Acoustic Theories of Speech Perception

Melissa A. Redford[a] & Melissa M. Baese-Berk

Department of Linguistics, University of Oregon, Eugene OR 97405

Approved for publication on March 3, 2023

Summary

Acoustic theories assume that speech perception begins with an acoustic signal transformed by auditory processing. In classical acoustic theory, this assumption entails perceptual primitives that are akin to those identified in the spectral analyses of speech. The research objective is to link these primitives with phonological units of traditional descriptive linguistics via sound categories and then to understand how these units/categories are bound together in time to recognize words. Achieving this objective is challenging because the signal is replete with variation making the mapping of signal to sound category nontrivial. Research that grapples with the mapping problem has led to many basic findings about speech perception, including the importance of cue redundancy to category identification and of differential cue-weighting to category formation. Research that grapples with the related problem of binding categories into words for speech processing motivates current neuropsychological work on speech perception. The central focus on the mapping problem in classical theory has also led to an alternative type of acoustic theory, namely, exemplar-based theory. According to this type of acoustic theory, variability is critical for processing talker-specific information during speech processing. The problems associated with mapping acoustic cues to sound categories is not addressed because exemplar-based theories assume that perceptual traces of whole words are perceptual primitives. Smaller units of speech sound representation as well as the phonology as a whole, are emergent from the word-based representations. Yet, like classical acoustic theories, exemplar-based theories assume that production is mediated by a phonology that has no inherent motor information. The presumed disconnect between acoustic and motor information during perceptual processing distinguishes acoustic theories as a class from other theories of speech perception.

Keywords: acoustic features; acoustic variability; cue weighting; lexical representations; linguistic features; perceptual learning; phonemes; speech sound categories; socio-indexicality; temporal integration

1. Introduction

Speech sound is the airborne transmission of a disturbance initiated with exhalation through the glottis and shaped by a vocal tract transformed through articulatory movements. The disturbance is

a) Address correspondence to M.A. Redford at redford@uoregon.edu

characterized as a continuously changing complex waveform. When it reaches the ear, the disturbance is transferred via middle ear mechanics to the fluid-filled cochlea where it undergoes a spectral analysis. The now fluid-borne disturbance sets up traveling waves along the basilar membrane, which peak at different amplitudes and locations corresponding to the amplitudes and frequencies of the complex waveform. The peaks displace the cilia of mechanoreceptors that ride atop the membrane, transducing a movement-based spectral analysis into neural firings. These firings are carried along the 8th nerve into the brain where sound is represented tonotopically in primary auditory cortex, preserving the analysis begun in the periphery. Acoustic theories treat the spectral analysis of early auditory processing as relevant to speech perception; in fact, they take the analysis to indicate that perception needs to be understood with reference to speech acoustics. Thus, the focus of these theories is how to connect the acoustic signal to linguistic representation. This challenge has been solved in different ways, resulting in different types of acoustic theory. In this chapter, we identify two types of theory that follow from different assumptions about the primary representations of speech perception and review the research pursued under each type.

2. Types of Acoustic Theory

Speech perception obviously relies on speech acoustics as an information source; there is no theory of perception that overlooks this fact. What is disputed is the extent to which speech perception should be understood as arising from auditory processing versus from the integration of perceptual and motor information about speech. Acoustic theories assume that speech perception begins with auditory processing (Diehl et al., 2004; Nearey, 1990; Ohala, 1996; Peelle, 2019; Pisoni, 1977; Poeppel & Monahan, 2008; Stevens & Klatt, 1974; Yi et al., 2019); other theories of perception, including the Motor Theory (Liberman & Whalen, 2000; Libermann et al., 1967; Libermann & Mattingly, 1985), the Direct Realist Theory (Fowler, 1986), and the Perception-for-Action-Control Theory (PACT; Schwartz et al., 2012), assume that it begins with the integration of sensory and motor information. This difference highlights another, equally fundamental one: acoustic theories assume that speech perception and production are independent systems mediated by a separately-constituted linguistic system. By contrast,

a) Address correspondence to M.A. Redford at redford@uoregon.edu

integration theories posit linguistic representations of speech that reference the dynamics of articulation (i.e., Motor Theory/Direct Realist Theory) or are functionally-linked to perceptual-motor units common to the speech perception and production system (i.e., PACT). This difference between acoustic and integration theories underscores the centrality of the perception—production relationship to speech. A complete theory of speech perception must make obvious how it interfaces with speech production. For acoustic theories, this entails linking perception to production via linguistic representation.

Although there is no single acoustic theory of speech perception, it is possible to identify two main types: classical theory and exemplar-based theory. In classical acoustic theory, the linguistic representations that mediate between perception and production are the phonological units of traditional descriptive and generative grammars (Chomsky & Halle, 1968; Jakobson et al., 1951; Pike, 1947). In contrast, exemplar-based acoustic theories embrace the fully complexity of the signal, taking their inspiration from exemplar approaches to memory and classification (Hintzman, 1986; Medin & Schaffer, 1978; Nosofsky, 1986). In exemplar-based theory, structure emerges from similarity relations between individual examples (exemplars) of experienced items. The perceptual traces of these items are stored in a multidimensional perceptual space. The relevant items in speech perception are words, each experienced example of which is stored in a lexicon organized according to the acoustic dimensions of the perceptual space. An emergent phonology from the lexicon then governs production (Pierrehumbert, 2002).

The different representational assumptions of the two main types of acoustic theory have implications for the research questions that motivate discovery and knowledge accrual. In classical acoustic theory, speech perception proceeds via the listener's identification of phonemes and distinctive features. Research from this perspective has therefore focused on mapping acoustic features derived from the spectral analysis of speech onto discrete linguistic units of sound (Diehl et al., 2004). This mapping is complex and problematic because of a lack of invariance between the acoustics of speech and discrete linguistic units. The lack of invariance problem arises because speech sounds are always articulated in the context of other sounds. The context-dependent articulation of sound, or coarticulation, is reflected in

a) Address correspondence to M.A. Redford at redford@uoregon.edu

speech acoustics. It results in the "same" sound having a different spectral shape in different contexts. By contrast, discrete linguistic units are symbolic units that are thought to remain the same regardless of context. In a later section, we review some of the fundamental discoveries that have followed from attempts to solve the lack of invariance problem. We also review research based on other sources of variation in the signal and the problem of how the perceiver manages to bind meaningless units of perception into larger, meaningful ones (McClelland & Elman, 1986; Norris et al., 2000; Yi et al., 2019).

Exemplar-based theory side-steps the lack of invariance problem by embracing variability in the form of detailed lexical representations (Casserly & Pisoni, 2010). For example, Goldinger (1998) successfully simulated the facilitative effect of different speakers' voices on listeners' memory for real words using a pure version of an exemplar model, Hintzman's MINERA 2 model (1986); one that assumes no abstract representations beyond those built up from individual traces of experienced perceptual objects and events. Johnson (1997) modeled the emergence of phonological structure from an exemplar-based lexicon, represented as a set of remembered spectral sequences attached to word nodes. The individual spectral sequences were distributed along acoustically-defined axes in a multidimensional space. This distribution entails an organization based on acoustic-phonetic similarity. Given this organization, Johnson showed how speech sounds and metrical structure could be emergent properties in an exemplar-based theory of perception. Consistent with pure exemplar theory, he characterized phonological structure as a fleeting phenomenon that emerged and then disappeared with word recognition.

Unlike classical acoustic theory, a pure exemplar-based theory is not easily aligned with mainstream theories of speech production (Dell et al., 1997; Goldrick, 2006; Guenther, 2016; Levelt, 1989; Roelofs, 1999). After all, these theories also assume a fundamental role for phonological units of traditional descriptive and generative grammars. By and large, the problem of interfacing with speech production has not impacted research on speech perception from an exemplar-based acoustic theoretic perspective. And the problem of how to account for speech production given an exemplar-based theory of

a) Address correspondence to M.A. Redford at redford@uoregon.edu

perception remains unsolved. Pierrehumbert (2002; 2016) acknowledged this and proposed a solution. She argued for a theory of speech perception—production that incorporates lexical representations with fine acoustic and contextual detail, consistent with an exemplar-based approach to perception. On top of the exemplar layer of representation, there is another stable layer of representation that is emergent from the first but codes abstract phonological structure. The phonological representations are then used to explain spoken language behavior, includingthe regularity with which a sound change sweeps across an individual's lexicon. Pierrehumbert's argument for phonological representation is that the multidimensional perceptual space within which exemplars reside will contain experience-based discontinuities that would disrupt the sweeping generalizations that are observed in speech production. Overall, Pierrehumbert's argument echoes one that exists in the psychological literature where it is asserted that hybrid rule-and-exemplar models are better able to explain generalization than exemplar-only models (see, e.g., Denton et al., 2008; Erickson & Kruschke, 1998).

In sum, acoustic theories assume that speech perception begins with the spectral analysis of the waveform that is a feature of early auditory processing. Theories then differ on how to connect the result of this processing to representations that allow for language comprehension and production. Classical acoustic theory assumes that perceptual primitives can be mapped on to discrete phonological units. Exemplar-based theory assumes that the perceptual detail is stored as acoustic-phonetic exemplars of words. Classic acoustic theory easily interfaces with mainstream theories of production; exemplar theory requires additional assumptions to do so. Research conducted from a classical acoustic theoretic perspective is largely aimed at overcoming variability in the signal that confuses the mapping between acoustic features and linguistic units. Research conducted from an exemplar-based acoustic theoretic perspective embraces this variability to show that it accounts for important behaviors that cannot be explained within the classical theory. In the following sections, we review the major lines of research in speech perception that have been motivated by the two main types of acoustic theory. Our goal is to celebrate the breadth of phenomena that continue to be explored given an acoustic theoretic view of

a) Address correspondence to M.A. Redford at redford@uoregon.edu

speech perception and to emphasize how theory-driven research contributes to fundamental knowledge about speech.

3. Research Motivated by Classical Acoustic Theory

Early research within the classical acoustic theoretic framework focused on finding structure in the acoustic signal that maps onto distinctive features and phonemes. Current research focuses on the neurophysiological underpinnings of perception. Both broad lines of research are motivated by the conceptual problems that ensue from assuming that discrete, context-independent sublexical units are fundamental to speech perception. These problems and their proposed solutions are reviewed below.

The Lack of Invariance Problem

Individual speech sounds that listeners categorize as the same consonant or vowel are produced differently across different phonological contexts. Naturally, with production differences come acoustic differences that obscure the direct mapping between a sound and its perceived segmental category. In classical theory where these categories mediate access to linguistic units, the variability with which speech sounds are produced presents a real problem for perception. Attempts to solve this lack of invariance problem have resulted in the discovery of categorical perception and in a deeper understanding of articulatory-acoustic relations, as well as in neuroscientific hypotheses that are still under investigation.

In the 1950s, researchers at Haskins Laboratories in New Haven, Connecticut invented the first speech synthesis technique (Cooper et al., 1951) and used it to explore possible invariant cues to category perception in the time-varying acoustic waveform. They focused their initial investigations on formant transitions, especially on the second formant (F2) onset to midpoint in stop-vowel sequences. The hypothesis was that F2 originated at a "locus" in frequency space that corresponded to place of articulation even if F2 itself was not necessarily present in the acoustic signal at that locus (Delattre et al., 1955). The hypothesis led to the discovery that listeners perceive continuous changes to F2 onsets categorically when F2 midpoints are held constant (Liberman et al., 1957). Specifically, listeners' stop consonant identification with respect to place of articulation was found to shift abruptly at certain points

a) Address correspondence to M.A. Redford at redford@uoregon.edu

along a continuum when F2 was varied from a lower to higher frequency at the onset to the same vowel; within stimulus discrimination was also better at these points than between them. This set of results is known as categorical perception. Since the absolute value of an F2 onset will vary with the value of the F2 midpoint which varies with vowel quality (e.g., [i] versus [u]), Liberman and colleagues proposed that listeners recover mostly hidden loci in frequency space by recovering the more nearly categorical articulatory commands (e.g., alveolar versus bilabial; Liberman et al., 1967). This proposal is at odds with the view that the object of perception is the acoustic signal itself. And so those who adopt an acoustic theoretic view of speech perception have sought a different explanation for categorical perception: one where the cues to stop place of articulation are found in the signal itself.

Lindblom (1963) approached the specific problem of variable F2 onsets by looking for systematicity in the variable signal, not for a single cue to place of articulation. He found that while the steepness and direction of an F2 transition from onset to midpoint depends on the vowel target, the relationship between these two points varies systematically such that the slope of a best-fit line through points associated with a single place of articulation for stop consonants and multiple vowels captures consonant-vowel (CV) coarticulation specific to that place of articulation. Sussman (1989) connected this observation to findings from neuroethological investigations of auditory localization to suggest an auditory neuroscience hypothesis for stop place of articulation; namely, that frequency-specific neurons arranged in slope-defined arrays within primary auditory cortex selectively respond to F2 transitions in such a way as to solve the lack of invariance problem, at least for stop place of articulation. This focus on neural mechanisms as explanation for category perception foreshadows a later and ongoing surge in auditory neuroscientific studies motivated by a classical acoustic theoretic view of speech perception (see below).

Other attempts to solve the lack of invariance problem within an acoustic theoretic perspective have focused on acoustic correlates of linguistic (distinctive) features. A search for invariant correlates was led by Blumstein and Stevens in the 1970s and 1980s. The team reported invariant characteristics in

a) Address correspondence to M.A. Redford at redford@uoregon.edu

the spectral shape of stop releases (Blumstein & Stevens, 1979, 1980, 1981; Stevens & Blumstein, 1975) and fricatives (Stevens et al., 1992) for place of articulation. Their search for invariant correlates of features was based on two assumptions (Stevens & Blumstein, 1981): the first was that the relationship between articulation and acoustic physics is such that there is acoustic stability across different constriction locations in some regions of the vocal tract as well as moments of rapid acoustic transition across constriction locations in other regions (i.e., the Quantal Theory of speech production; Stevens, 1989); the second was that perceptual processing is similarly quantized, resulting in the type of discontinuities that give rise to categorical perception (Stevens, 1981; Stevens & House, 1972). Both assumptions have shaped subsequent work in the field; for example, the first has given rise to sophisticated models of vocal tract acoustics (e.g., Honda et al., 2010; Mrayati et al., 1988), and the second to an interest in the specifics of early auditory processing of speech sound as a means to understand the psychophysical underpinnings of speech perception (e.g., Delgutte & Kiang, 1984; Ghitza, 1995; Kluender, 1994; Seneff, 1988).

Of course, acoustic correlates of linguistic features are only useful if they are present in the signal. And some putatively invariant cues, such as the spectral shape of stop bursts identified by Blumstein and Stevens (1979; 1980), are frequently not available. In fact, stop consonants are rarely released at the ends of words in running speech, which means that they have no bursts with which to convey place of articulation. Despite this, listeners are perfectly able to distinguish words such as *bad* from *bag* in the speech stream. This ability argues against an approach to the lack of invariance problem that assumes a single invariant cue to every linguistic feature; instead, it supports an approach that assumes cue redundancy in the signal (e.g., Bailey & Summerfield, 1980; Mann & Repp, 1980).

There is ample evidence for cue redundancy to speech sound categories in the speech signal. Consider, for example, the two cues to stop place of articulation discussed above. Current approaches to speech sound perception assume both cue redundancy and the integration of these cues to derive speech sound categories (e.g., Bailey & Summerfield, 1980; Toscano & McMurray, 2010). The approach has

a) Address correspondence to M.A. Redford at redford@uoregon.edu

motivated studies of perceptual learning where language-specific patterns of acquisition are understood as differential cue weighting (e.g., Guion & Pederson, 2007; Lim & Holt, 2011; Mayo et al., 2003; Nearey, 1997; Toscano & McMurray, 2010). For example, modeling data (e.g., Toscano & McMurray 2010) demonstrate that categories can be learned using cues weighted as a function of their reliability. That is, simulations using a mixture of Gaussian models are able to capture the trading relations among cues that are evidenced in human behavior. The statistics of the input appear to be sufficient to derive the appropriate cue weights that humans demonstrate. Critically, however, statistics alone are not sufficient to capture cue weighting. Learning is therefore an important component of the simulations; specifically, the history and structure of the learning system critically impacts how speech category learning takes place.

In behavioral data, the relative weighting of cues shifts as a function of the reliability of these cues. For example, Lim & Holt (2011) demonstrate that when a more-preferred cue (e.g., F2 for discrimination of liquids by Japanese learners of English) is less reliable (i.e., more variable), listeners are more likely to strongly weight a less-preferred cue (e.g., F3 for the discrimination of liquids). Critically, this shift in cue weighting happens quickly and without explicit instruction. Neuroscientific studies have attempted to investigate the neural underpinnings of this type of category learning in order to account for the behavioral results (Lim et al., 2014; Feng et al., 2019; Yi et al., 2016). Among other structures, the basal ganglia and the left superior temporal gyrus (LTSG) have been implicated in learning. The basal ganglia, for example, has been long-understood to be crucial for learning with explicit feedback of non-speech categories, but its role in speech learning has not been investigated until more recently (Lim et al., 2014). Similarly, the LTSG and the broader auditory corticostriatal circuitry have been shown to be crucial for mediating acquisition of categories with appropriate cue weights (Yi et al., 2016).

The cue-weighting accounts of category learning can be contrasted with other accounts that focus on feature detection and the representation of these features in the brain. The bulk of the latter work has used electrocorticogram recordings, using intracranial measurements, to investigate neural responses to auditory stimuli. Some of this work has demonstrated that specific brain areas (e.g., posterior STG;

a) Address correspondence to M.A. Redford at redford@uoregon.edu

pSTG) are organized according to linguistic features (e.g., Chang et al., 2010). For example, both F2 onset frequency (which correlates with place of articulation) and F2 formant transitions (which correlates with non-coronal speech sounds) are robustly represented in the pSTG, suggesting that features may be extracted and represented in the brain. Similarly, manner of articulation also appears to be directly represented in the STG (Mesgarani et al., 2014).

To summarize, the lack of invariance problem has motivated research into speech perception for nearly three-quarters of a century. Early work focused on identifying acoustic cues to distinctive features and phonemes. Evidence for multiple cues suggested sound categories that then link to phonemes. The emphasis on categories suggests perceptual learning, specifically the perceptal learning of relative cue weighting. An interest in learning motivates one line of research into the neural mechanisms that underlie speech perception. The original interest in feature detection based on a spectral analysis of the signal has also persisted and motivates another line of research into central neural responses to acoustic features.

Speaker and Rate Normalization

The lack of invariance problem arises due to coarticulation. But context-dependency is not the only source of variability in the signal that affects segmental acoustics. Other important sources of variability are the speaker's vocal tract and the default rate at which an individual speaks. These sources of variability must also be resolved for listeners to hear the same sound when that sound is produced by different speakers, speaking at different rates.

One source of speaker variability is vocal tract morphology. Different morphologies give rise to different segmental acoustics, including to those that would seemingly distort the phonetic-phonological characteristics of the segment itself (e.g., child versus adult vowel formant frequencies; Peterson & Barney, 1952). Relatedly, different fundamental frequencies will affect the shape of the sound spectrum even in the same vocal tract (Fant, 1970). In classical acoustic theory, speaker variability has been treated separately from context-dependent variability because the speaker effect is thought to apply equally to all speech sounds. If this is the case, then listeners may be able to apply a simple transformation to the signal

a) Address correspondence to M.A. Redford at redford@uoregon.edu

before continuing with sound category / phoneme identification (see, e.g., Ladefoged & Broadbent, 1957; Mullennix et al., 1989; Nygaard et al., 1994). The specific hypothesis is that listeners extract information about the speaker based on characteristics of the signal and then use this information to establish the speaker-specific acoustic cues to speech sound categories and their corresponding context-independent phonological representation (Johnson, 2005; Nusbaum & Magnuson, 1997; Pisoni, 1997). Behavioral evidence is consistent with the hypothesis (Choi et al., 2018; Magnuson & Nusbaum, 2007; Mullennix & Pisoni, 1990). For example, Choi et al. (2018) showed that speeded classification of phonetically similar words (e.g., *boot* vs. *boat*) was always slower when the words were produced by multiple talkers than when produced by a single talker. The effect, always present in the results, increased with the acoustic-phonetic similarity of the words in question. The results are interpreted to reflect the processing costs associated with talker normalization.

Individual speakers not only have different vocal tract morphologies, they also speak at different default rates (Bradlow et al, 2017; Kendall, 2013; Tsao et al., 2006). The different speaking rates affect the absolute duration of individual speech segments. This is problematic for classical acoustic theory because duration is often as important an acoustic cue to phonemic contrast as spectral and amplitude changes, even when the contrasts in question are not specifically temporal (Klatt, 1976). For example, the distinction between the vowels in British English words *who'd* and *hood* are as confusable when their duration is altered as when their formant structure is altered (Ainsworth, 1972). Like the problem introduced by different vocal tract morphologies, the problem of speaking rate can be solved if the signal can be normalized. In particular, a listener may be able to use the average speaking rate (e.g., in syllables per second) to calibrate their perception of segmental duration. Again, behavioral evidence is consistent with the hypothesis. Studies that investigate the effect of speech rate manipulations on segment (Miller & Volaitis, 1989; Summerfield, 1975), syllable (Baese-Berk et al., 2019), and word (Dilley & Pitt, 2010) identification show that perceptually ambiguous stretches of speech are perceived differently as a function of the speech rate context: when the ambiguous portion of speech is held constant and the rate of adjacent

a) Address correspondence to M.A. Redford at redford@uoregon.edu

speech is either increased or decreased, the listener will either hear or not hear the segment of interest (e.g., *Don must see the harbor or boats* is perceived as *Don must see the harbor boats*). Results such as these are taken as evidence that listeners track speech rate and use this information to interpret the sounds they hear.

In sum, variability that is due to global speaker effects on the signal is treated differently from variability due to context-dependent articulation. Within a classical acoustic theoretic approach to perception, the listener is hypothesized to normalize the signal across speakers so that correspondences between acoustic cues and phonological representations can be identified.

Temporal Binding Problem

Another problem for the classical acoustic theory of speech perception is the problem of how to bind acoustic cues to features and phonemes into percepts that allow for communication (e.g., words). While temporal integration is known to occur at early stages of auditory processing, it occurs only at time scales that could account for the percept of phonemes or syllables (Shamma, 2003). At some point during perception, the listener must be able to integrate information across larger temporal intervals to extract words from the speech stream. At a minimum, the temporal ordering of detected features/phonemes much be preserved. Early attempts to explain the integration process assumed grouping by similarity following a type of Auditory Scene Analysis (Bregman, 1994). For example, Darwin (1997) suggested that listeners use continuity of pitch and spatial location to render a speech stream coherent; at the same time, the listener applies top-down knowledge to identify meaningful patterns in the frequency modulated signal. When applied to the full spectrum of speech perception, however, Gestalt grouping principles based on bottom-up processes fail in a number of ways to account for stable percepts. For example, Remez and colleagues (1994) make the point that there is little coherent or similar across intervals of speech. They argue that, even if a listener might be able to use pitch and the timbre of someone's voice to track their speech, the context-dependent variability of individual sounds, coupled with the interruption of continuous formant trajectories due to consonantal articulation, leave little that would allow a listener to

a) Address correspondence to M.A. Redford at redford@uoregon.edu

automatically group particular sounds together into word units. Further, Remez (2003) notes that solving the problem with reference to top-down knowledge only begs the question of how that knowledge might be acquired in the first place.

More recently, the issue of temporal integration has been taken up in neuroscientific investigations of speech perception. For example, Chang and colleagues (Yi et al., 2019) have suggested that the superior temporal gyrus (STG) is parcellated into regions that track different temporal landmarks in speech. Building on the idea of STG as part of the higher-order associative auditory cortex, they propose that the neural populations that serve as acoustic-phonetic feature detectors (e.g., place of articulation) are integrated at more local and global timescales by the differential temporal sensitivity of middle-to-anterior and posterior regions of the STG.

Of course, neural theories of temporal integration will vary to some extent with the tools that are used to investigate speech processing. No doubt that one reason Chang and colleagues have suggested a largely spatial explanation for the temporal binding of features into larger units is because their studies rely on electrocorticography (ECoG), which uses electrodes that are placed directly on the surface of the exposed brain to preoperatively assess function across different cortical regions in patients with severe drug-resistant epilepsy who are undergoing focal resection treatment. Alternate theories have emerged with different neuroscientific techniques. For example, Poeppel (2003) proposed the asymmetric sampling in time (AST) hypothesis to suggest that "auditory signals in general are quantized in the time domain" differently by different neuronal populations, which are distributed differently across the two hemispheres (p. 246). The smaller and larger temporal windows of integration (quantization) that results, say, in the percept of phones versus prosodic words, are assumed to be reflected oscillatory neural activity at a variety of frequency bands (i.e., time scales; Poeppel, 2003). The hypothesis makes sense of a body of findings based on electroencephalography (EEG) and magnetoencephalography (MEG), which record the unfolding of neural activity at extremely high temporal resolutions, but with significantly less spatial resolution than ECoG. Interestingly, recent ECoG work has suggested that even contrasts typically

a) Address correspondence to M.A. Redford at redford@uoregon.edu

represented by timing cues (e.g., VOT) are transformed into a spatial code in auditory cortext (Fox et al., 2020).

Summary

The classical acoustic theory of speech perception assumes a fundamental role for distinctive features and phonemes in speech perception. This assumption gives rise to two broad problems: the problem of variability for mapping acoustic features onto speech sound categories and the problem of integration over time. Whereas the problem of temporal integration exists for all theories of perception at some level — the speech signal is, after all, a time-varying signal that must be resolved at some point into meaning, which is atemporal — the mapping problem is only a problem for theories of perception that assume a primary role for features or phonemes. In the next section, we consider research motivated by an exemplar-based theory, which treats acoustic variability as an important and valuable source of information in speech processing.

4. Research Motivated by Exemplar-Based Theory

Whereas classical acoustic theory treats acoustic variability as a problem to overcome, exemplar-based theory embraces acoustic variability as an important source of information in speech perception. Research within an exemplar-based theoretic perspective largely ignores context-dependent variability in the signal, since lexical representations are primary, and instead focuses on mainly on speaker variability and its contributions to the perceptual processing of speech.

Variability Across Speakers

Exemplar-based theory recognizes that the specific acoustic characteristics of an individual speaker's voice conveys crucial social information and other language-independent information during communication (Perrachione et al., 2009; Winters et al., 2008). Moreover, the evidence suggests that his information can alter segmental perception in a variety of tasks (Eisner & McQueen, 2005; Kraljic & Samuel, 2005, 2007; Ladefoged, 1978; Ladefoged & Broadbent, 1957). For example, Kraljic and colleagues (2008) demonstrate that perception of an ambiguous phoneme (e.g., a speech sound between

a) Address correspondence to M.A. Redford at redford@uoregon.edu

/s/ and /ʃ/) is perceived differently depending on whether the use of this ambiguous phoneme can be attributed to an individual difference (i.e., idiolect) or is a feature of the speaker's dialect.

More generally, listeners demonstrate improved intelligibility for familiar talkers (Nygaard et al., 1994; Nygaard & Pisoni, 1998). These talker familiarity effects suggest that variation between talkers is tracked and represented by listeners, a prediction that is not compatible with the classical acoustic theory of perception. Indeed, discoveries of facilitiative talker effects on speech perception were an original motivation for exemplar-based theories (see Casserly & Pisoni, 2010). The talker normalization assumed in classical acoustic theory leaves no clear reason why perception of a familiar talker would be easier than perception of a novel talker. Specifically, if listeners are asked to make judgments about phonological features or structure, including transcription of whole words during recognition (i.e., intelligibility), classical acoustic theory predicts that listeners have already "abstracted away" from the talker-specific details that are not directly relevant to perception of these phonological features.

Exemplar-based theory accounts for positive effects of speaker variability by positing that the forms produced by individual speakers are stored as perceptual traces in memory. The traces can then be indexed for speaker characteristics, which allows for the emergence of socio-indexical information (see below). Also, because talker information is integrated with linguistic representations via traces in memory, a familiar talker will have more robust traces, thus improving perception of speech produced by that talker. Exemplar-based theory has accounted for the processing costs associated with new talkers by positing that newer traces (i.e., more recently heard exemplars) carry more weight than older exemplars. Talker switch costs are explained as due to the reduced activation of traces from memory as successive speech samples from an old talker then new talker are less similar to one another than when the successive samples are produced by the same talker (but see Magnuson et al. (2021) for evidence that current exemplar-based theory cannot account for all talker familiarity results).

Of course, exemplar-based theory must also account for the experience of phonetic constancy; that is, for the experience that an /s/ is perceived as an /s/ regardless of speaker. Whereas this constancy is

a) Address correspondence to M.A. Redford at redford@uoregon.edu

fundamental to the assumption of feature/phoneme detection in classical acoustic theory, it must be abstracted from more detailed representations in exemplar-based theory. On approach is to statistically sample the space of talker characteristics in the traces which are stored in memory. But, there are also hybrid approaches. For example, Kleinschmidt (2019) presents models for quantifying variability across talkers, and examines how useful such variability might be for recognizing segmental features across talkers. This work uses a computational framework (the Ideal Adapter Framework; Kleinschmidt & Jaeger, 2015) to combine theories of normalization over variability with theories that posit that (some) relevant socio-indexical information is stored with speech input (Kleinschmidt et al., 2018).

Socio-Indexicality

Because exemplar-based acoustic theory handles speaker variability so well, this perspective has supercharged interest in the perception of socio-indexical information — something that research from a classical acoustic theoretic perspective has largely ignored. Indexical information, broadly speaking is the non-linguistic social information that co-occurs with linguistic structure (Abercrombie, 1967) — information such as gender, age, socio-economic status and other individually identifying information, including that which may be context specific. In the past 15 years, researchers have discovered exactly how much this information about the speaker impacts speech perception (e.g., Hay et al., 2006; McGowan & Babel, 2020; Niedzielski, 1999; Strand & Johnson, 1996; Sumner, 2015). To take just one example: listeners perceive vowels differently depending on who they believe is producing those vowels (Hay et al., 2006); that is, identical acoustics are perceived differently depending on their perceived source. Exemplar-based theory not only provides a framework for integrating socio-indexical information with other linguistic information, it entails that such information is co-processed and stored during perception.

In addition to behavioral work that suggests joint processing of linguistic and non-linguistic information during perception, neuroimaging work has shown that linguistic and indexical information (i.e., 'what' and 'who' information about the speech) is tightly integrated during processing; for example, in the posterior left middle temporal gyrus (Chandrasekaran et al., 2011). Relatedly, Kreitewolf et al.

a) Address correspondence to M.A. Redford at redford@uoregon.edu

(2014) have suggested that interdependencies between speaker and speech processing are correlated with interactions between the right and left hemisphere: when recognizing linguistic prosody from different speakers, the functional connections between right and left Heschl's gyri are activated, suggesting a deep interconnection between speaker information and linguistic information.

Foulkes (2010) lays out the implications of the exemplar-based perspective for our understanding of language processing and acquisition, stating that "one of the attractions of exemplar theory is its capacity to predict and model learning of linguistic and non-linguistic structures through the same mechanism" (p. 15). Other work has suggested that such models can account for learning both lexical and socio-indexical information (e.g., Munson et al., 2011) and for types of language change that are typically relegated to socio-phonetic investigations (e.g., near-mergers; Nycz, 2011; 2015).

In pure exemplar-based theory, lexical representations are a subset of more general auditory memory representations. This view suggests that non-speech information may be co-stored (or at least co-processed) with linguistic information. And there is evidence for this suggestion: for example, Pufahl & Samuel (2014) demonstrated that environmental sounds co-produced during speech perception impact perception in similar ways to socio-indexical information. Computational models of speech perception have frequently adopted this purer view of the lexicon, by proposing models that do not require any sub-lexical levels of representation (e.g., dynamic cohort model: Gaskell & Marslen-Wilson, 2002; the lexical access from spectra model: Klatt, 1989). In fact, in many of these models, sub-lexical representations are disruptive to successful recognition (e.g., Gaskell & Marslen-Wilson, 2002).

Despite these successes, hybrid models of perception have also been proposed under the assumption that sub-lexical representations are needed to account for spoken language behavior. Early, it was noted that phoneme-level representations allow exemplar-based theory of perception to interface with mainstream theories of speech production. But, hybrid models have also been proposed to account for the same problems that motivate research from a classical acoustic theoretic perspective. For example, hybrid models have been proposed to solve the "lack of invariance" problem. The specific suggestion is that

a) Address correspondence to M.A. Redford at redford@uoregon.edu

listeners may use social categories to separate sounds that are acoustically identical into distinct linguistic categories (Pierrehumbert, 2016; Sumner et al., 2014). In these models, both linguistic information and relevant socio-indexical information is stored in the lexicon and abstract information about speech sounds (e.g., phonemes and features) are emergent from the lexicon, but also stored separately and so psychologically real.

Summary

The discovery of talker-specific effects on perception led to an alternative exemplar-based acoustic theory of perception, which has in turn led to research on the perception of social-indexical information in speech. But even though the importance of extra-linguistic information to speech processing appears settled, questions still remain about how to handle other phenomena within this framework. For example, it is not entirely clear how listeners are able to form categories within the socio-indexical domain for factors including age, gender, race, sexual orientation, and the intersections between these and many other factors. Thus, while an exemplar-based theory of perception provides a framework for research into the effects of socio-indexicality on speech, there is still substantial work to be done to account for the interactions between linguistic and socio-indexical information within this framework.

5. Conclusion

In this chapter, we identified two types of acoustic theory of speech perception based on their assumptions regarding linguistic representation and how the speech signal interfaces with these representations. We then showed how these assumptions have driven speech perception research in different directions. The underlying theories will continue to evolve with research findings, but the assumption that speech perception is based on auditory processing of an acoustic signal will remain. This most fundamental assumption of acoustic theories provides a useful complement to the contrasting assumption that speech perception begins after sensory-motor integration is achieved. Just as classical and exemplar-based acoustic theory have promoted research that covers separate ground, so too do the

a) Address correspondence to M.A. Redford at redford@uoregon.edu

acoustic and integration theories of perception. We encourage the reader to explore these differences by reading the chapter in this volume on motor theories of speech perception.

Further Reading

Casserly, E. D., & Pisoni, D. B. (2010). Speech perception and production. Wiley Interdisciplinary Reviews: Cognitive Science, 1(5), 629-647.

Diehl, R. L., Lotto, A. J., & Holt, L. L. (2004). Speech Perception. *Annual Review of Psychology*, *55*(1), 149–179.

Hickok, G., & Poeppel, D. (2016). Neural basis of speech perception. In G. Hickok & S.L. Small (ed), *Neurobiology of language* (pp. 299-310). San Diego: Academic Press.

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

Samuel, A. G. (2011). Speech perception. *Annual Review of Psychology*, 62, 49-72.

References

Abercrombie, D. (1967). *Elements of General Phonetics*. Edinburgh University Press.

Ainsworth, W. A. (1972). Duration as a cue in the recognition of synthetic vowels. *The Journal of the Acoustical Society of America*, *51*(2B), 648–651.

Baese-Berk, M. M., Dilley, L. C., Henry, M. J., Vinke, L., & Banzina, E. (2019). Not just a function of function words: Distal speech rate influences perception of prosodically weak syllables. *Attention, Perception, & Psychophysics*, *81*(2), 571–589. https://doi.org/10.3758/s13414-018-1626-4

Bailey, P. J., & Summerfield, Q. (1980). Information in speech: Observations on the perception of [s]-stop clusters. *Journal of Experimental Psychology: Human Perception and Performance*, *6*(3), 536.

a) Address correspondence to M.A. Redford at redford@uoregon.edu

Blumstein, S. E., & Stevens, K. N. (1979). Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *The Journal of the Acoustical Society of America*, *66*(4), 1001–1017.

Blumstein, S. E., & Stevens, K. N. (1980). Perceptual invariance and onset spectra for stop consonants in different vowel environments. *The Journal of the Acoustical Society of America*, *67*(2), 648–662.

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

Bradlow, A. R., Kim, M., & Blasingame, M. (2017). Language-independent talker-specificity in first-language and second-language speech production by bilingual talkers: L1 speaking rate predicts L2 speaking rate. *The Journal of the Acoustical Society of America*, 141(2), 886-899.

Bregman, A. S. (1994). *Auditory scene analysis: The perceptual organization of sound*. MIT press.

Casserly, E. D., & Pisoni, D. B. (2010). Speech perception and production. Wiley Interdisciplinary Reviews: Cognitive Science, 1(5), 629-647.

Chandrasekaran, B., Chan, A. H., & Wong, P. C. (2011). Neural processing of what and who information in speech. *Journal of Cognitive Neuroscience*, *23*(10), 2690–2700.

Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., & Knight, R. T. (2010). Categorical speech representation in human superior temporal gyrus. *Nature Neuroscience*, *13*(11), 1428.

Choi, J. Y., Hu, E. R., & Perrachione, T. K. (2018). Varying acoustic-phonemic ambiguity reveals that talker normalization is obligatory in speech processing. *Attention, Perception, & Psychophysics*, *80*(3), 784–797. https://doi.org/10.3758/s13414-017-1395-5

Chomsky, N., & Halle, M. (1968). *The sound pattern of English.* Harper and Row.

Cooper, F. S., Liberman, A. M., & Borst, J. M. (1951). The Interconversion of Audible and Visible Patterns as a Basis for Research in the Perception of Speech. *Proceedings of the National Academy of Sciences of the United States of America*, *37*(5), 318–325.

a) Address correspondence to M.A. Redford at redford@uoregon.edu

Darwin, C. J. (1997). Auditory grouping. *Trends in Cognitive Sciences*, *1*(9), 327–333. https://doi.org/10.1016/S1364-6613(97)01097-8

Delattre, P. C., Liberman, A. M., & Cooper, F. S. (1955). Acoustic loci and transitional cues for consonants. *The Journal of the Acoustical Society of America*, 27(4), 769-773.

Delgutte, B. (1980). Representation of speech-like sounds in the discharge patterns of auditory-nerve fibers. *The Journal of the Acoustical Society of America*, *68*(3), 843–857.

Delgutte, B., & Kiang, N. Y. (1984). Speech coding in the auditory nerve: IV. Sounds with consonant-like dynamic characteristics. *The Journal of the Acoustical Society of America*, *75*(3), 897–907.

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological Review*, 104(4), 801.

Denton, S. E., Kruschke, J. K., & Erickson, M. A. (2008). Rule-based extrapolation: A continuing challenge for exemplar models. *Psychonomic Bulletin & Review*, *15*(4), 780–786.

Diehl, R. L., Lotto, A. J., & Holt, L. L. (2004). Speech Perception. *Annual Review of Psychology*, *55*(1), 149–179. https://doi.org/10.1146/annurev.psych.55.090902.142028

Dilley, L. C., & Pitt, M. A. (2010). Altering Context Speech Rate Can Cause Words to Appear or Disappear. *Psychological Science*, *21*(11), 1664–1670. https://doi.org/10.1177/0956797610384743

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Attention, Perception & Psychophysics*, *67*(2), 224–238.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*(2), 107.

Fant, G. (1970). *Acoustic theory of speech production*. Walter de Gruyter.

Feng, G., Yi, H. G., & Chandrasekaran, B. (2019). The role of the human auditory corticostriatal network in speech learning. *Cerebral Cortex*, 29(10), 4077-4089

a) Address correspondence to M.A. Redford at redford@uoregon.edu

Foulkes, P. (2010). Exploring social-indexical knowledge: A long past but a short history. *Laboratory Phonology*, *1*(1), 5–39.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*. http://files.eric.ed.gov/fulltext/ED274022.pdf#page=144

Fox, N. P., Leonard, M., Sjerps, M. J., & Chang, E. F. (2020). Transformation of a temporal speech cue to a spatial neural code in human auditory cortex. *Elife*, 9.

Gaskell, M. G., & Marslen-Wilson, W. D. (2002). Representation and competition in the perception of spoken words. *Cognitive Psychology*, *45*(2), 220–266. https://doi.org/10.1016/S0010-0285(02)00003-8

Gay, T. (1981). Mechanisms in the Control of Speech Rate. *Phonetica*, *38*(1–3), 148–158. https://doi.org/10.1159/000260020

Ghitza, O. (1995). Auditory models and human performance in tasks related to speech coding and speech recognition. In *Modern Methods of Speech Processing* (pp. 401-448). Springer, Boston, MA.

Goldinger, S. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251–279.

Goldrick, M. (2006). Limited interaction in speech production: Chronometric, speech error, and neuropsychological evidence. Language and Cognitive Processes, 21(7-8), 817-855.

Guenther, F. H. (2016). *Neural control of speech*. MIT Press.

Guion, S. G., & Pederson, E. (2007). Investigating the role of attention in phonetic learning. O.-S. Bohn & M. Munro (eds.), *Language experience in second language speech learning* (pp. 57-77). Amsterdam: John Benjamins.

Hay, J., Warren, P., & Drager, K. (2006). Factors influencing speech perception in the context of a merger-in-progress. *Journal of Phonetics*, *34*(4), 458–484.

Hintzman, D. L. (1986). " Schema abstraction" in a multiple-trace memory model. *Psychological Review*, *93*(4), 411.

a) Address correspondence to M.A. Redford at redford@uoregon.edu

Honda, K., Takano, S., & Takemoto, H. (2010). Effects of side cavities and tongue stabilization: Possible extensions of the quantal theory. *Journal of Phonetics*, *38*(1), 33–43.

Jakobson, R., Fant, C. G., & Halle, M. (1951). *Preliminaries to speech analysis: The distinctive features and their correlates*. MIT Press.

Johnson, K. (1997). *The auditory/perceptual basis for speech segmentation*.

Johnson, K. (2005). Speaker normalization in speech perception. In D. B. Pisoni & R. E. Remez (Eds.), *The Handbook of Speech Perception* (pp. 363–389). Malden, MA: Blackwell.

Kendall, T. (2013). *Speech rate, pause, and sociolinguistic variation*. Palgrave Macmillan, New York.

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America*. http://scitation.aip.org/content/asa/journal/jasa/59/5/10.1121/1.380986

Klatt, D. H. (1989). Review of selected models of speech perception. In *Lexical representation and process* (pp. 169–226). The MIT Press.

Kleinschmidt, D. F. (2019). Structure in talker variability: How much is there and how much can it help? *Language, Cognition and Neuroscience*, *34*(1), 43–68. https://doi.org/10.1080/23273798.2018.1500698

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*. https://doi.org/10.1037/a0038695.supp

Kleinschmidt, D. F., Weatherholtz, K., & Florian Jaeger, T. (2018). Sociolinguistic perception as inference under uncertainty. *Topics in Cognitive Science*, *10*(4), 818–834.

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

Kraljic, T., Brennan, S. E., & Samuel, A. G. (2008). Accommodating variation: Dialects, idiolects, and speech processing. *Cognition*, *107*(1), 54-81.

a) Address correspondence to M.A. Redford at redford@uoregon.edu

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, *51*, 141–178.

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, *56*(1), 1–15. https://doi.org/10.1016/j.jml.2006.07.010

Kreitewolf, J., Gaudrain, E., & von Kriegstein, K. (2014). A neural mechanism for recognizing speech spoken by different speakers. *NeuroImage*, *91*, 375–385. https://doi.org/10.1016/j.neuroimage.2014.01.005

Ladefoged, P. (1978). Expectation Affects Identification by Listening. *Language and Speech*, *21*(4), 373–374. https://doi.org/10.1177/002383097802100412

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

Levelt, W.J.M. (1989). *Speaking: From intention to articulation*. MIT Press.

Liberman, A. M., & Whalen, D. H. (2000). On the relation of speech to language. *Trends in Cognitive Sciences*, *4*(5), 187–196. https://doi.org/10.1016/S1364-6613(00)01471-6

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

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

Liberman, A. M., & Mattingly, I. G. (1985). *The motor theory of speech perception revised*. *21*(1), 1–36.

Lim, S. J., Fiez, J. A., & Holt, L. L. (2014). How may the basal ganglia contribute to auditory categorization and speech perception?. *Frontiers in Neuroscience*, 8, 230.

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

a) Address correspondence to M.A. Redford at redford@uoregon.edu

Lindblom, B. (1963). Spectrographic Study of Vowel Reduction. *The Journal of the Acoustical Society of America*, *35*(11), 1773–1781. https://doi.org/10.1121/1.1918816

Magnuson, J. S., Nusbaum, H. C., Akahane-Yamada, R., & Saltzman, D. (2021). Talker familiarity and the accommodation of talker variability. *Attention, Perception, & Psychophysics*, 1–19.

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

Mayo, C., Scobbie, J. M., Hewlett, N., & Waters, D. (2003). The influence of phonemic awareness development on acoustic cue weighting strategies in children's speech perception. Journal of Speech, Language, and Hearing Research, 46, 1184-1196.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1–86. https://doi.org/10.1016/0010-0285(86)90015-0

McGowan, K. B., & Babel, A. M. (2020). Perceiving isn't believing: Divergence in levels of sociolinguistic awareness. *Language in Society*, *49*(2), 231–256. https://doi.org/10.1017/S0047404519000782

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207–238. https://doi.org/10.1037/0033-295X.85.3.207

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic Feature Encoding in Human Superior Temporal Gyrus. *Science*, *343*(6174), 1006–1010. https://doi.org/10.1126/science.1245994

Miller, J. L., & Volaitis, L. E. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception and Psychophysics*, *46*(6), 505–512. https://doi.org/10.3758/BF03208147

Mrayati, M., Carré, R., & Guérin, B. (1988). Distinctive regions and modes: A new theory of speech production. *Speech Communication*, *7*(3), 257–286.

Mullenix, J. W., D. B. Pisoni 1990. Stimulus variability and processing dependencies in speech perception. *Perception and Psychophysics*, 47, 379–390.

a) Address correspondence to M.A. Redford at redford@uoregon.edu

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America*, *85*(1), 365–378. https://doi.org/10.1121/1.397688

Munson, B., Edwards, J., & Beckman, M. E. (2011). *Phonological Representationsin Language Acquisition: Climbing The Ladder of Abstraction*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199575039.013.0012

Nearey, T. M. (1990). The segment as a unit of speech perception. *Journal of Phonetics*, *18*(3), 347–373.

Nearey, T. M. (1997). Speech perception as pattern recognition. *The Journal of the Acoustical Society of America*, *101*(6), 3241–3254.

Niedzielski, N. (1999). The effect of social information on the perception of sociolinguistic variables. *Journal of Language and Social Psychology*, *18*(1), 62–85.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences*, *23*, 299–325.

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

Nusbaum, H. C., & Magnuson, J. S. (1997). Talker normalization: Phonetic constancy as a cognitive process. Talker variability in speech processing, 109-132.

Nycz, J. (2011). *Second Dialect Acquisition: Implications for Theories of Phonological Representation - ProQuest* [New York University]. https://search.proquest.com/openview/286dbcc561b203e335ffe6865eb03036/1?pq-origsite=gscholar&cbl=18750&diss=y

Nycz, J. (2015). Second Dialect Acquisition: A Sociophonetic Perspective. *Language and Linguistics Compass*, *9*(11), 469–482. https://doi.org/10.1111/lnc3.12163

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, *60*(3), 355–376. https://doi.org/10.3758/BF03206860

a) Address correspondence to M.A. Redford at redford@uoregon.edu

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech Perception as a Talker-Contingent Process. *Psychological Science*, *5*(1), 42–46. https://doi.org/10.1111/j.1467-9280.1994.tb00612.x

Ohala, J. J. (1996). Speech perception is hearing sounds, not tongues. *The Journal of the Acoustical Society of America*, *99*(3), 1718–1725.

Peelle, J. E. (2019). The neural basis for auditory and audiovisual speech perception. In *The Routledge Handbook of Phonetics* (pp. 193–216). Taylor and Francis.

Perrachione, T. K., Pierrehumbert, J. B., & Wong, P. C. M. (2009). Differential neural contributions to native- and foreign-language talker identification. *Journal of Experimental Psychology. Human Perception and Performance*, *35*(6), 1950–1960. https://doi.org/10.1037/a0015869

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, *24*(2), 175–184.

Pierrehumbert, J. (2002). Word-specific phonetics. *Laboratory Phonology*, 7(1), 101-140.

Pierrehumbert, J. B. (2016). Phonological representation: Beyond abstract versus episodic. *Annual Review of Linguistics*, *2*, 33–52.

Pike, K. L. (1947). *Phonemics: A technique for reducing languages to writing.* University of Michigan Press.

Pisoni, D. B. (1977). Identification and discrimination of the relative onset time of two component tones: Implications for voicing perception in stops. *The Journal of the Acoustical Society of America*, *61*(5), 1352–1361.

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: Cerebral lateralization as 'asymmetric sampling in time.' *Speech Communication*, *41*(1), 245–255. https://doi.org/10.1016/S0167-6393(02)00107-3

Poeppel, D., & Monahan, P. J. (2008). Speech perception: Cognitive foundations and cortical implementation. *Current Directions in Psychological Science*, *17*(2), 80–85.

a) Address correspondence to M.A. Redford at redford@uoregon.edu

Pufahl, A., & Samuel, A. G. (2014). How lexical is the lexicon? Evidence for integrated auditory memory representations. *Cognitive Psychology*, *70*, 1–30. https://doi.org/10.1016/j.cogpsych.2014.01.001

Remez, R. E. (2003). Establishing and maintaining perceptual coherence: Unimodal and multimodal evidence. *Journal of Phonetics*, *31*(3), 293–304. https://doi.org/10.1016/S0095-4470(03)00042-1

Remez, R. E., Ferro, D. F., Dubowski, K. R., Meer, J., Broder, R. S., & Davids, M. L. (2010). Is desynchrony tolerance adaptable in the perceptual organization of speech? *Attention, Perception, & Psychophysics*, *72*(8), 2054–2058. https://doi.org/10.3758/BF03196682

Remez, R. E., Rubin, P. E., Berns, S. M., Pardo, J. S., & Lang, J. M. (1994). On the perceptual organization of speech. *Psychological Review*, 101(1), 129.

Roelofs, A. (1999). Phonological segments and features as planning units in speech production. *Language and Cognitive Processes*, *14(2)*, 173-200.

Schwartz, J.-L., Basirat, A., Ménard, L., & Sato, M. (2012). The Perception-for-Action-Control Theory (PACT): A perceptuo-motor theory of speech perception. *Journal of Neurolinguistics*, *25*(5), 336–354.

Seneff, S. (1988). A joint synchrony/mean-rate model of auditory speech processing. *Journal of Phonetics*, *16(1)*, 55-76.

Shamma, S. (2003). Physiological foundations of temporal integration in the perception of speech. *Journal of Phonetics*, *31*(3), 495–501. https://doi.org/10.1016/j.wocn.2003.09.001

Stevens, K. N. (1981). Constraints imposed by the auditory system on the properties used to classify speech sounds: Data from phonology, acoustics, and psychoacoustics. *Advances in Psychology*, *7*, 61–74.

Stevens, K. N. (1989). On the quantal nature of speech. *Journal of Phonetics*, *17*(1–2), 3–45.

Stevens, K. N., & Blumstein, S. E. (1975). Quantal aspects of consonant production and perception: A study of retroflex stop consonants. *Journal of Phonetics*, *3*(4), 215–233.

a) Address correspondence to M.A. Redford at redford@uoregon.edu

Stevens, K. N., & Blumstein, S. E. (1981). The search for invariant acoustic correlates of phonetic features. *Perspectives on the Study of Speech*, 1–38.

Stevens, K. N., Blumstein, S. E., Glicksman, L., Burton, M., & Kurowski, K. (1992). Acoustic and perceptual characteristics of voicing in fricatives and fricative clusters. *The Journal of the Acoustical Society of America*, *91*(5), 2979–3000. https://doi.org/10.1121/1.402933

Stevens, K. N., & House, A. S. (1972). Speech perception(Acoustic model and linguistic, syntactic, lexical and semantic factors in speech perception and production process). *Foundations of Modern Auditory Theory.*, *2*, 3–62.

Stevens, K. N., & Klatt, D. H. (1974). Role of formant transitions in the voiced-voiceless distinction for stops. *The Journal of the Acoustical Society of America*, *55*(3), 653–659.

Strand, E., & Johnson, K. (1996). Gradient and Visual Speaker Normalization in the Perception of Fricatives. In *Natural Language Processing and Speech Technology: Results of the 3rd KONVENS Conference* (p. 26).

Summerfield, Q. (1975). How a full account of segmental perception depends on prosody and vice versa. In *Structure and process in speech perception* (pp. 51–68). Springer.

Sumner, M. (2015). The social weight of spoken words. *Trends in Cognitive Sciences*, *19*(5), 238–239.

Sumner, M., Kim, S. K., King, E., & McGowan, K. B. (2014). The socially weighted encoding of spoken words: A dual-route approach to speech perception. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.01015

Sussman, H. M., McCaffrey, H. A., & Matthews, S. A. (1991). An investigation of locus equations as a source of relational invariance for stop place categorization. *The Journal of the Acoustical Society of America*, *90*(3), 1309–1325. https://doi.org/10.1121/1.401923

Sussman, H. M., & Shore, J. (1996). Locus equations as phonetic descriptors of consonantal place of articulation. *Perception & Psychophysics*, *58*(6), 936–946. https://doi.org/10.3758/BF03205495

a) Address correspondence to M.A. Redford at redford@uoregon.edu

Toscano, J. C., & McMurray, B. (2010). Cue Integration With Categories: Weighting Acoustic Cues in Speech Using Unsupervised Learning and Distributional Statistics. *Cognitive Science: A Multidisciplinary Journal*, *34*(3), 434–464. https://doi.org/10.1111/j.1551-6709.2009.01077.x

Toscano, J. C., & McMurray, B. (2012). Cue-integration and context effects in speech: Evidence against speaking-rate normalization. *Attention, Perception & Psychophysics*, *74*(6), 1284–1301. https://doi.org/10.3758/s13414-012-0306-z

Winters, S. J., Levi, S. V., & Pisoni, D. B. (2008). Identification and discrimination of bilingual talkers across languages. *Journal of the Acoustical Society of America*, *123*(6), 4524. https://doi.org/10.1121/1.2913046

Yi, H. G., Leonard, M. K., & Chang, E. F. (2019). The encoding of speech sounds in the superior temporal gyrus. *Neuron*, *102*(6), 1096–1110.

Yi, H. G., Maddox, W. T., Mumford, J. A., & Chandrasekaran, B. (2016). The role of corticostriatal systems in speech category learning. *Cerebral Cortex*, 26(4), 1409-1420.

a) Address correspondence to M.A. Redford at redford@uoregon.edu