auditory cortex neurons to frequencies at the center of the stimulus frequency range adapted the most and there was relative enhancement of responses at the eccentric frequencies furthest away from the center of the frequency range. This created a U-shape in the neural tuning curves, with maximal adaptation at the central frequencies and relative enhancement at the edges. This appears to arise because adaptation strength is negatively associated with the frequency difference between the present stimulus and the stimulus from the preceding trial (Brosch & Schreiner, 1997; Ulanovsky et al., 2003). Thus, adaptation is greatest for central frequencies because central frequencies, on average, have smaller frequency differences from the preceding trials compared to eccentric frequencies. Holt (2006) argued that this may relate to the observation that the mean frequency of a distribution of preceding tones is the best predictor of the impact of context on speech categorization. In line with proposals made by Ulanovsky and colleagues (2003; 2004), the depression of neural response to regularity and corresponding exaggeration of change with enhanced neural response may provide a means by which the system responds to regularity across specific input dimensions. SSA seems to have some of the right properties to support the contrastive, dimension-specific contrast effects evident in speech perception (Holt, 2006).

Moving animal neurobiological studies even closer to the behavioral phenomena of speech perception, a recent study of songbird forebrain demonstrates that rapid discrimination of behaviorally-relevant vocalizations depends not only on specific stimulus features, but also on expectations generated from context about upcoming events (Lu & Vicario, 2017). When acoustic features of a target songbird vocalization differed from the statistical distribution of a preceding context song, auditory response to the target vocalization was significantly enhanced relative to when it shared the same acoustic distribution as preceding context. Thus, songbird auditory forebrain is dynamically modulated by acoustic context to emphasize complex acoustic dimensions that depart from the regularities build up across prior context. In mammalian species, human and nonhuman animal auditory cortex also is sensitive to statistical context across extended time scales (Yaron, Hershenhoren, & Nelken, 2012).

In this way, the distribution of acoustic dimensions evolving in incoming input provide a means of modulating auditory processing to bias the system to down-weight the significance of dimensions well-sampled in prior input and enhance those that are novel. Although these effects are not often spoken of as selective attention, this pattern of bias toward (or away from) a particular input dimension may be another way that the auditory system directs dimension-based selective attention to behaviorally-relevant objects and events. Indeed, Jääskeläinen et al. (2011) have made the case that the tuning of auditory cortical feature-specific neural populations via SSA is especially intriguing in light of the fact that such cortical tuning has been implicated as a mechanism of auditory selective attention (Fritz et al. 2007).

1.4 Building a Bridge from Animal Neurobiology to Human Phenomena

Like the speech perception phenomena reviewed above, these illustrative examples from nonhuman animal neuroscience demonstrate that the very mapping of acoustics to auditory dimensions and objects is dependent upon an organism's prior history of experience, the short-term experience evolving in the local input, and statistical relationships relating the present sound exemplar to those experienced previously. Yet, despite the intriguing connections reviewed above, there remains a gulf between the speech perception phenomena and the paradigms of animal neurobiological research in examining putative roles for dimension-based auditory attention. It would be highly desirable to have a human behavioral paradigm that could build a bridge this gulf in constructing a neurobiological model of human auditory perception, including speech perception, that incorporate dimension-based auditory attention.

To be clear, the goal need not be to model the speech phenomenon described above directly. Rather the aim would be to develop a productive test-bed for investigating non-spatial auditory dimension-based attention in human listeners that might inform us about the auditory mechanisms available to speech perception. In this context, any such paradigm would need to include several important elements.

First, nonspeech stimuli would be desirable as the use of speech complicates direct connections with the informative neurobiological research with nonhuman animals. Speech also makes it challenging to isolate specific auditory dimensions of selective attention and assessments across speech can be 'contaminated' by individual differences in language ability, native-language background, and other factors. Nonspeech sounds, in contrast, allow for fine-grained manipulation of acoustic parameters.

Second, task demands should require directing attention along a specific acoustic dimension. In humans, the most straightforward means of directing attention is to instruct participants to focus on a particular dimension (e.g., 'pay attention to the higher sounds), or on some sub-region of that dimension while ignoring another sub-region (e.g., 'the cue to press the button will be a high sound, and not a low sound'). Overtly guiding participants' attention to a part of the spectrum is an attractive possibility. From a practical perspective, participants' attention to frequency band can be directed using relative height terms. More importantly, frequency is the primary dimension of auditory representation and it has been used so productively in animal electrophysiology research on dimension-based auditory attention. In addition, it relates naturally to the formant-frequency-band-specific effects so common in speech perception, as well as to visual neuroscience paradigms that overtly direct attention to parts of retinotopic space.

Such explicit, symbolic (language-directed), and *endogenously driven* attention is experimentally convenient in that little to no training is required for participants to understand the task. However, it does not capture more *exogenous* attentional effects, such as those driven by the acoustic saliency or informational structure of the auditory scene. These effects are vital to account for, in that decades of psychoacoustic research using variants of the 'probe-signal' paradigm (Greenberg, 1968) have shown that detection and processing of isolated or embedded tones is strongly modulated by the presence and reliability of the preceding spectral context (Cusack, Decks, Aikman, & Carlyon, 2004; Dai, Scharf, & Buus, 1991; Green & McKeown, 2001; Hafter, Schlauch, & Tang, 1993; Hübner & Hafter, 1995; Larkin & Greenberg, 1970; Mondor, 1999; Mondor & Breau, 1999; Mondor, Breau, & Milliken, 1998; Reeves & Scharf, 2010; Richards & Neff, 2004; Scharf, Quigley, Aoki, Peachey, & Reeves, 1987; Scharf, Reeves, & Giovanetti, 2008; Scharf, Reeves, & Suciu, 2007; Tan, Robertson, & Hammond, 2008; Woods, Alain, Diaz, Rhodes, & Ogawa, 2001; Wright, 2005). Such findings are highly reminiscent of those using endogenous and exogenous spatial attentional cues in vision research (reviewed in Carrasco, 2011). Nor do explicitly cued attention paradigms get at the putatively attentional mechanisms underlying the dynamic perceptual reweighting *along multiple dimensions*, as discussed above for speech phenomena. Thus, a good experimental model of dimension-selective auditory attention should allow for *simultaneous* driving of more sustained, endogenous, and explicitly cued attention along with moment-to-moment manipulation of acoustic and informational parameters that transiently guide exogenous and endogenous attention along different auditory dimensions.

Finally, it would be desirable to utilize sounds that make strong demands on integration of information within a dimension, and to be able to manipulate the difficulty of this integration to place greater or lesser demands on the system, as this is surely a factor in speech processing. At the same time, it would be advantageous to be able to manipulate the relationship of a target input dimension with competing 'distractor' dimensions across sustained sound input. This would assist in bringing studies of non-spatial dimension-based auditory attention of the sort directed to brief segments of speech in closer alignment with more common studies of auditory attention across sustained sounds, as in the classic cocktail party phenomenon. An additional benefit is that this approach would align well with human neuroimaging tools and the demands of listening to continuous, fluent speech in everyday listening.

Building a bridge between mechanisms of auditory attention in spoken language comprehension to those revealed by non-human electrophysiology has been an active research agenda in the EEG/MEG field (e.g., Ding & Simon, 2013; Forte et al., 2017; Kong et al., 2014; O'Sullivan et al., 2014; Skoe & Kraus, 2010; Zion Golumbic et al., 2012). In the same spirit, here we present a novel experimental approach that meets these desiderata and we share four insights from preliminary research.

1.5 Sustained Auditory Selective Attention (SASA), A Novel Approach to Investigating Non-Spatial Dimension-based Auditory Selective Attention

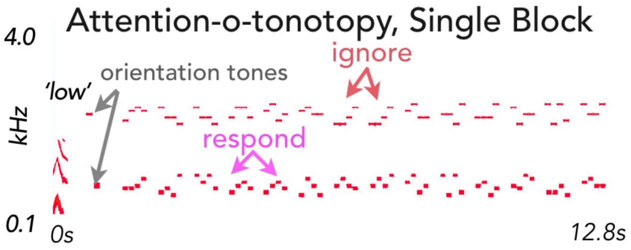

Figure 3. Example Stimuli from SASA Behavioral Paradigm. Spectrograms (Time x Frequency) plot an example stimulus. Stimuli consisted of four-tone 'mini-sequences within a 'target' frequency band paired with a 'distractor' frequency band. A verbal cue (*high / low*) prompted listeners to monitor a specific band for mini-sequence repeats. This required listeners to maintain sustained auditory selective attention to the evolution of spectral structure across time within a specific frequency band in the context of similarly-complex sounds in a distractor band.

In recent work, we developed a novel behavioral paradigm we refer to as *SASA*, the Sustained Auditory Selective Attention paradigm. In the SASA paradigm, listeners direct attention to a series of four-tone 'mini-sequences' that fall within a specific spectra band, without any auditory spatial cues (see Figure 3). Listeners monitor for temporally-adjacent mini-sequence repeats within the attended band. This puts a high demand on encoding and integrating information across a delimited frequency range, the center frequency of which varies across trials. Adding to the challenge, target mini-sequences are accompanied by mini-sequences in a distractor frequency band that varies in its spectral distance from the target frequency band. The distractor band may also contain mini-sequence repeats. A verbal cue (*high, low*) directs attention to a specific frequency band and brief 'orientation tones' alert listeners to the mean frequency of each band. Listeners report mini-sequence repeats in this target band with a key press.

The task meets the experimental *desiderata* outlined above in that it requires directing attention to a specific acoustic dimension, namely spectral band. (We discuss other manipulable dimensions below). The task involves nonspeech stimuli that make strong demands on integrating information (the mini-sequences) across an input dimension (the frequency band) and that can be extended across time to require sustained selective attention. Likewise, SASA requires spectrally-selective attention to a particular frequency band. In this, it aligns well with the nonhuman animal literature that has similarly capitalized on frequency as a significant acoustic input dimension across which selective attention can be directed (see Fritz et al. 2007).

In the next section, we describe four insights from utilizing this SASA paradigm among adult human listeners and describe how future work might exploit the approach further to make closer connections between the speech phenomena reviewed above and animal neurobiological models.

1.6 Four Insights from SASA

1.6.1 There are Substantial Individual Differences Even in Typical Young Adults

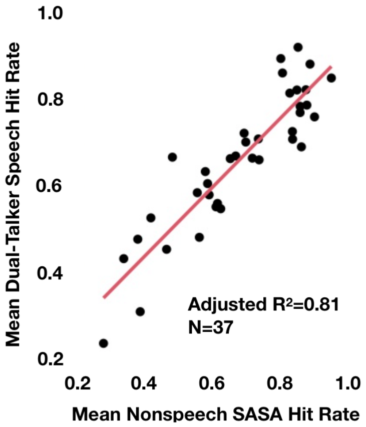

Figure 4. There is substantial individual variability in nonspeech SASA performance that is associated with speech comprehension in dual-talker conditions. Here, we plot each participant's average hit rates for nonspeech SASA versus dual-talker speech. The same relationship holds ($R^2 = 0.81$) when participants' accuracy for a standard mental rotation are included in the regression model, suggesting that the tight relationship between nonspeech SASA and dual-talker speech is not simply driven by individual differences in participants' generic ability to perform experimental tasks.

A first study examined the range of individual variation in SASA performance among healthy young-adult (N=37) university students. In this study, listeners completed a temporally-interleaved version of the nonspeech SASA task that complemented the simultaneous version shown in Figure 3. In this version, the high and low frequency bands alternated in time (every 125 ms) and listeners' task was to monitor one of the bands for mini-sequence repeats. The results are plotted in Figure 4. Even among this rather homogeneous sample of young-adult university students, there were substantial individual differences in performance on the nonspeech SASA task (apparent across the range variation on the Figure 4 x-axis). This is important in that it reveals that even healthy listeners differ in baseline ability to direct and sustain auditory selective attention to a specific acoustic dimension. Larger-scale future studies sampling a more diverse participant population have the potential to establish the range of individual variability evident among healthy listeners. This would be highly desirable as a benchmark for clinical assessment of dimension-based auditory selective attention among healthy older listeners who exhibit auditory selective attention difficulties, and among individuals with neurodevelopmental or neurodegenerative disorders that impact auditory attention (Shinn-Cunningham, 2017). It may be especially valuable that the SASA task is unlikely to be contaminated by language ability, native-language background, and other speech-specific factors.

1.6.2 Performance in the Nonspeech SASA Paradigm is Associated with Speech Comprehension in Dual-Talker Conditions

In the same study, we also sought to examine whether the novel SASA task demanding selective attention to a specific frequency band across nonspeech stimuli relates to more common measures of auditory selective attention, specifically in the speech domain. For this reason, the same participants also completed a dual-talker speech task, similar to canonical multitalker studies of real-world listening challenges (Brungart, Simpson, Darwin, Arbogast, & Kidd, 2005). In this task, listeners attempted to detect exact repetitions in a string of 3 key words in the attended talker stream (male/female), while ignoring the other talker. As a control for overall performance, listeners also completed a version of a classic mental rotation task (Shepard & Metzler, 1971).

The strong relationship between SASA performance and dual-talker speech performance illustrated in Figure 4 indicates auditory selective attention to specific frequency bands, as measured using the novel nonspeech SASA task, is strongly associated with dual-talker speech comprehension and holds even when mental rotation is included as a factor in the general linear model to control for overall performance differences. This is important in that it indicates that performance in the nonspeech SASA paradigm is robustly associated with a multi-talker speech comprehension challenge that demands dimension-based auditory selective attention. This is exciting because it suggests that the nonspeech SASA paradigm can serve as a proxy for everyday listening challenges. Whereas comprehension of speech in noise is a common model of auditory selective attention, the use of speech complicates direct connections with informative neurobiological research with nonhuman animal models, makes it challenging to isolate specific auditory dimensions of selective attention, and can be contaminated by individual differences in language ability. The nonspeech SASA paradigm allows greater experimental control over details of the target and distractor dimensions than is possible with natural speech stimuli and connects directly to productive animal neurobiological models. Future studies more directly connecting this approach to the speech phenomena reviewed above might, for example, take the approach of manipulating regularities across which task-relevant information appears in a specific frequency band (to tap into perceptual weighting and associated plasticity). Just as importantly, there is considerable opportunity to carefully manipulate demands upon human spectrally-selective attention in order to address the many open questions regarding basic mechanism.

1.6.3 Listeners Can Learn to Better Deploy Dimension-based Auditory Attention

Especially intriguing, training can improve listeners' ability to deploy non-spatial dimension-based auditory selective attention. In a separate cohort of listeners sampled from the same population of healthy university students, we provided two 1-hour sessions of training with feedback on the nonspeech SASA task. As shown in Figure 5, most listeners improved in their ability to integrate information in the target frequency band in the context of complex acoustic information in a distractor frequency band. This implicates behavioral training as a viable intervention that may improve dimension-based auditory selective attention among those with poor baseline abilities, or clinical impairment of auditory selective attention. An exciting, as yet unexplored, possibility is that such training might improve listeners' ability to direct attention to *specific* frequency bands. It might be possible, for example, to redirect spectral attention to higher frequencies that carry significant speech information (Monson, Hunter, & Story, 2012; Monson, Lotto, & Story, 2014; Vitela, Monson, & Lotto, 2015) in the context of noisy surroundings that mask lower frequencies, thereby encouraging new perceptual weighting schemes beneficial to behavior.

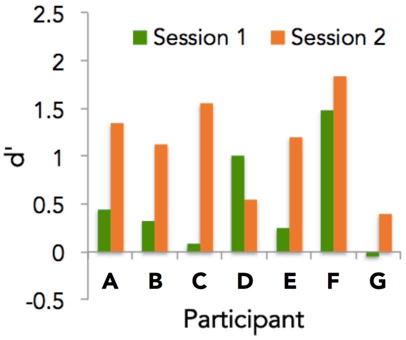

Figure 5. Training Improves SASA Performance. Brief (two-session, 90 minute) training with the SASA model paradigm improves performance in most participants.

1.6.4 Dimension-based Auditory Attention can be Topographically Mapped in Human Primary and Non-Primary Auditory Cortex

Acoustic frequency is a particularly attractive model for dimension-based auditory attention in that (a) informative and / or disambiguating acoustic cues in ecologically-relevant environmental sounds and intentional communicative signals are unevenly distributed across the spectrum and (b) frequency is topographically mapped across multiple auditory areas that differentially contribute to perceptual and decision processes. However, as noted by Schwartz and David (2017), it has been challenging to come up with paradigms in nonhuman animals that isolate frequency-selective attention from other attentional factors -- a primary goal of our human SASA paradigm. In humans, recent work on spectrally-selective attention (Da Costa et al., 2013, see also Paltoglou, Sumner, & Hall, 2009) has shown that when listeners attend to either a high or low frequency stream containing behavioral targets (with both streams presented simultaneously, but to different ears), voxels in auditory regions with preferred frequencies near an attended frequency band show increased blood-oxygen-level-dependent (BOLD) activation, whereas voxels with preferred frequencies far from the attended frequency band show decreased BOLD activity. Using an innovative melody-monitoring paradigm in a three-frequency-band stimulus, Riecke et al. (2016) showed that the topography of spectral attention significantly echoed tonotopic maps in early auditory areas; in putative secondary areas, attended frequency could be decoded using multivoxel-pattern classification approaches, but did not seem to follow tonotopic progressions.

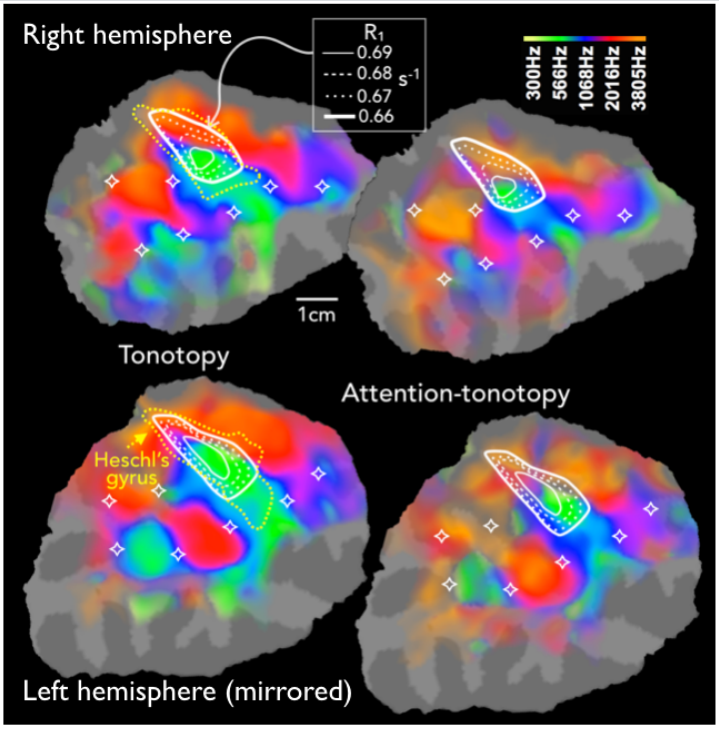

Figure 6. Group average Tonotopy and Attention-Tonotopy maps on flattened superior temporal lobe patches, with R1 contours showing putative auditory core, from Dick et al. (2017). Isocontour lines show quantitative R1 values for the group-averaged putative auditory core, and color maps showing group-averaged best frequency in both conditions. The stars are fiduciary points to assist in visual comparisons of maps across conditions; the outline of Heschl's gyrus is in yellow dashed lines. The average tonotopic map is characterized by two pairs of three interlacing best-frequency 'fingers,' with the high-frequency fingers (red/orange colormap) showing greatest frequency preference medially and extending laterally, where they meet interdigitated lower-frequency fingers (green/yellow colormap) extending lateral to medial, with the longest 'middle' lower-frequency finger extending about halfway into auditory core. This pattern is evident in Fourier-analysis-derived maps of the Attention-tonotopy condition but not in the 'randomized control' for which the attentional response was phase-cancelled (not shown here)

Using our non-speech SASA paradigm, we have recently examined spectral-based auditory selective attention in human cortex, combining functional MRI with high-resolution quantitative MRI in order to identify putative auditory core (Dick et al. 2017). Here, we observed that human primary and much of non-primary auditory cortical activation is strongly modulated by spectrally-directed auditory selective attention to five different frequency bands, in a manner that recapitulates its tonotopic sensory organization. The detailed, graded activation profiles elicited by single frequency bands (without distractors) were strongly associated with attentionally-driven activation when these frequency bands were accompanied by distractors (acoustic stimuli as in Figure 3, Figure 6 shows group average maps for tonotopic and 'attention-o-tonotopic' conditions from Dick et al., 2017). Moreover, systematic spatial maps of 'dis-preferred frequency' (the frequency that drove the smallest response at each voxel) could also be recapitulated by frequency-directed attention to those same frequencies. Finally, the graded frequency preferences observed in small patches across auditory cortex were closely aligned to those evoked by attention to those frequencies in the presence of distractor frequency bands.

1.6.5 SASA Overview and Future Directions

These initial studies using the SASA paradigm demonstrate that we can non-invasively observe dimension-based auditory selective attention in the human brain by embedding task-relevant information in different regions of the frequency spectrum - here the dimension along which attention is directed. A major advantage of this approach is that brings human auditory cortical paradigms into closer alignment with informative animal electrophysiological research. Additionally, since behavioral research using the same paradigm indicates the close association of performance in this nonspeech SASA task with comprehension of multi-talker speech it builds a bridge across which to connect traditional approaches in human listeners like perception of speech in noise with these productive animal paradigms. Since training in the nonspeech SASA paradigm leads to improvements in the ability to direct attention to specific frequency bands, the pairing of training with these neuroimaging approaches can present new opportunities for understanding how dimension-based auditory selective attention relates to short- and long-term plasticity.

Notably, these first studies using the SASA paradigm did not manipulate listeners' attention to auditory dimensions other than spectral band, nor did they explore any other means of directing attention than through specific verbal instruction. The SASA paradigm can accommodate explicit attention to other dimensions through varying the acoustic character of the individual sequence elements, which are not limited to pure tones but can be complex tones or synthetic sound objects. For instance, attention can be directed to durational or timbral characteristics that define the task-relevant mini-sequence stream - similar to the way that listeners at a concert will attend to spectrally and temporally overlapping flute or oboe lines in an orchestral piece. A SASA variant more analogous to the dimension-based dynamic reweighting effects discussed above might provide multiple probabilistic acoustic cues that would predict the occurrence of a mini-sequence repeat. As an example, explicit attention could be directed a given spectral band (as in the original SASA), but the acoustic characteristics of the constituent tone elements would vary constantly in two dimensions (duration and envelope) in both attended and unattended bands. Target mini-sequences would be more likely to occur when preceded by tones of shorter duration, or a combination cue of shorter duration and sharper onset envelope. Such a configuration would allow for listeners to discover and selectively direct attention along the acoustic dimension(s) that are task-informative, as in the speech examples above. The dynamics of this (putative) functionally-driven attentional reweighting could be directly compared to parallel manipulations in speech or speech-like domains.

1.7 Summary and Conclusions

Was that a *beach* or a *peach*? This rather simple example, the auditory dimensions of which evolve across just 10s of milliseconds, proves to involve more complex processing that has traditionally been described. In contrast to early accounts of speech processing that emphasized rather static mapping of input dimensions to discrete phonemic representations, contemporary research highlights that speech perception involves selective weighting of acoustic input dimensions as a function of context and both short- and long-term input regularities. We have attempted to make a case that selective attention to specific, non-spatial auditory dimensions may be an important contributor in this dynamic mapping of speech input to behaviorally-relevant representations and actions. Yet, the state of our understanding is such that there remain many open questions regarding this putative link. We do not yet have a deep understanding of human auditory selective attention, especially as it relates to directing attention to specific, non-spatial dimensions evolving within a sound object of the sort potentially demanded by speech phenomena reviewed above. Nevertheless, there are important parallels emerging in animal neurobiological research. This work suggests that the phenomena we refer to as involving *selective attention* are likely to draw from multiple neurobiological mechanisms. The hope is that paradigms that put human and nonhuman animal research into closer alignment, as in the case of the SASA paradigm we reviewed above, can facilitate progress in discovering the basic mechanisms of auditory selective attention available to support higher-level processing like that demanded by speech to move us beyond selective attention as a cognitive placeholder.

Author Note

Data presented involving human subjects were collected in accordance with the Declaration of Helsinki, with informed consent obtained. The authors have no competing interests to declare. Support for the preparation of this manuscript came from the National Institutes of Health, the Rothberg Research Award in Human Brain Imaging of Carnegie Mellon University, and the Marie Curie Sklodowska Actions of the European Commission.

References

Abramson, A. S., & Whalen, D. H. (2017). Voice Onset Time (VOT) at 50: Theoretical and practical issues in measuring voicing distinctions. *Journal of Phonetics*, *63*, 75–86. http://doi.org/10.1016/j.wocn.2017.05.002

Assgari, A. A., & Stilp, C. E. (2015). Talker information influences spectral contrast effects in speech categorization. *The Journal of the Acoustical Society of America*, *138*(5), 3023–3032. http://doi.org/10.1121/1.4934559

Atencio, C. A., Sharpee, T. O., & Schreiner, C. E. (2008). Cooperative nonlinearities in auditory cortical neurons. *Neuron*, *58*(6), 956–966. http://doi.org/10.1016/j.neuron.2008.04.026

Atiani, S., David, S. V., Elgueda, D., Locastro, M., Radtke-Schuller, S., Shamma, S. A., & Fritz, J. B. (2014). Emergent Selectivity for Task-Relevant Stimuli in Higher-Order Auditory Cortex. *Neuron*, *82*(2), 486–499. http://doi.org/10.1016/j.neuron.2014.02.029

Aydelott, J., & Bates, E. (2004). Effects of acoustic distortion and semantic context on lexical access. *Language and Cognitive Processes*, *19*(1), 29–56. http://doi.org/10.1080/01690960344000099

Bieszczad, K. M., & Weinberger, N. M. (2010). Representational gain in cortical area underlies increase of memory strength. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(8), 3793–3798. http://doi.org/10.1073/pnas.1000159107

Blair, M. R., Watson, M. R., Walshe, R. C., & Maj, F. (2009). Extremely selective attention: Eye-tracking studies of the dynamic allocation of attention to stimulus features in categorization. *Journal of Experimental Psychology Learning, Memory, and Cognition*, *35*(5), 1196–1206. http://doi.org/10.1037/a0016272

Blumstein, S. E., & Stevens, K. N. (1985). On some issues in the pursuit of acoustic invariance in speech: a reply to Lisker. *The Journal of the Acoustical Society of America*, *77*(3), 1203–1204.

Bradlow, A. R., Nygaard, L. C., & Pisoni, D. B. (1999). Effects of talker, rate, and amplitude variation on recognition memory for spoken words. *Perception & Psychophysics*, *61*(2), 206–219.

Brosch, M., & Schreiner, C. E. (1997). Time course of forward masking tuning curves in cat primary auditory cortex. *Journal of Neurophysiology*, *77*(2), 923–943. http://doi.org/10.1152/jn.1997.77.2.923

Brungart, D. S., Simpson, B. D., Darwin, C. J., Arbogast, T. L., & Kidd, G. (2005). Across-ear interference from parametrically degraded synthetic speech signals in a dichotic cocktail-party listening task. *The Journal of the Acoustical Society of America*, *117*(1), 292–304.

Carcea, I., Insanally, M. N., & Froemke, R. C. (2017). Dynamics of auditory cortical activity during behavioural engagement and auditory perception. *Nature Communications*, *8*, 14412. http://doi.org/10.1152/jn.00048.2016

Carrasco, M. (2011). Visual attention: the past 25 years. *Vision Research*, *51*(13), 1484–1525. http://doi.org/10.1016/j.visres.2011.04.012

Cherry, C. E. (1953). Some experiments on the recognition of speech, with one and two ears. *The Journal of the Acoustical Society of America, 25*, 975-979.

Chua, K.-W., Richler, J. J., & Gauthier, I. (2015). Holistic processing from learned attention to parts. *Journal of Experimental Psychology: General*, *144*(4), 723–729. http://doi.org/10.1037/xge0000063

Cohen, J. D., Romero, R. D., Servan-Schreiber, D., & Farah, M. J. (1994). Mechanisms of spatial attention: the relation of macrostructure to microstructure in parietal neglect. *Journal of Cognitive Neuroscience*, 6, 377-387.

Conboy, B. T., & Kuhl, P. K. (2011). Impact of second-language experience in infancy: brain measures of first- and second-language speech perception. *Developmental Science*, *14*(2), 242–248. http://doi.org/10.1111/j.1467-7687.2010.00973.x

Cusack, R., Decks, J., Aikman, G., & Carlyon, R. P. (2004). Effects of Location, Frequency Region, and Time Course of Selective Attention on Auditory Scene Analysis. *Journal of Experimental Psychology Human Perception and Performance*, *30*(4), 643–656. http://doi.org/10.1037/0096-1523.30.4.643

Da Costa, S., Van Der Zwaag, W., Miller, L. M., Clarke, S., & Saenz, M. (2013). Tuning in to sound: frequency-selective attentional filter in human primary auditory cortex. *Journal of Neuroscience*, *33*(5), 1858–1863. http://doi.org/10.1523/JNEUROSCI.4405-12.2013

Dai, H. P., Scharf, B., & Buus, S. (1991). Effective attenuation of signals in noise under focused attention. *The Journal of the Acoustical Society of America, 89*(6), 2837–2842. http://doi.org/10.1121/1.400721

David, S. V., Fritz, J. B., & Shamma, S. A. (2012). Task reward structure shapes rapid receptive field plasticity in auditory cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(6), 2144–2149. http://doi.org/10.1073/pnas.1117717109

Dehaene-Lambertz, G., Pallier, C., Serniclaes, W., Sprenger-Charolles, L., Jobert, A., & Dehaene, S. (2005). Neural correlates of switching from auditory to speech perception. *NeuroImage*, *24*(1), 21–33. http://doi.org/10.1016/j.neuroimage.2004.09.039

Desai, R., Liebenthal, E., Waldron, E., & Binder, J. R. (2008). Left posterior temporal regions are sensitive to auditory categorization. *Journal of Cognitive Neuroscience*, *20*(7), 1174–1188. http://doi.org/10.1162/jocn.2008.20081

Ding, N., and Simon, J. Z. (2013). Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. *Journal of Neuroscience, 33,* 5728-5735.

Forte, A. E., Etard, O., & Reichenbach, T. (2017). The human auditory brainstem response to running speech reveals a subcortical mechanism for selective attention. *eLife,* e27203.

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*(2), 349–366.

Francis, A. L., Ciocca, V., Wong, N. K. Y., Leung, W. H. Y., & Chu, P. C. Y. (2006). Extrinsic context affects perceptual normalization of lexical tone. *The Journal of the Acoustical Society of America*, *119*(3), 1712. http://doi.org/10.1121/1.2149768

Francis, A. L., Kaganovich, N., & Driscoll-Huber, C. (2008). Cue-specific effects of categorization training on the relative weighting of acoustic cues to consonant voicing in English. *The Journal of the Acoustical Society of America*, *124*(2), 1234–1251. http://doi.org/10.1121/1.2945161

Goldinger, S. D. (1996). Words and voices: episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology Learning, Memory, and Cognition*, *22*(5), 1166–1183.

Gordon, P. C., Eberhardt, J. L., & Rueckl, J. G. (1993). Attentional Modulation of the Phonetic Significance of Acoustic Cues. *Cognitive Psychology*, *25*(1), 1–42. http://doi.org/10.1006/cogp.1993.1001

Green, T. J., & McKeown, J. D. (2001). Capture of attention in selective frequency listening. *Journal of Experimental Psychology Human Perception and Performance*, *27*(5), 1197–1210.

Greenberg, G. Z. (1968). Frequency-Response Characteristic of Auditory Observers Detecting Signals of a Single

Frequency in Noise: The Probe-Signal Method. *The Journal of the Acoustical Society of America*, *44*(6), 1513. http://doi.org/10.1121/1.1911290

Grubb, M. A., White, A. L., Heeger, D. J., & Carrasco, M. (2014). Interactions between voluntary and involuntary attention modulate the quality and temporal dynamics of visual processing. *Psychonomic Bulletin & Review*, *22*(2), 437–444. http://doi.org/10.3758/s13423-014-0698-y

Hackett, T. A. (2011). Information flow in the auditory cortical network. *Hearing Research*, *271*(1-2), 133–146. http://doi.org/10.1016/j.heares.2010.01.011

Hafter, E. R., Schlauch, R. S., & Tang, J. (1993). Attending to auditory filters that were not stimulated directly. *The Journal of the Acoustical Society of America*, *94*(2), 743–747. http://doi.org/10.1121/1.408203

Heald, S. L. M., & Nusbaum, H. C. (2014). Speech perception as an active cognitive process. *Frontiers in Systems Neuroscience*, *8*, 35. http://doi.org/10.3389/fnsys.2014.00035

Henry, M. J., Herrmann, B., & Obleser, J. (2015). Selective attention to temporal features on nested time scales. *Cerebral Cortex*, *25*(2), 450–459. http://doi.org/10.1093/cercor/bht240

Herrmann, B., Henry, M. J., & Obleser, J. (2013a). Frequency-specific adaptation in human auditory cortex depends on the spectral variance in the acoustic stimulation. *Journal of Neurophysiology*, *109*(8), 2086–2096. http://doi.org/10.1152/jn.00907.2012

Herrmann, B., Henry, M. J., Scharinger, M., & Obleser, J. (2013b). Auditory filter width affects response magnitude but not frequency specificity in auditory cortex. *Hearing Research*, *304*, 128–136. http://doi.org/10.1016/j.heares.2013.07.005

Herrmann, B., Schlichting, N., & Obleser, J. (2013c). Dynamic Range Adaptation to Spectral Stimulus Statistics in Human Auditory Cortex. *Journal of Neuroscience*, *34*(1), 327–331. http://doi.org/10.1523/JNEUROSCI.3974-13.2014

Holt, L. L. (2005). Temporally nonadjacent nonlinguistic sounds affect speech categorization. *Psychological Science*, *16*(4), 305–312. http://doi.org/10.1111/j.0956-7976.2005.01532.x

Holt, L. L. (2006a). Speech categorization in context: joint effects of nonspeech and speech precursors. *The Journal of the Acoustical Society of America*, *119*(6), 4016–4026. http://doi.org/10.1121/1.2195119

Holt, L. L. (2006b). The mean matters: effects of statistically defined nonspeech spectral distributions on speech categorization. *The Journal of the Acoustical Society of America*, *120*(5 Pt 1), 2801–2817. http://doi.org/10.1121/1.2354071

Holt, L. L., & Lotto, A. J. (2006). Cue weighting in auditory categorization: implications for first and second language acquisition. *The Journal of the Acoustical Society of America*, *119*(5 Pt 1), 3059–3071. http://doi.org/10.1121/1.2188377

Holt, L. L., & Lotto, A. J. (2010). Speech perception as categorization. *Attention, Perception, & Psychophysics*, *72*(5), 1218–1227. http://doi.org/10.3758/APP.72.5.1218

Holt, L. L., Lotto, A. J., & Diehl, R. L. (2004). Auditory discontinuities interact with categorization: Implications for speech perception. *The Journal of the Acoustical Society of America*, *116*(3), 1763–1773. http://doi.org/10.1121/1.1778838

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.

Huang, J., & Holt, L. L. (2012). Listening for the Norm: Adaptive Coding in Speech Categorization. *Frontiers in Psychology*, *3*. http://doi.org/10.3389/fpsyg.2012.00010

Hübner, R., & Hafter, E. R. (1995). Cuing mechanisms in autitory signal detection. *Perception & Psychophysics*, *57*(2), 197–202. http://doi.org/10.3758/BF03206506

Idemaru, K., & Holt, L. L. (2011). Word recognition reflects dimension-based statistical learning. *Journal of Experimental Psychology Human Perception and Performance*, *37*(6), 1939–1956. http://doi.org/10.1037/a0025641

Idemaru, K., & Holt, L. L. (2013). The developmental trajectory of children's perception and production of English /r/-/l/. *The Journal of the Acoustical Society of America*, *133*(6), 4232–4246. http://doi.org/10.1121/1.4802905

Idemaru, K., & Holt, L. L. (2014). Specificity of dimension-based statistical learning in word recognition. *Journal of Experimental Psychology Human Perception and Performance*, *40*(3), 1009–1021. http://doi.org/10.1037/a0035269

Iverson, P., & Kuhl, P. K. (1995). Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. *The Journal of the Acoustical Society of America*, *97*(1), 553–562.

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., & Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition*, *87*(1), B47–B57. http://doi.org/10.1016/S0010-0277(02)00198-1

Jääskeläinen, I. P ., Ahveninen, J., Belliveau, J. W., Raij, T., & Sams, M. (2007). Short-term plasticity in auditory cognition. *Trends in Neuroscience*, 30, 653–661.

Jääskeläinen, I. P ., Ahveninen, J., Andermann, M. L., Belliveau, J. W., Raij, T., & Sams, M. (2011). Short-term plasticity as a neural mechanism supporting memory and attentional functions. *Brain Research*, 1422, 66–81.

Kondaurova, M. V., & Francis, A. L. (2008). The relationship between native allophonic experience with vowel duration and perception of the English tense/lax vowel contrast by Spanish and Russian listeners. *The Journal of the Acoustical Society of America*, *124*(6), 3959–3971. http://doi.org/10.1121/1.2999341

Kondaurova, M. V., & Francis, A. L. (2010). The role of selective attention in the acquisition of English tense and lax vowels by native Spanish listeners: comparison of three training methods. *Journal of Phonetics*, *38*(4), 569–587. http://doi.org/10.1016/j.wocn.2010.08.003

Kong, Y. Y., Mullangi, A., and Ding, N. (2014). Differential modulation of auditory responses to attended and unattended speech in different listening conditions. *Hearing Research, 316,* 73-81.

Kruschke, J. K., Kappenman, E. S., & Hetrick, W. P. (2005). Eye Gaze and Individual Differences Consistent With Learned Attention in Associative Blocking and Highlighting. *Journal of Experimental Psychology Learning, Memory, and Cognition*, *31*(5), 830–845. http://doi.org/10.1037/0278-7393.31.5.830

Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. *Nature Reviews Neuroscience*, *5*(11), 831–843. http://doi.org/10.1038/nrn1533

Ladefoged, P., & Broadbent, D. E. (1957). Information Conveyed by Vowels. *The Journal of the Acoustical Society of America*, *29*(1), 98–104. http://doi.org/10.1121/1.1908694

Laing, E. J. C., Liu, R., Lotto, A. J., & Holt, L. L. (2012). Tuned with a Tune: Talker Normalization via General Auditory Processes. *Frontiers in Psychology*, *3*. http://doi.org/10.3389/fpsyg.2012.00203

Larkin, W., & Greenberg, G. Z. (1970). Selective attention in uncertain frequency detection1. *Perception & Psychophysics*, *8*(3), 179–184. http://doi.org/10.3758/BF03210201

Leech, R., Gygi, B., Aydelott, J., & Dick, F. (2009a). Informational factors in identifying environmental sounds in natural auditory scenes. *The Journal of the Acoustical Society of America*, *126*(6), 3147–3155. http://doi.org/10.1121/1.3238160

Leech, R., Holt, L. L., Devlin, J. T., & Dick, F. (2009b). Expertise with artificial nonspeech sounds recruits speech-sensitive cortical regions. *Journal of Neuroscience*, *29*(16), 5234–5239. http://doi.org/10.1523/JNEUROSCI.5758-08.2009

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), 358–368.

Lisker, L. (1985). The pursuit of invariance in speech signals. *The Journal of the Acoustical Society of America*, *77*(3), 1199–1202.

Lisker, L. (1986). "Voicing" in English: a catalogue of acoustic features signaling/b/versus/p/in trochees. *Language and Speech*, *29*(1), 3–11. http://doi.org/10.1177/002383098602900102

Liu, R., & Holt, L. L. (2011). Neural changes associated with nonspeech auditory category learning parallel those of speech category acquisition. *Journal of Cognitive Neuroscience*, *23*(3), 683–698. http://doi.org/10.1162/jocn.2009.21392

Lotto, A. J., & Holt, L. L. (2006). Putting phonetic context effects into context: a commentary on Fowler (2006). *Perception & Psychophysics*, *68*(2), 178–183.

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

Lu, K., & Vicario, D. S. (2017). Familiar But Unexpected: Effects of Sound Context Statistics on Auditory Responses in the Songbird Forebrain. *Journal of Neuroscience*, *37*(49), 12006–12017. http://doi.org/10.1523/JNEUROSCI.5722-12.2017

McMurray, B., Aslin, R. N., Tanenhaus, M. K., Spivey, M. J., & Subik, D. (2008). Gradient sensitivity to within-category variation in words and syllables. *Journal of Experimental Psychology Human Perception and Performance*, *34*(6), 1609–1631. http://doi.org/10.1037/a0011747

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition*, *86*(2), B33–42.

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic Feature Encoding in Human Superior Temporal Gyrus. *Science (New York, NY)*. http://doi.org/10.1126/science.1245994

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

Mondor, T. A. (1999). Predictability of the cue-target relation and the time-course of auditory inhibition of return. *Perception & Psychophysics*, *61*(8), 1501–1509. http://doi.org/10.3758/BF03213113

Mondor, T. A., & Breau, L. M. (1999). Facultative and inhibitory effects of location and frequency cues: Evidence of a modulation in perceptual sensitivity. *Perception & Psychophysics, 61*(3), 438–444. http://doi.org/10.3758/BF03211964

Mondor, T. A., Breau, L. M., & Milliken, B. (1998). Inhibitory processes in auditory selective attention: Evidence of location-based and frequency-based inhibition of return - Springer. *Perception & Psychophysics*.

Monson, B. B., Hunter, E. J., & Story, B. H. (2012). Horizontal directivity of low- and high-frequency energy in speech and singing. *The Journal of the Acoustical Society of America, 132*(1), 433–441. http://doi.org/10.1121/1.4725963

Monson, B. B., Lotto, A. J., & Story, B. H. (2014). Detection of high-frequency energy level changes in speech and singing. *The Journal of the Acoustical Society of America*, *135*(1), 400–406. http://doi.org/10.1121/1.4829525

Morillon, B., Hackett, T. A., Kajikawa, Y., & Schroeder, C. E. (2015). Predictive motor control of sensory dynamics in auditory active sensing - ScienceDirect. *Current Opinion in Neurobiology*, *31*, 230–238. http://doi.org/10.1016/j.conb.2014.12.005

Nittrouer, S., Lowenstein, J. H., & Packer, R. R. (2009). Children discover the spectral skeletons in their native language before the amplitude envelopes. *Journal of Experimental Psychology Human Perception and Performance*, *35*(4), 1245–1253. http://doi.org/10.1037/a0015020

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–61.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1995). Effects of stimulus variability on perception and representation of spoken words in memory. *Perception & Psychophysics*, *57*(7), 989–1001.

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., et al. (2014). Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cerebral Cortex, 25,* 1697-1706.

Paltoglou, A. E., Sumner, C. J., & Hall, D. A. (2009). Examining the role of frequency specificity in the enhancement and suppression of human cortical activity by auditory selective attention. *Hearing Research*, *257*(1-2), 106–118. http://doi.org/10.1016/j.heares.2009.08.007

Reeves, A., & Scharf, B. (2010). Auditory frequency focusing is very rapid. *The Journal of the Acoustical Society of America*, *128*(2), 795–803. http://doi.org/10.1121/1.3458823

Richards, V. M., & Neff, D. L. (2004). Cuing effects for informational masking. *The Journal of the Acoustical Society of America*, *115*(1), 289–300. http://doi.org/10.1121/1.1631942

Riecke, L., Peters, J. C., Valente, G., Kemper, V. G., Formisano, E., & Sorger, B. (2016). Frequency-Selective Attention in Auditory Scenes Recruits Frequency Representations Throughout Human Superior Temporal Cortex. *Cerebral Cortex*. http://doi.org/10.1093/cercor/bhw160

Scharf, B., Quigley, S., Aoki, C., Peachey, N., & Reeves, A. (1987). Focused auditory attention and frequency selectivity. *Perception & Psychophysics*, *42*(3), 215–223. http://doi.org/10.3758/BF03203073

Scharf, B., Reeves, A., & Giovanetti, H. (2008). Role of attention in overshoot: frequency certainty versus uncertainty. *The Journal of the Acoustical Society of America*, *123*(3), 1555–1561. http://doi.org/10.1121/1.2835436

Scharf, B., Reeves, A., & Suciu, J. (2007). The time required to focus on a cued signal frequency. *The Journal of the Acoustical Society of America*, *121*(4), 2149–2157. http://doi.org/10.1121/1.2537461

Scharinger, M., Herrmann, B., Nierhaus, T., & Obleser, J. (2014). Simultaneous EEG-fMRI brain signatures of auditory cue utilization. *Frontiers in Neuroscience*, *8*, 137. http://doi.org/10.3389/fnins.2014.00137

Schwartz, Z. P., & David, S. V. (2017). Focal Suppression of Distractor Sounds by Selective Attention in Auditory Cortex. *Cerebral Cortex (New York, NY : 1991)*, *28*(1), 323–339. http://doi.org/10.1093/cercor/bhx288

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

Shepard, K. N., Lin, F. G., Zhao, C. L., Chong, K. K., & Liu, R. C. (2015). Behavioral Relevance Helps Untangle Natural Vocal Categories in a Specific Subset of Core Auditory Cortical Pyramidal Neurons. *Journal of Neuroscience*, *35*(6), 2636–2645. http://doi.org/10.1523/JNEUROSCI.3803-14.2015

Shepard, R. B., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science (New York, NY)*, *171*, 701–703.

Shinn-Cunningham, B. (2017). Cortical and Sensory Causes of Individual Differences in Selective Attention Ability Among Listeners With Normal Hearing Thresholds. *Journal of Speech, Language, and Hearing Research : JSLHR*, *60*(10), 2976. http://doi.org/10.1044/2017_JSLHR-H-17-0080

Skoe, E. & Kraus, N. (2010) Hearing It Again and Again: On-Line Subcortical Plasticity in Humans. *PLoS One, 26*, e13645.

Tan, M. N., Robertson, D., & Hammond, G. R. (2008). Separate contributions of enhanced and suppressed sensitivity to the auditory attentional filter. *Hearing Research*, *241*(1-2), 18–25. http://doi.org/10.1016/j.heares.2008.04.003

Toscano, J. C., & Mcmurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science*, *34*(3), 434–464. http://doi.org/10.1111/j.1551-6709.2009.01077.x

Tsunada, J., Liu, A. S. K., Gold, J. I., & Cohen, Y. E. (2015). Causal contribution of primate auditory cortex to auditory perceptual decision-making. *Nature Neuroscience*. http://doi.org/10.1038/nn.4195

Ulanovsky, N., Las, L., & Nelken, I. (2003). Processing of low-probability sounds by cortical neurons. *Nature Neuroscience*, *6*(4), 391–398. http://doi.org/10.1038/nn1032

Utman, J. A. (1998). Effects of local speaking rate context on the perception of voice-onset time in initial stop consonants. *The Journal of the Acoustical Society of America*, *103*(3), 1640–1653. http://doi.org/10.1121/1.421297

Utman, J. A., Blumstein, S. E., & Burton, M. W. (2000). Effects of subphonetic and syllable structure variation on word recognition. *Perception & Psychophysics*, *62*(6), 1297–1311.

Utman, J. A., Blumstein, S. E., & Sullivan, K. (2001). Mapping from Sound to Meaning: Reduced Lexical Activation in Broca's Aphasics. *Brain and Language*, *79*(3), 444–472. http://doi.org/10.1006/brln.2001.2500

Van Gulick, A. E., & Gauthier, I. (2014). The perceptual effects of learning object categories that predict perceptual goals. *Journal of Experimental Psychology Learning, Memory, and Cognition*, *40*(5), 1307–1320. http://doi.org/10.1037/a0036822

Vitela, A. D., Monson, B. B., & Lotto, A. J. (2015). Phoneme categorization relying solely on high-frequency energy. *The Journal of the Acoustical Society of America*, *137*(1), EL65–EL70. http://doi.org/10.1121/1.4903917

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

Weinberger, N. M. (2007). Auditory associative memory and representational plasticity in the primary auditory cortex. *Hearing Research*, *229*(1-2), 54–68. http://doi.org/10.1016/j.heares.2007.01.004

Wellmann, C., Holzgrefe, J., Truckenbrodt, H., Wartenburger, I., & Höhle, B. (2012). How each prosodic boundary cue matters: evidence from german infants. *Frontiers in Psychology*, *3*, 580. http://doi.org/10.3389/fpsyg.2012.00580

Winkowski, D. E., Bandyopadhyay, S., Shamma, S. A., & Kanold, P. O. (2013). Frontal Cortex Activation Causes Rapid Plasticity of Auditory Cortical Processing. *Journal of Neuroscience*, *33*(46), 18134–18148. http://doi.org/10.1523/JNEUROSCI.0180-13.2013

Woods, D. L., Alain, C., Diaz, R., Rhodes, D., & Ogawa, K. H. (2001). Location and frequency cues in auditory selective attention. *Journal of Experimental Psychology Human Perception and Performance*, *27*(1), 65–74.

Wright, B. A. (2005). Combined representations for frequency and duration in detection templates for expected signals. *The Journal of the Acoustical Society of America*, *117*(3), 1299–1304. http://doi.org/10.1121/1.1855771

Yaron, A., Hershenhoren, I., & Nelken, I. (2012). Sensitivity to Complex Statistical Regularities in Rat Auditory Cortex. *Neuron*, *76*(3), 603–615. http://doi.org/10.1016/j.neuron.2012.08.025

Yin, P., Fritz, J. B., & Shamma, S. A. (2014). Rapid Spectrotemporal Plasticity in Primary Auditory Cortex during Behavior. *Journal of Neuroscience*, *34*(12), 4396–4408. http://doi.org/10.1523/JNEUROSCI.2799-13.2014

Ylinen, S., Uther, M., Latvala, A., Vepsäläinen, S., Iverson, P., Akahane-Yamada, R., & Näätänen, R. (2010). Training the brain to weight speech cues differently: a study of Finnish second-language users of English. *Journal of Cognitive Neuroscience*, *22*(6), 1319–1332. http://doi.org/10.1162/jocn.2009.21272

Zevin, J. D. (2012). A sensitive period for shibboleths: the long tail and changing goals of speech perception over the course of development. *Developmental Psychobiology*, *54*(6), 632–642. http://doi.org/10.1002/dev.20611

Zion Golumbic, E. M. Z., Poeppel, D., and Schroeder, C. E. (2012). Temporal context in speech processing and attentional stream selection: a behavioral and neural perspective. *Brain & Language*, *122*, 151-161.