Attention, Perception, & Psychophysics
https://doi.org/10.3758/s13414-019-01702-x

PERCEPTUAL/COGNITIVE CONSTRAINTS ON THE STRUCTURE OF SPEECH COMMUNICATION: IN HONOR OF RANDY DIEHL

Long-standing problems in speech perception dissolve within an information-theoretic perspective

Keith R. Kluender[1] · Christian E. Stilp[2] · Fernando Llanos Lucas[3]

Abstract
An information theoretic framework is proposed to have the potential to dissolve (rather than attempt to solve) multiple long-standing problems concerning speech perception. By this view, speech perception can be reframed as a series of processes through which sensitivity to information—that which changes and/or is unpredictable—becomes increasingly sophisticated and shaped by experience. Problems concerning appropriate objects of perception (gestures vs. sounds), rate normalization, variance consequent to articulation, and talker normalization are reframed, or even dissolved, within this information-theoretic framework. Application of discriminative models founded on information theory provides a productive approach to answer questions concerning perception of speech, and perception most broadly.

Keywords Speech perception · Psychoacoustics · Perceptual learning

Randy L. Diehl is a true scholar. Most contributions to this volume address Randy Diehl the scientist and theorist. In this role, his highest aspiration is parsimony, and few things are more enjoyable than serious scientific argument. To be a student of Diehl included mentorship by scientist and theorist, and also historian and philosopher. Randy is as likely to make an analogy to Ptolemy or a pre-Socratic Greek as to precisely cite findings from a disparate field. It is only natural for him to have been such a successful scholar-leader as Dean of the College of Liberal Arts.

Among philosophers, the highest aspiration is to bring a perspective that does more than solve a problem. Ideally, one's philosophical project serves to resolve or dissolve a debate. British empiricist David Hume, the most important philosopher ever to write in English (Brown & Morris, 1988), is most frequently depicted in this way.

Within an empiricist tradition, we aspire to begin to provide parsimonious explanations for problems concerning speech perception by adopting a perspective that ideally serves to dissolve some long-standing problems.

Here, we invite the reader to consider an alternative perspective that, while well established in other domains, enjoys less application to speech perception. We advocate that by applying an information-theoretic framework to perception of speech, apparent problems such as rate normalization, talker normalization, and variability (aka lack of invariance) are satisfactorily dissolved. Further, insights into development as a native listener and talker are provided.

Objects versus objectives in speech perception

If one's problem is finding the right fencing to corral a unicorn, then there is really no problem at all. Instead, the problem is dissolved upon discovery that unicorns do not exist.

We begin with the historically contentious division concerning the proper objects of speech perception. Early consensus was that objects of speech perception are articulatory gestures. The most nativist of these perspectives, motor

✉ Keith R. Kluender
kkluender@purdue.edu

Christian E. Stilp
christian.stilp@louisville.edu

Fernando Llanos Lucas
f.llanos@pitt.edu

[1] Department of Speech, Language, and Hearing Sciences, Purdue University, 715 Clinic Drive, West Lafayette, IN 47907, USA

[2] Department of Psychological and Brain Sciences, University of Louisville, 308 Life Sciences Building, Louisville, KY 40292, USA

[3] Department of Communication Sciences and Disorders, University of Pittsburgh, Pittsburgh, PA 15213, USA

Published online: 01 April 2019

theory, maintained that perception of speech is special because perceivers had access to operations required to produce speech (Liberman, Cooper, & Studdert-Kennedy, 1967; Liberman & Mattingly 1985). Among other difficulties, this approach was challenged by demonstrations that nonhuman animals, that cannot produce speech, could learn to respond to speech distinctions much as humans do (e.g., Kluender, Diehl, & Killeen, 1987; Kluender, Lotto, Holt, & Bloedel, 1998; Kuhl & Miller, 1975).

In a related but thoroughly generalist approach, direct realism maintained that proper objects for perception are distal objects and events—articulatory acts for the case of speech (Fowler, 1986). By this view, perception of speech was not special at all. Direct realism is attractive because it is internally consistent theoretically and applies across senses and domains. Despite holding a different perspective (below), Diehl (1986) admired these strengths and encouraged his students to aspire to them in their own work. Facts of physical acoustics, however, presented the greatest problem for direct realism, because energy impinging upon transducers is insufficient to recover even an approximation to a distal acoustic source (Gordon, Webb, & Wolpert, 1992; Kluender & Alexander, 2008), a problem that is not unique to audition (Berkeley, 1709/1975).

Contra gestural accounts was evidence that some perceptual effects, initially hypothesized to be consequences of listeners' access to articulatory maneuvers, were adequately explained by general characteristics of auditory processing (e.g., Diehl, Kluender, & Walsh, 1990; Kluender, Diehl, & Wright, 1988; Parker, Kluender, & Diehl, 1986). These and other findings encouraged a perspective that acoustic patterns or patterns of sensory stimulation serve as proper objects of speech perception (e.g., Diehl & Kluender, 1989). This "auditorist" approach has largely supplanted earlier gesturalist accounts.

Auditorist accounts have their own limitations. It is true that fundamental operating characteristics of auditory systems do an admirable job of capturing some basic phenomena of speech perception. At their best, auditory considerations predict why languages adopt the particular sets of speech sounds that they do (e.g., Kingston & Diehl, 1994; Liljencrantz & Lindblom, 1972; Lindblom, 1986). They do not, however, capture the richly detailed correspondence between facts of articulation and perceptual consequences. There is a great deal more to speech perception than meets the generic ear, because speech perception is notoriously dependent on experience within a native language environment.

Those who shared any of the above perspectives appreciated the importance of learning from the start. For auditorists, effects of experience are subsumed within general principles of learning. There may be much to admire in this thoroughly generalist approach, given the parsimony of not requiring any processes specific to speech, and Kluender (1994) presented a somewhat comprehensive strategy much like this. However, this double-pronged attack of ears and experience may be criticized to the extent that unbridled deployment of learning serves as a general purpose "mop-up" operation after all that could be wrung out of the auditory system had been wrung out. In this regard, the auditorist approach may be unfalsifiable absent well-defined constraints on what counts as learning.

Here, we ask the reader to consider the possibility that there are no objects of perception, neither for speech nor for perception in general. Like unicorns, they do not exist at all. Instead, there are *objectives* for perception. Within this transparently functional perspective, perceptual objectives are to maintain adequate agreement between organisms and their worlds to facilitate adaptive behavior. Perceptual success does not require recovery or representations of the world per se. Perceivers' subjective impressions may be of objects and events in the world, and the study of perceptual processes benefits from inspection of real-world objects and events, patterns of light, sound pressure waves, transduction properties, and neural responses. By and large, however, viewing perception with a focus on either distal (environmental) or proximal (sensory) properties falls short of capturing the essential functional characteristic of perception—the relationship between an organism's world and its behavior.

This may strike the reader as a feel-good ecological depiction of organisms' ongoings within their respective milieus. Less pejorative, it harkens to the approach developed by Gibson (1950, 1966) and hews most closely to his characterization of affordances (Gibson, 1979). Here, we take advantage of a rigorous framework that, for 7 decades, has existed to characterize transmission of information between parties much in the same way that perception facilitates the relationship between organisms and their worlds.

Shannon information theory

Working for Bell Laboratories, Claude Shannon (1948) published *A Mathematical Theory of Communication*, which he developed for practical application to telephone bandwidth.[1] For present purposes, no math is required, as basic principles of information theory are straightforward and reasonably intuitive. A fundamental premise of Shannon's information theory is that information exists only in the relationship between transmitters and receivers. Information does not exist in either per se, and information does not portray any essential characteristics about either transmitters or receivers. In the same fashion, perceptual information exists in the relationship

[1] Shannon's work was developed in part based upon work by fellow Bell Labs researchers Harry Nyquist and Ralph Hartley. When Shannon's 1948 articles were published in a 1949 book, the title was changed modestly to *The Mathematical Theory of Communication* in recognition of the broad generalizability of the approach.

between organisms and their environments—not in either proximal or distal properties.

Shannon's definition of information is pivotal. Potential information is defined as a lack of predictability, formalized as *entropy*. With respect to a message, the probability of an item relates to what he called information content, with entropy being the average information content of the source. In this way, entropy is a metric of how uncertain things are with respect to a given transmitter and period of time. Let us suppose that the transmitter is the weather channel. If it is always sunny, then the entropy will be zero—no new information is transmitted after the first report. However, if there is constant mix of equally probable sunny, cloudy, rainy, and snowy days, entropy would be maximal because no a particular weather event is more probable than the others.

Of course, in actuality there are weather events (e.g., sunny) that predominate depending on the region or the season (e.g., summer in Arizona). In regions and seasons that are mostly sunny, knowing that it is going to snow has dramatically greater functional importance than knowing that it will remain sunny, because there are far greater consequences for subsequent actions. In a similar way, signal patterns that are less predictable from the same environment are more meaningful to the sensory system, as they encourage greater behavioral adjustments. Because Shannon's entropy provides a baseline of how predictable things are on average in a given environment, it can be used to estimate the relative impact of specific signal properties for sensory functioning.

Absent quantification, this relationship between entropy and information corresponds to the simple fact that there is no information in something one already knows or can predict. The less one knows or can predict (greater entropy), the greater the potential information. Simple examples of uncertainty are flipping a coin, rolling a die, and picking from a deck of playing cards. A coin yields two possible outcomes, a die yields six, and a deck of cards (sans jokers) yields 52. If measuring information in bits, as Shannon did, a coin, a die, and a deck of cards can convey 1.0, 2.6, and 5.7 bits of information, respectively. However, if flipping a trick coin or rolling a loaded die, then zero bits of information can be transmitted because the outcomes are certain and predictable.

Here, no particular characters or values, as would be required in a formal message, are implied. It will be sufficient for the reader to think about information as a simple lack of predictability. Finally, information is *transmitted* when uncertainty is reduced and agreement is achieved between receivers and transmitters, or in the case of perception, between organisms and their environments. Within a sea of alternative perceptual endpoints, agreement between the organism and environment is functionally successful to the extent that the organism arrives at the alternative that gives rise to adaptive behavior.

Primacy of change

Given these facts about information, it is true and fortunate that sensorineural systems respond only to change relative to what is stationary or predictable (Kluender, Coady, & Kiefte, 2003). Perceptual systems do not record absolute levels, whether loudness, pitch, brightness, or color. At least since Ernst Weber in the mid-19th century,[2] it has been widely appreciated that perception of differences is primitive. Sacrifice of absolute encoding has enormous benefits along the way to optimizing information transmission. Biological transducers have impressive dynamic range given their evolution via borrowed parts (e.g., gill arches to middle ear bones); however, biological dynamic range is always dwarfed by the physical range of absolute levels available from the environment. The beauty of sensory systems is that, by responding to relative change, a limited dynamic range can shift to optimize the amount of change that can be detected in the environment at a given moment. There are increasingly sophisticated mechanisms supporting sensitivity to change with ascending levels of processing, and several will be discussed in this contribution.

Readers who use cellular phones have direct experience with a communication system that transmits only change. This is *delta coding*, referring to the mathematical symbols δ and Δ denoting differences, and it is used for its efficiency. When there is a long pause between talkers on the phone, the subtle background noise drops to dead silence. No bits are wasted when there is no change in sound to convey, making delta coding most efficient. Application of delta (and double delta) features dramatically improves accuracy of automatic speech recognition (ASR; Furui, 1986).

Relative change, of course, requires context from which to change. Context itself is relatively uninformative; it is what already exists or can be predicted. Context can be very brief—the present or immediate past from which change arises. Context can be extended, such as predictable characteristics of listening conditions. Context can be measured in milliseconds, minutes, months, or even a lifetime of experience with predictable properties of a structured world. In all cases, perceptual systems are more efficient to the extent that predictable elements of context are registered in ways that enhance sensitivity to that which is less predictable and more informative.

By adopting this way of viewing information for perception, traditional distinctions between sensation, perception, and learning seamlessly extend through a series of processes that operate over broader ranges of time and experience. From

[2] John Locke (1690) reaches a similar conclusion in *An Essay Concerning Human Understanding* with what came to be known as the "three bowl experiment" through which he reflects upon the relative nature of sensation and anticipates what we now know as sensory adaptation.

peripheral sensory transduction through cortical organization consequent to experience, a series of successively more sophisticated processes extract predictability to make unpredictable (informative) changes easier to detect. Classic, albeit slippery to define, dissociations between sensation, perception, and learning dissolve within a common framework of extracting predictability to maximize sensitivity to change across expanding time scales.

Before employing this modest set of first principles to explore particular examples in perception of speech, credit is due to others. Some readers may be familiar with Fletcher's pioneering applications of information theory to speech (Fletcher 1953/1995) or G. A. Miller and Nicely's (1955) use of information theory to analyze consonant confusion data. For the present discussion, our application here will be most akin to early approaches of vision scientists such as Fred Attneave (1954, 1959) and Horace Barlow (1961). In keeping with Attneave and Barlow, we depart from Shannon's conceptualization of entropy as calculated on the basis of a sequence of discrete characters, and adopt a conceptualization of entropy more akin to thermodynamics by which white noise constitutes maximum entropy.

Attneave (1959) was an evangelist for applications of information theory to psychology most broadly. As a vision scientist, Attneave (1954) emphasized the highly redundant nature of sensory input before applying information-theoretic principles to argue for more economical perceptual processes that, among other things, subsume gestalt principles for perceptual organization (e.g., Koffka, 1935). Attneave's emphasis on redundant properties of a structured world figure prominently in the latter part of this contribution.

Barlow is a pioneering sensory neuroscientist, and he adopted an especially functional approach to sensorineural processing emphasizing the role of sensory processing—not for enriching the subjective experience of the world, but instead for modifying behavior in ways that encourage survival. This is in keeping with the functional emphasis advanced here when advocating for functional objectives of perception and eschewing objects of perception per se. Sympathetic to Attneave's arguments, Barlow (1961) introduced the "redundancy-reducing hypothesis" to capture the principle that successive sensory relays should prioritize signals that cannot be predicted by past and current events (ecologically most significant). For this, he adopts the language and mathematical formulations of information theory as his framework.

Attneave's and Barlow's approaches continue to be productive in contemporary theories of "efficient coding" (e.g., Barlow 1997, 2001; Schwartz & Simoncelli 2001; Simoncelli 2003; Simoncelli & Olshausen, 2001; Stilp & Assgari (2019); Stilp & Kluender, 2012, 2016); although, most all address questions concerning visual perception.

Potential information and speech intelligibility

We begin by considering the importance of sensory change at the lowest levels of the auditory system. Stilp and Kluender (2010) evaluated the extent to which measures of sensory change, tailored by the cochlea, may serve to explain intelligibility of connected speech. Other investigators attempted to parcel relative perceptual contributions of consonants, vowels, and transitions between them by removing and replacing segments with noise (Cole, Yan, Mak, Fanty, & Bailey, 1996; Fogerty & Kewley-Port, 2009; Fogerty, Kewley-Port, & Humes, 2012; Kewley-Port, Burkle, & Lee, 2007). Stilp and Kluender (2010) tested whether a simple metric, not defined by linguistic conventions, could best predict speech intelligibility. They tested whether the amount of spectral change over time would serve to define potential information, because signals that change more across time are locally less predictable from previous spectra. They quantified these psychoacoustic spectral changes as cochlea-scaled entropy (CSE).

CSE was quantified as Euclidean distances between adjacent psychoacoustically scaled spectral slices. Euclidean distances between adjacent 16-ms slices were calculated, and distances were then summed across durations of either 80 ms (approximating the mean duration of consonants in the TIMIT database, Garofolo et al., 1990) or 112 ms (approximate mean vowel duration). Cumulative Euclidean distances within a boxcar function[3] were taken as measures of spectral entropy and served as a psychoacoustic metric of potential information.

As is shown in Fig. 1, amount of potential information (CSE) robustly predicts intelligibility (r^2 = .80, p < .01). Although significantly more vowels were replaced with each increase in CSE, proportion of vowels or consonants replaced are not significant predictors of intelligibility (r^2 = .55, *ns*). Despite the fact that there is greater CSE during transitions into or out of higher amplitude portions of the speech signal, CSE is superior to intensity when predicting speech intelligibility (Aubanel, Cooke, Davis, & Kim, 2018).

Information normalizes speaking rate

Speech is highly intelligible across at least a four-fold range in speaking rate, and acoustic consequences of increasing rate of

[3] Here, a boxcar function is a rectangular window, convolved across a sentence, which includes values for only five or seven 16-ms spectral slices corresponding roughly to consonant and vowel durations, respectively, for the TIMIT corpus.

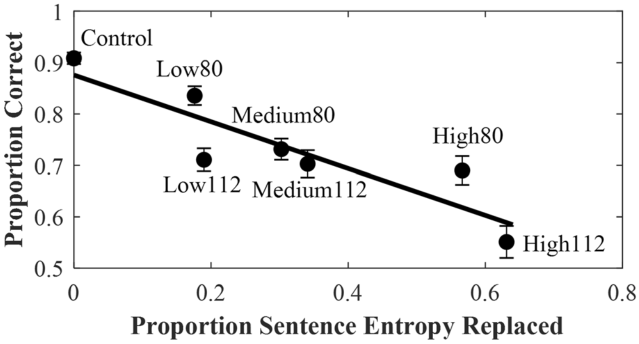

Fig. 1 Results of Stilp and Kluender (2010). Sentence intelligibility (proportion of words correctly identified) is plotted as a function of proportion of sentence entropy (CSE) replaced by noise. Data points are labeled by level of CSE replaced by noise (Low, Medium, High) and duration of each replacement (80-ms, 112-ms). Potential information is a significant predictor of sentence intelligibility ($r^2 = .80$, $p < .01$). Error bars represent standard error of the mean

speech are complex. In general, vowels compress more than consonants. Some consonants compress more than others do. Plosives reduce very little owing to their explosive release of air pressure, and continuants behave more like vowels. To make matters more challenging, perception of consonants in CVs depends upon vowel duration (e.g., J. L. Miller & Liberman, 1979), and perception of vowels depends on consonant duration (e.g., Ainsworth, 1975). Does a measure of relative change continue to predict intelligibility despite these uneven and interdependent consequences of changes in speaking rate?

When one adopts relative change as a metric for perceptual information, units corresponding to absolute frequency and intensity are discarded. Relative change (Weber's law) is depicted as ΔS/S. S in numerator is cancelled by S in denominator, leaving unitless delta (Δ). To the extent that change is a fundamentally unitless measure, absolute time also ought not matter within limits. To the extent that relative change is the most useful measure of potential information for perception, one should be able to warp time (slower or faster), and intelligibility should be predictable on the basis of amount of relative change.

Stilp and colleagues (Stilp, Kiefte, Alexander, & Kluender, 2010b) conducted a series of experiments in which measures of relative change were used to predict sentence intelligibility across a four-fold variation in rate of speech. Next, they imposed temporal distortion by time-reversing equal-duration segments (20 ms, 40 ms, 80 ms, and 160 ms) of every sentence (see, e.g., Saberi & Perrott, 1999). As is shown in Fig. 2 (left), listeners tolerated longer durations of temporal distortion at slower rates of speech. Listener performance across conditions is very well predicted based upon proportion of the utterance distorted (see Fig. 2, right) and not absolute duration (left).

Stilp and colleagues created a slightly modified measure of CSE to evaluate the extent to which relative change accounted for these data. Euclidean distances were measured between a given cochlea-scaled spectral slice and each slice thereafter until the end of the sentence. This provided a profile of the amount of cochlea-scaled spectral change across increasing intervals within an utterance. There are two notable features to these sentence analyses. First, just like listener performance, all three rates of speech converge to a single function when time is scaled as proportion of the utterance (see Fig. 3, right). Second, this extended measure of CSE is predictably related to syllable structure. This is because physical acoustic properties of speech have a local dependence (similarity) due to coarticulation. Owing to mass and inertia of articulators (as well as planning), articulatory movements are compromises between where articulators have been and where they are headed. Because the acoustic signal directly reflects these articulatory facts, the frequency spectrum assimilates in the same fashion as speech articulation. Cochlea-scaled spectra are more similar (less Euclidean distance) near in time and become more distinct (greater Euclidean distance) at longer intervals, and these time frames are proportional to both syllable duration and vowel-to-vowel intervals.

CSE functions peak at roughly two-thirds of mean syllable duration reflecting the fact that acoustic realizations of consonant and vowel sounds are largely conditioned by preceding vowels or consonants until they begin to assimilate to the next speech sound (see Fig. 3, right). For English VCVs, acoustics of the second vowel are largely independent of the first vowel, and identities of vowels in successive syllables are also largely independent. Consequently, beyond these relative maxima, distances regress toward the mean Euclidean distance of any spectral sample to the long-term spectrum of speech from the same talker. This simple, limited measure of information conveyed by spectral change accounts for a substantial proportion of variance in listener performance across all rate conditions ($r^2 = .89$, $p < .001$).

We now see that another attractive property of CSE is that it requires no explicit rate normalization. This measure of potential information naturally accommodates variable-rate speech. There have been substantial efforts to better understand how listeners normalize across speaking rate when identifying individual consonants (e.g., J. L. Miller, 1981; J. L. Miller & Liberman, 1979), vowels (e.g., Ainsworth 1972, 1974, 1975; Gottfried, Miller, & Payton, 1990), or words (e.g., J. L. Miller & Dexter, 1988), and all of these efforts have concentrated upon absolute physical changes in frequency and time. Within natural limits, CSE appears to capture scale-invariant transmission of information. This scale-invariance is not restricted to speech. Gervain and colleagues (Gervain, Werker, & Geffen, 2014) have shown that 5-month-old infants perceive commonality among varying "water sounds" defined by scale-invariant spectro-temporal structure. Using near-infrared spectroscopy, they (Gervain, Werker, Black, & Geffen,

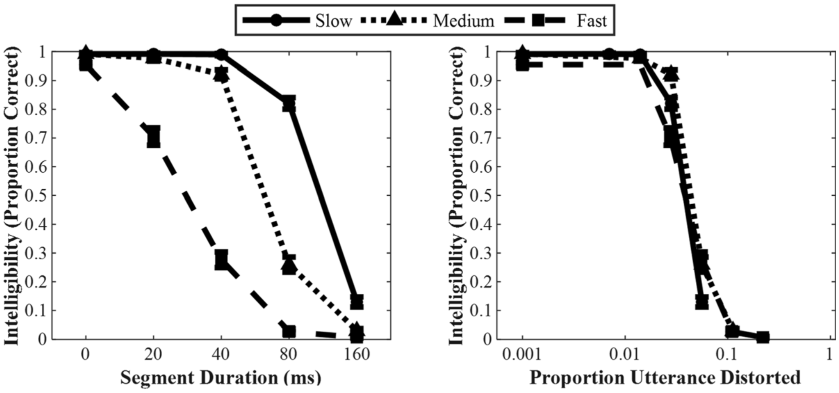

Fig. 2 Results from Stilp et al. (2010b). (Left) Intelligibility of sentences at a wide range of speaking rates when fixed-duration segments were temporally reversed. Performance declines faster (slower) for sentences at faster (slower) rates relative to medium-rate speech. (Right) Data converge to a common function when plotted using the relative measure of proportion of utterance distorted (segment duration divided by mean sentence duration)

2016) revealed cortical activity in 1-day-old to 3-day-old infants that is consistent with behavioral findings with older infants, and they situate their results within a framework of efficient coding. To the extent that potential information, not time or frequency per se, accounts for perception, concerns about normalization of time or frequency toward some iconic standard dissolve. While durations and frequencies may vary, potential information remains relatively constant and requires no such normalization. Figure 4.

The fact that CSE predicts intelligibility across a four-fold span of rate does not, in itself, capture successive steps from signal to a message composed of a string of words. As we discuss later, this information-theoretic approach does not make claims about extraction of phonetic segments per se. CSE provides insights into the earliest stages of a cascade of processes on the way to a linguistic message. These results illustrate that one will be most successful in understanding successive processes if they are posited to operate with respect to unitless relative change, not frequency, intensity, or loudness.

Spectral contrast and lack of invariance

Sensitivity to change is revealed most starkly in contrast effects. For example, a gray region appears darker against a white background and lighter next to a black background (see, e.g., Anderson & Winawer, 2005; see Fig, 5). While referred to as contrast "effects," within an information theoretic perspective, relative changes (contrast) are not so much effects as they are primitives. When environmental stimuli differ in some dimension (in Fig. 5, luminances of the neuron vs. its respective background), sensory systems do not register these differences in an absolute fashion. Instead, these differences are amplified beyond absolute differences. Sensory systems are attuned to what is informative (relative change), and enhancement of these differences heightens sensitivity.

Contrast effects are ubiquitous, and of course, they exist for audition (Cathcart & Dawson, 1928–1929; Christman, 1954). Spectral contrast—enhancement of differences between successive spectral compositions—has been shown to contribute to solving one of the most, if not the most, difficult questions

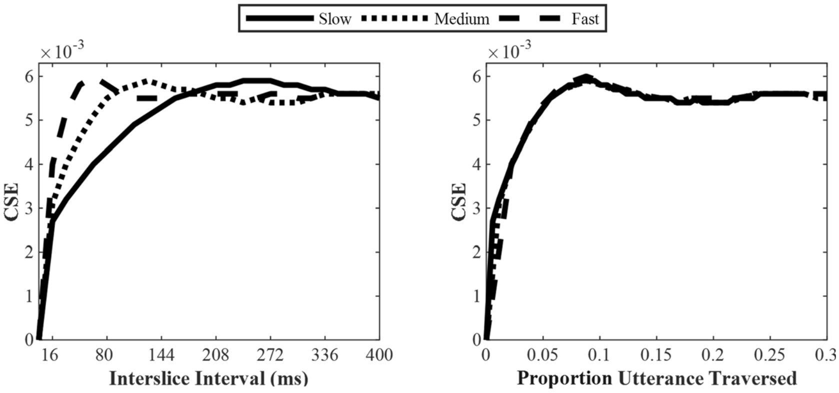

Fig. 3 CSE analyses of variable-rate sentences from Stilp et al. (2010b). (Left) CSE measured in 16-ms slices with increasing interslice intervals. Larger CSE values indicate greater spectral dissimilarity. Fast-rate speech peaks first (64 ms), followed by medium (128 ms), then slow (256 ms). Distances regress to the mean spectral distance between any two slices spoken by the same talker. (Right) Like behavioral data, CSE functions converge when plotted using the relative measure of proportion utterance traversed (interslice interval divided by mean sentence duration). All functions peak at approximately two-thirds of mean syllable duration

Fig. 4 Perception of speech across changes in speaking rates may be thought of much like viewing a photograph of the same image, in this case, Claude Shannon, across different sizes. The content of the speech and image are essentially unchanged as the patterns of changes remain intact

concerning speech perception: coarticulated speech. How do listeners hear a speech sound such as [d] when acoustic characteristics change dramatically depending upon sounds that precede and follow (e.g., vowels [e] versus [o]?; see Fig. 6).

Adjacent sounds assimilate toward the spectral characteristics of one another. Because the acoustic signal directly reflects these articulatory facts, the frequency spectrum assimilates in the same fashion that speech articulation assimilates.

Lindblom (1963) provided some of the best early evidence concerning how context systematically influences speech production. He observed that the frequency of the second formant (F_2) was higher in the productions of [dId] (“dɪd”) and [dʌd] (“dud”) than for the vowels [ɪ] and [ʌ] in isolation, and that F_2 was lower for vowels in [bɪb] and [bʌb]. In both contexts, F_2 frequency approached that of flanking consonants, which are higher for [d] than for [b]. In a subsequent study, Lindblom and Studdert-Kennedy (1967) demonstrated that perception of coarticulated vowels is complementary to these facts of articulation. Listeners reported hearing /ɪ/ (higher F_2) more often in [wVw] (lower consonant F_2)

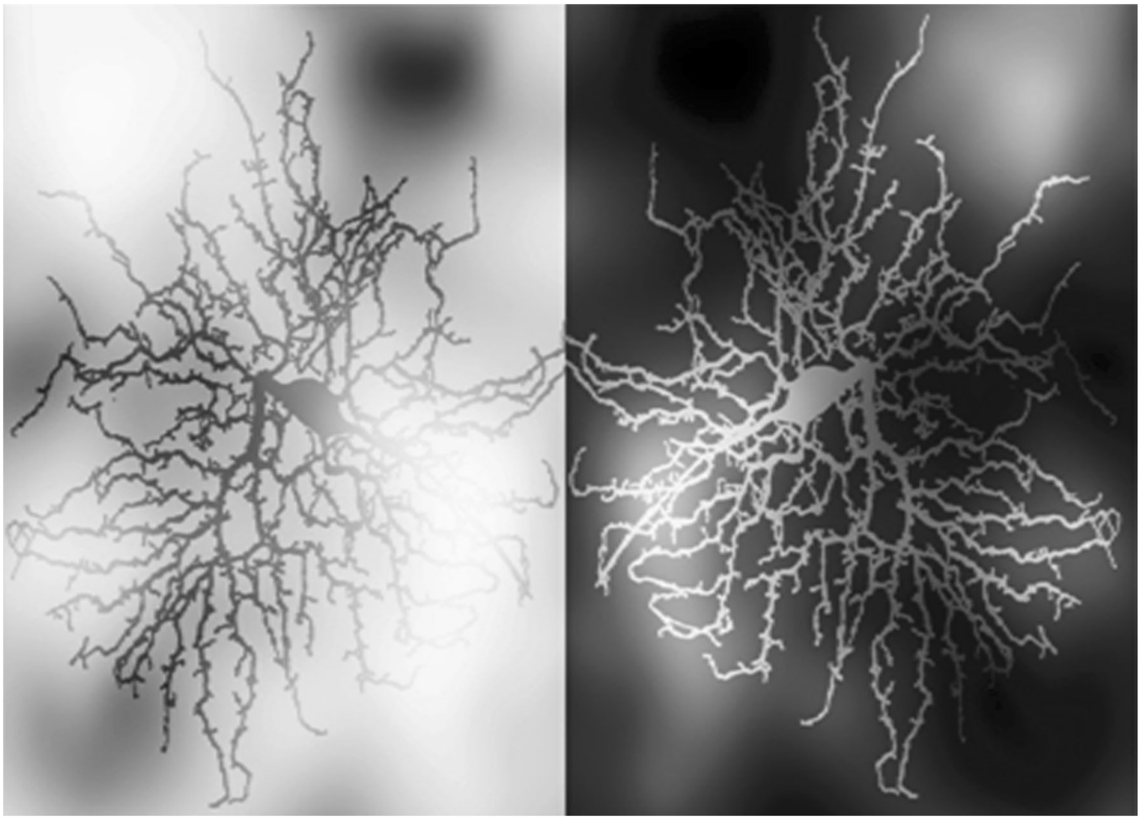

Fig. 5 Absolute luminance of neurons on left and right are equivalent. Processes of brightness contrast inherent to the visual system provide the sensation that the neuron on the left is much darker than the one on the right

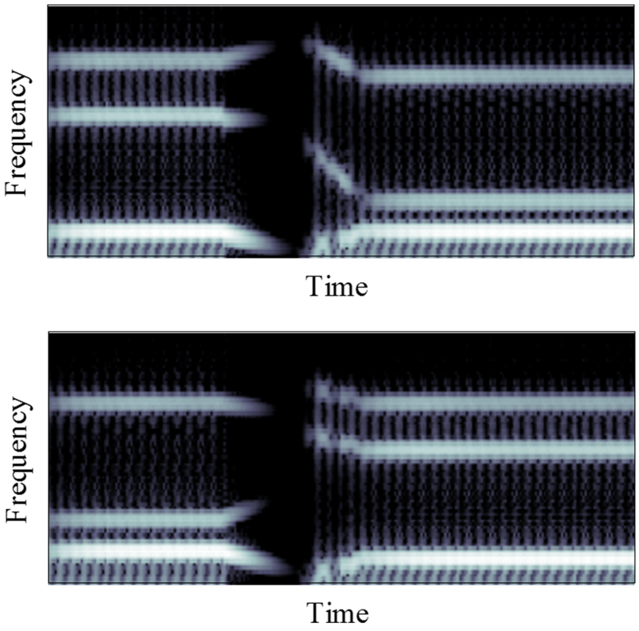

Fig. 6 Schematic spectrograms of [edo] (top) and [ode] (bottom.) Note that acoustic properties of [d] depend upon characteristics of preceding and following vowel sounds

context, and /ʌ/ more often in [jVj] (higher consonant F_2) context. Consonant context affected vowel perception in a manner complementary to the assimilative effects of coarticulation. Lindblom and Studdert-Kennedy (1967) wrote,

> It is worth reiterating . . . that mechanisms of perceptual analysis whose operations contribute to enhancing contrast in the above-mentioned sense are precisely the type of mechanisms that seem well suited to their purpose given the fact that the slurred and sluggish manner in which human speech sound stimuli are often generated tends to reduce rather than sharpen contrast. (p. 842)

One of the most thoroughly investigated cases for perceptual context dependence concerns the realization of [d] and [g] as a function of preceding liquid (Mann, 1980) or fricative (Mann & Repp, 1980). Perception of /d/ as contrasted with perception of /g/ can be largely signaled by the onset frequency and trajectory of the third formant (F_3). In the context of a following [a], a higher F_3 onset encourages perception of /da/ while a lower F_3 onset results in perception of /ga/. Onset frequency of the F_3 transition varies as a function of the preceding consonant in connected speech. For example, F_3-onset frequency for [da] is higher following [al] in [alda] than when following [ar] in [arda]. The offset frequency of F_3 is higher for [al] owing to a more forward place of articulation, and is lower for [ar] due to a more posterior place.

Perception of /da/ and /ga/ is complementary to the facts of production much as it is for CVCs. Listeners are more likely to report hearing /da/ (high F_3) when preceded by the syllable [ar] (low F_3), and hearing /ga/ (low F_3) when preceded by [al] (high F_3) (Lotto & Kluender, 1998; Mann, 1980). In subsequent studies, the effect has been found for speakers of Japanese who cannot distinguish [l] and [r] (Mann, 1986), for prelinguistic infants (Fowler, Best, & McRoberts, 1990), and for avian subjects (Lotto, Kluender, & Holt, 1997).

Coarticulation per se can be dissociated from its acoustic consequences by concatenating synthetic speech targets to nonspeech flanking energy that captures minimal essential spectral aspects of speech. Lotto and Kluender (1998) replaced [al] and [ar] precursors with nothing more than constant-frequency sinusoids set to the offset frequencies of F_3 for [al] and [ar] syllables. Perception of following [da-ga] shifted just as it did following full-spectrum [al] and [ar].

Holt, Lotto, and Kluender (2000) replicated the Lindblom and Studdert-Kennedy findings with CVCs using the vowels [ʌ] and [ɛ] flanked by stop consonants [b] and [d]. They replaced flanking [b] and [d] with FM glides that tracked the center frequency of only F_2 for [b] or [d]. Again, the pattern of results for flanking nonspeech FM glides mimicked that for full-spectrum [b] and [d] syllable-initial and syllable-final transitions. Finally, Holt (1999) demonstrated that perception of consonants following FM glides modeling F_2 transitions of natural vowels shift in patterns consistent with those for consonants following natural vowels in VCVs.

Based upon effects of nonspeech energy in VCCV, CVC, and VCV contexts, one can conclude that much of perceptual accommodation for coarticulation can be explained as little more than contrast. All of the findings are consistent with spectral contrast, whereby the spectral composition of context serves to enhance the perceptual efficacy of spectral components for adjacent sounds.

In keeping with typical usage, the term contrast has been used in a largely descriptive way this far. Kluender and Alexander (2008) provide extended descriptions of psychoacoustic precedents such as enhancements effects (e.g., Cardozo, 1967; Green, McKey, & Licklider, 1959; Houtgast, 1972;Schouten, 1940; Viemeister, 1980; Viemeister & Bacon, 1982). At levels as early as auditory nerve (AN), responses to test tones are dependent upon conditioning tones (Smith, 1977; Smith & Zwislocki, 1975). In these experiments, increments or decrements of test tone intensity are introduced atop preceding "conditioning" tones, and sensitivity to changes in intensity are maintained relative to these preceding "pedestals." Dynamic ranges of individual AN fibers adjust up and down depending on tonic levels of stimulation. Delgutte and colleagues (Delgutte, 1980, 1986, 1996; Delgutte, Hammond, Kalluri, Litvak, & Cariani, 1996; Delgutte & Kiang, 1984) have established the case for a broad role of peripheral adaptation in perception of speech, and they (Delgutte et al., 1996) argue that neurophysiological evidence demonstrates that "adaptation enhances spectral contrast between successive speech segments" (p. 3.)

Exceptions to this spectral contrast account for coarticulated speech should be noted. Viswanathan and colleagues (Viswanathan, Magnuson, & Fowler, 2010) did not find contrast effects for Tamil liquids. Based on this study and others (Viswanathan, Fowler, & Magnuson, 2009; Viswanathan, Magnuson, & Fowler, 2013, 2014), they suggest greater emphasis upon gestural substrates to speech. In particular, they argue that results from studies using nonspeech sounds such as sinewaves, but not speech, may be better accounted for by sensory masking. They also show that patterns of frequency change (formant kinematics) play an important role in ways that are consistent with speech gestures.

With respect to sensory masking, psychoacoustic explanations of enhancement effects (e.g., Viemeister & Bacon, 1982) do explicitly incorporate masking as well as adaptation of suppression/inhibition as processes through which spectral contrast occurs. Within these explanations, it is expected that continuously changing signals such as speech should be more resistant to masking effects, because all fluctuating signals (including sinewaves) produce less masking than steady-state counterparts. This, of course, does not encourage one to posit that speech signals somehow avoid these relatively low-level (e.g., auditory nerve, cochlear nucleus) processes altogether. It is difficult to argue that the frequency spectrum of speech is the sole instance in which a sensorineural system does not operate in a relative (contrastive) way, and the mechanisms described above are not in force.

The emphasis upon change is, of course, wholly consistent with all of the forgoing arguments. Where the present approach parts ways with Viswanathan and colleagues concerns whether sensitivity to formant variation emblematic of speech requires explicit appeal to articulatory maneuvers that create these changes. Speech signals that listeners hear are unerringly coarticulated, and listeners should be sensitive to these highly reliable patterns of change. In later sections of this contribution, especially those concerning ways through which listeners exploit reliable second-order statistics, a nongesturalist approach to these reliable patterns of change will be presented.

Auditory color constancy

Contributions of spectral contrast to perception of coarticulated speech focus narrowly in both time and frequency. Conversely, it is advantageous to adapt to reliable characteristics of the listening environment that extend across time and frequency.

By analogy, intensity and spectral composition of reflected light entering the eye vary dramatically depending upon illumination, yet viewers perceive objects as having relatively constant brightness and color. Spectral distribution of light entering the eye depends on both the spectrum of illumination and spectral characteristics light encounters on its path to the eye (Nassau, 1983). In order to achieve color constancy, the visual system must extract reliable spectral properties across the entire image in order to determine inherent spectral properties of objects within the scene (Boynton, 1988; Foster et al., 2006). One can consider the spectrum of illumination as a filter imposed on the full context of viewing. Perceptual color constancy is maintained by relative differences between the spectral composition of the object being viewed versus reliable spectral characteristics of the viewing context. In this way, perception is normalized or calibrated with respect to the imposing filter common to both context and object.

Readers may recall the great controversy concerning the colors of "the dress" in early 2015. Viewers of a photograph of the striped dress reported vastly different coloration of the dress, either black and blue stripes or white and gold stripes. The best scientific explanation of such radical differences in subjective impression is that visual systems are supposed to cancel contributions of the illuminant. For viewers whose visual systems "assumed" yellow-tinted illumination, the dress appeared black and blue, while the appearance of white and gold is consistent with blue-tinted illumination.

Many studies have demonstrated how auditory perception calibrates to properties of the listening context in ways quite similar to visual color constancy. In a classic study on context effects in vowel perception, Ladefoged and Broadbent (1957) showed that identification of a target vowel, [bɪt] (lower first formant frequency, F_1) versus [bɛt] (higher F_1), was affected by manipulations of average F_1 in a preceding context sentence. Raising average F_1 frequency in the context sentence elicited more /bIt/ (lower F_1) responses. Ladefoged and Broadbent (1957) drew explicit analogies between their findings and color constancy in vision. They wrote,

> It is obvious that this experiment provides a demonstration of perceptual constancy in the auditory field; that is an auditory phenomenon somewhat parallel to the visual case in which the response evoked by a stimulus is influenced by the stimuli with which it is closely associated. An example is the correct identification of the color of an object in widely differing illuminations. Consequently it is hoped that further investigation of the auditory phenomenon will provide data which are of general psychological interest. (p. 102)

Watkins and Makin (1994) argued that it was not specific F_1 frequencies per se that shifted responses in the studies by Ladefoged and Broadbent (1957), but rather the long-term spectrum of the context sentence. Watkins (1991) demonstrated effects similar to those of Ladefoged and Broadbent by filtering a precursor sentence with the difference between spectral envelopes of two vowels. This resulted in a context colored by spectral peaks of one vowel and by spectral notches corresponding to the peaks of the other vowel. When the context sentence was processed by a difference filter with the spectral shape of /ɪ/ minus /ɛ/, there was an increase in the number of /ɛ/ responses to an /ɪt /–/ɛt /series. This perceptual shift was observed across widely differing speech contexts varying in talker gender, spatial position, and ear of presentation, and whether the context was forward or time-reversed, or even if it was speech-shaped signal-correlated noise. In each case, perception calibrated to persistent spectral peaks and notches of the context emphasizing /ɪ/ such that listeners were more likely to hear target vowels as /ɛ/.

Kiefte and Kluender (2008) designed stimuli to assess relative contributions of spectrally global (spectral tilt) versus local (spectral peak) characteristics of a listening context on identification of vowel sounds. They varied both spectral tilt and center frequency of F_2 to generate a matrix of vowel sounds that perceptually varied from /u/ to /i/. Listeners identified these vowels following filtered forward or time-reversed precursor sentences. When precursor sentences were filtered to share the same long-term spectral tilt as the target vowel, tilt information was neglected and listeners identified vowels principally on the basis of F_2. Conversely, when precursors were filtered with a single pole centered at the F_2 frequency of the target vowel, perception instead relied upon tilt. These results demonstrate calibration to reliable global and local spectral features across both intelligible and unintelligible speech-like contexts. Stilp and colleagues (Stilp, Anderson, Assgari, Ellis, & Zahorik, 2016) later demonstrated that calibration is broadly indifferent to the source of reliable spectral properties. After replicating calibration to reliable spectral peaks in context sentences, they found even greater effects when stimuli were highly reverberant or when a pure tone at the F_2 frequency of a target vowel was added to the nonreverberant context.

Alexander and Kluender (2010) found the same patterns of performance as Kiefte and Kluender (2008) using nonspeech precursor contexts consisting of a harmonic spectrum filtered by four frequency-modulated resonances (somewhat akin to formants). Precursors filtered to match F_2 or tilt of following vowels induced perceptual calibration (i.e., diminished perceptual weight) to F_2 and tilt, respectively. Perceptual calibration to F_2 and tilt followed different time courses. Calibration to F_2 (spectrally local) was greatest for shorter duration precursors; in contrast, calibration to tilt was greatest for precursors that provided greater opportunities to sample the spectrum (longer duration and/or higher resonance-modulation rates).

Stilp and colleagues (Stilp, Alexander, Kiefte, & Kluender, 2010a) demonstrated that calibration to listening context is not limited to perception of speech. They asked listeners to identify edited notes from a French horn and tenor saxophone following either resynthesized speech or a short passage of music. Preceding contexts were "colored" by spectral-envelope-difference filters created to emphasize differences between horn and saxophone spectra. Listeners were more likely to report hearing a saxophone when following a context filtered to emphasize spectral characteristics of the French horn and vice versa. Despite clear changes in apparent acoustic source, perception calibrates to relatively predictable spectral characteristics of filtered context, differentially affecting perception of subsequent target nonspeech sounds (see also Frazier, Assgari, & Stilp, 2019).

Calibrating to acoustic context in the service of enhancing sensitivity to change would have been efficacious since the very first auditory systems, even before the advent of neocortex. So, it is likely that brainstem processes play an important role. Projections from superior olive to outer hair cells, collectively called the medial olivocochlear efferent system (MOC), have been hypothesized to adjust basilar membrane tuning to improve resolution of signals against background noise. Kirk and Smith (2003), for example, hypothesized the MOC evolved as a mechanism for "unmasking" biologically significant acoustic stimuli. Psychoacousticians (e.g., Champlin & McFadden, 1989; Krull & Strickland, 2008; McFadden & Champlin, 1990; Roverud & Strickland, 2010, 2014; Strickland, 2001; von Klitzing & Kohlrausch, 1994) have developed clever psychoacoustic techniques to characterize these MOC effects in human listeners.

Effects of the acoustic context on early sensorineural plasticity are further supported by animal models showing that activity from neurons in the rostral brainstem (e.g., inferior colliculus and medial geniculate body) is selectively inhibited from the deepest layers of the auditory cortex to enhance the acoustic properties that are functionally more relevant to the organism (Keuroghlian & Knudsen, 2007; Malmierca, Anderson, & Antunes, 2015). This is typically referred to as stimulus-specific adaptation (SSA)—a reduction in neural response to a relatively predictable sound and enhanced response to a relatively unpredictable sound. The typical form of such an experiment employs an oddball paradigm. After a neuron's responses have been characterized as equally responsive to two sounds, the sounds are presented at complementary ratios (e.g., 8:2 and 2:8). Despite the fact that each sound elicits the same neural response (firing rate) when odds are even, responses to the rarer sound for a given series of presentations are substantially greater than those to the more common sound. Because neural firing increases in response to the less frequent sound, fatigue cannot be the cause.

Stimulus-specific adaptation has been demonstrated in the inferior colliculus (IC; Malmierca, Cristaudo, Pérez-González, & Covey, 2009), auditory thalamus (Antunes, Nelken, Covey, & Malmierca, 2010), and auditory cortex (Ulanovsky, Las, & Nelken, 2003). Through and through, the auditory system is remarkably sensitive to predictability and its counterpart, unpredictability (information).

Exploiting redundancy in the ascending auditory system

It may be a slippery slope from foregoing descriptions of adaptation to processes of learning. Often, when introducing the topic of learning, teachers begin with the simplest example of habituation. Effects of a stimulus that once resulted in a behavioral change diminish absent consequences. In examples above, when stimuli become predictable, the auditory system neglects them relative to stimuli that are less predictable, and hence, potentially more informative.

Instead of splitting hairs concerning sensation, perception, and learning, we move to a useful distinction. All of the instances described this far, predictability from one spectral slice to another and predictable properties of a listening environment can be quantified using first-order statistics—measures of central tendency. Much predictability in sounds is not about temporal adjacencies or listening contexts, but instead is inherent in the ways sounds are structured.

Most sounds in the environment are well structured in frequency and in time due to physical constraints on the sources that create them. Listeners are keenly sensitive to acoustic structure (predictability, nonrandomness) in spectro-temporal composition, even when sounds are novel (Stilp, Kiefte, & Kluender, 2018). The ability to recognize novel sounds amidst a background of competing sounds can be predicted by the amount of relative entropy. Sounds with lower entropy (more structure, redundancy) are easier to discover and recognize. Greater redundancy (less entropy) across either time or frequency improves performance, as does different degrees of entropy in targets versus background distractors.

For sounds created by real structures including musical instruments and vocal tracts, there is greater predictability than one may first expect. Changes along different acoustic dimensions cohere in accordance with physical laws governing sound-producing sources. Articulatory maneuvers that produce consonant and vowel sounds give rise to multiple acoustic attributes, but this complexity is predictable in accordance with constraints imposed by vocal tracts. This redundancy across attributes contributes to robust speech perception despite substantial signal degradation (Assmann & Summerfield, 2004; Kluender & Alexander, 2008; Kluender & Kiefte, 2006).

Redundancy in speech epitomizes the general fact that objects and events in the world have structure. Attneave (1954) emphasized how patterns of stimulation upon the visual system are redundant because sensory events are highly interdependent in both space and time. This is simply because "the world as we know it is lawful" (p. 183). Adopters of information theory as an explanatory construct for human perception quickly came to appreciate the significance of capturing predictability among stimulus attributes in the interest of increasing sensitivity to relatively unpredictable changes between signals. Attneave argued that "it appears likely that a major function of the perceptual machinery is to strip away some of the redundancy of stimulation, to describe or encode incoming information *in a form more economical* than that in which it impinges on the receptors" (p. 189, emphasis added). Within an emphasis upon neural encoding, Barlow (1959) hypothesized that "it is supposed that the sensory messages are submitted to a succession of recoding operations which result in *reduction of redundancy* and increase of relative entropy of the messages which get through" (p. 536, emphasis added). By detecting and exploiting redundancy in the environment (predictability), perceptual systems enhance sensitivity to new information (unpredictability, or change).

These principles lie at the heart of contemporary models of efficient coding, and there have been many supporting findings in visual perception. Some studies concern adaptation to images varying in simpler aspects such as color, orientation, or directional movement, and extend to complex images including faces (see Clifford et al., 2007, for review). In their highly influential paper concerning efficient coding, Barlow and Földiák (1989) argued that populations of cortical neurons should organize in ways that capture correlations across inputs so that perceptual dimensions are more nearly orthogonal (decorrelated) and better able to detect changes in the environment that are not predictable (more informative) based upon prior experience. Barlow and Földiák proposed that absorption of correlations makes it easier to detect newly appearing associations resulting from new causal factors in the environment, and can account for effects of experience during cortical development. To these, one might add a simple, perhaps obvious, observation concerning neurons most broadly. Most neurons have many synapses along their dendrites. Whether a neuron fires depends on the joint contributions of many inputs, excitatory and inhibitory, along those dendrites. This simple fact of neural architecture requires that responses depend critically upon coincident activity across synapses.

There is growing physiological evidence that responses of neurons at successive stages of processing become increasingly independent from one another, and such demonstrations have been clearest in the auditory system (Chechik et al., 2006). Most recently, Liu and colleagues (Liu, Montes-Lourido, Wang, & Sadagopan, 2019) employed a combination of information-theoretic analyses of animal-call acoustics and recordings of responses to calls in the primary auditory cortex. They demonstrated that acoustic features identified using information-theoretic analyses of calls, to determine the most informative (and least redundant) features, effectively predict observed neural responses.

Second-order statistics and speech

It is well-attested that all contrasts between speech sounds are multiply specified. No single attribute is, in itself, both necessary and sufficient to support perception of a consonant or vowel. Kluender and colleagues (Kluender & Kiefte, 2006; Kluender & Alexander, 2008; Kluender & Lotto, 1999) have argued that one way in which multiple attributes are important to perception is the extent to which they are correlated with one another, and hence, provide redundancies that are central to sensorineural encoding of speech sounds.

We begin with vowels sounds. Amongst the most ubiquitous graphical presentations in all of speech research are the distributions of F_1 and F_2 formant peak frequencies across talkers provided by Peterson and Barney (1952) and later Hillenbrand, Getty, Clark, and Wheeler (1995), shown in Fig. 7. This morass of overlapping distributions for the simplest of speech sounds is emblematic of an apparent challenge for speech perceivers.

This apparent mess, however, is the consequence of a small number of rather straightforward physical acoustic constraints. Spectra of vowel sounds include peaks (formants) corresponding to resonances in the vocal tract. Center frequencies of these peaks depend principally upon two physical properties of vocal tracts. First, formant frequencies depend upon the shape of the vocal tract. The center frequency of the F_1 depends primarily upon how low or high the tongue and jaw are positioned. Open vowels with low tongue body such as /æ/ and /ɑ/ have higher F_1 frequencies, and close vowels with high tongue body such as /i/ and /u/ have lower F_1s. When the tongue is placed relatively forward in the vocal tract, the frequency of F_2 for front vowels such as /æ/ and /i/ is higher, but for vowels in which the tongue is placed relatively farther back such as /u/ and /ɔ/, F_2 is lower in frequency. While the center frequency of F_3 also varies across vowel sounds in perceptually significant ways, all vowel sounds can be depicted roughly by relative frequencies of F_1 and F_2 with the exception of /ɝ/. In addition, some vowels are produced with rounded lips (e.g., /u/ as in “boot”) or with different fundamental frequencies, among other variations.

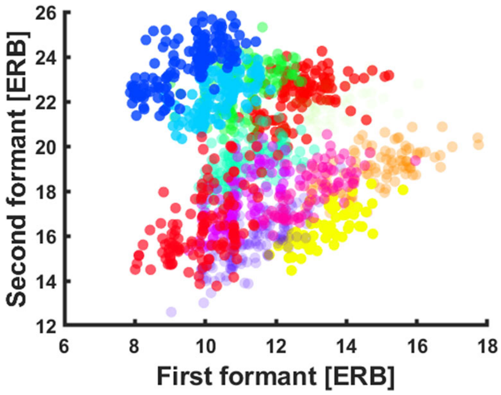

Fig. 7 The $_{\mathrm{ERB}}F_1$ and $_{\mathrm{ERB}}F_2$ of vowels in the Hillenbrand et al. (1995) database measured at vowel midpoint (50% of overall duration) for men, women, and children. Each color corresponds to a different vowel. (Color figure online)

The second major physical characteristic for vowel sounds is length of the vocal tract. When vocal tracts are shorter or longer, center frequencies of formants are higher or lower, respectively. It is given by the physical acoustics of tubes, vocal tracts included, that for a proportional increase or decrease in length, center frequencies of resonances decrease or increase by the same proportion (Nordström & Lindblom, 1975). One consequence of this dependency between vocal-tract length and vowel acoustics is that vowel sounds are very different across talkers. Vowels judged perceptually to be the same phonemically, such as /æ/ produced by men, women. and children, differ greatly in acoustic properties according to vocal-tract length. Vocal-tract length differences are not the only significant differences in renditions of the same vowel spoken by different talkers because other supralaryngeal properties also vary across talkers (Fant, 1966) properties, but they are the most substantial and easiest to characterize acoustically. Variation across talkers is so extreme that clear renditions of any given vowel overlap extensively with different vowels by talkers with vocal tracts of different lengths.

The approach taken here has a very long history, extending back to Lloyd (1890a, 1890b, 1891, 1892, cf. J. D. Miller, 1989) who claimed that vowels with common articulations result in common perceptions of vowel quality because they share common ratios among formants. Variants of this formant-ratio theory have appeared and reappeared with regularity (e.g., Broad, 1976; Chiba & Kajiyama, 1941; Kent, 1979; J. D. Miller, 1989; Minifie, 1973; Okamura, 1966).

Much of talker-dependent differences in vowel sounds, or at least those accounted for by vocal-tract length, decreases following two operations. Most true to physical acoustics, formant center frequencies may be converted to a logarithmic scale because proportional (fractional) changes in length correspond to proportional changes in center frequencies. Alternatively, center frequencies may be converted to a psychoacoustically realistic quasi-logarithmic scale such as Bark (Zwicker, 1961) or, more recently, equal rectangular bandwidth (ERB; Moore & Glasberg, 1983). Conversion to a psychoacoustically realistic scale has been productive in work concerning vowel sounds. For example, Syrdal and Gopal (1986) employed the Bark scale when demonstrating that differences between formants of a given vowel (more than or less than 3 Bark) could be used to classify vowels similarly to traditional high/low and front/back features.

Second, after adopting a logarithmic or quasi-logarithmic scale, relational measures must be used to capture systematicities across talkers. For example, J. D. Miller (1989) and Nearey (1989) employed measures of log(F_2/F_1) and log(F_3/F_2). Here, the ERB scale (Glasberg & Moore, 1990) is employed. Kluender and colleagues (Kluender, Stilp, & Kiefte, 2013) employed principal components analysis (PCA) to measure the amount of shared covariance for $_{\mathrm{ERB}}F_1$, $_{\mathrm{ERB}}F_2$, and $_{\mathrm{ERB}}F_3$ for each of the twelve vowels spoken by 139 men, women, and children and reported by Hillenbrand et al. (1995). Covariance between ERB F_1, F_2, and F_3 captures over three fourths of the substantial variability across men, women, and child talkers. (Figure 8).

Llanos and colleagues (Llanos, Jiang, & Kluender, 2014) demonstrated the functional significance of these covariance structures. They used three unsupervised clustering algorithms for learning minimal contrasts between English vowel pairs in the Hillenbrand et al. (1995) data set with formant frequencies converted to ERB. Two first-order models assumed uniform or Gaussian distributions of vowels in an F_1–F_2 space. The third model employed second-order statistics by encoding covariance between F_1 and F_2. The first-order Gaussian model performed better than a uniform distribution model for most vowel contrasts. By far the best performance (almost 90% accuracy) was achieved with the second-order model that outperformed both first-order models for every vowel pair. By exploiting correlations inherent across talkers, the challenge presented by Peterson and Barney (1952) is effectively dissolved.

When one embraces covariance, variability is no longer a problem. Instead, variability is a necessity. If one wishes to learn the relation between vocal-tract length and formant frequencies, one will not get far measuring central tendencies. Variability is essential to any model that depends upon covariance. Correlations between stimulus attributes are discovered only because there is variability across talkers, and as will be demonstrated next, across phonetic context.

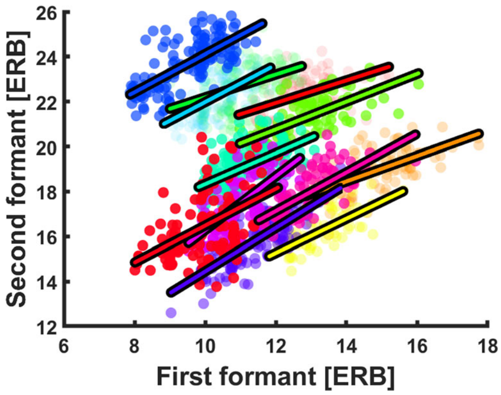

Fig. 8 Mean $_{\mathrm{ERB}}F_1$ and $_{\mathrm{ERB}}F_2$ of vowels in the Hillenbrand et al. (1995) database measured at vowel midpoint (50% of overall duration). Each color corresponds to a different vowel. Lines correspond to regressions between $_{\mathrm{ERB}}F_1$ and $_{\mathrm{ERB}}F_2$ for each vowel. (Color figure online)

Perceptual sensitivity to correlations among stimulus attributes may well account in part for listeners' solution to the lack of invariance across consonantal place of articulation. For example, acoustic information specifying /d/ is dramatically different depending upon the following vowel sound. Perceiving speech despite such variation was once thought to suggest that it was a uniquely human perceptual achievement (but see Kluender et al., 1987).

To address some of the variability for place of articulation across vowel contexts, Sussman and colleagues (e.g., Sussman, Fruchter, Hilbert, & Sirosh, 1998) reintroduced the idea of locus equationslocus equations (Delattre, Liberman, & Cooper, 1955) as part of an explanation for perception of place of articulation. They made exhaustive measurements of thousands of tokens of /b/, /d/, and /g/ produced before multiple vowels, by many different talkers, and found that the correlations between onset frequency of F_2 and F_2 frequency of the following vowel efficiently captured differences between /b/, /d/, and the two allophones of /g/ (front and back). Regression lines between F_2 onset and F_2 of the following vowel were distinct between /b/, /d/, and the two allophones of /g/. Correlation coefficients were relatively strong (r = .75–.96). Kiefte (2000) replicated these findings using ten Western Canadian English vowels measuring correlations of r = .98 for /b/, r = .87 for /d/, and r = .95 for /g/. Later efforts (Iskarous, Fowler, & Whalen, 2010; Lindblom & Sussman, 2012) revealed ways through which these acoustic systematicities are natural consequences of constraints on vocal tract maneuvers.

Vowels in VCs, like consonants in CVs, can also be characterized by reliable relationships between F_2 values as a function of time (Kluender, Stilp, & Kiefte, 2013). Directly analogous to locus equations for stop consonants, there is remarkable correlation between F_2 values for each vowel across variation in preceding plosive, averaging r = .82. Of course, owing to the fact that there are more vowels than there are consonantal places of articulation, differences between slopes of regression lines across vowels cannot be as profound as those found for /b/, /d/, and /g/, and the extent to which listeners exploit these correlations in vowel perception has yet to be investigated. Nearey (2010) demonstrated that, for all cases he tested (Hillenbrand et al., 2001), it is possible to decompose CVC syllables into locus constituents (CV and VC) and kinematic representations of vowels (vowel inherent spectral change).

There have been justifiable criticisms of the locus equation concept, perhaps most importantly the fact that other acoustic characteristics contribute to perception of place of articulation (e.g., Blumstein 1998). However, within a proposal that redundancy between correlated stimulus attributes should be efficiently coded, there is no formal upper bound on the number of attributes that can contribute to the overall covariance structure.

Learning correlations among stimulus attributes

As noted above, a legitimate criticism of appeals to general processes of audition and learning is that learning may serve mostly as a quasi-explanation for all things not captured by the auditory system per se. It is insufficient to appeal to the fact that behavior changes consequent to experience with a structured input. For example, with respect to multiple demonstrations of statistical learning (e.g., Aslin, Saffran, & Newport, 1998; Evans, Saffran, & Robe-Torres, 2009; Kirkham, Slemmer, & Johnson, 2002; Saffran, Aslin, & Newport, 1996), what types of learning, other than statistical, may be considered? In what way does the adjective "statistical" provide greater precision to an appeal to learning?

Within the scope of language learning, some investigators (e.g., Frost, Armstrong, Siegelman, & Christiansen, 2015; Siegelman, Bogaerts, Christiansen, & Frost, 2017; Siegelman, Bogaerts, & Frost, 2016) recently have been making important progress illuminating distinctions between and across multiple findings that have fallen within the broad scope of statistical learning. Here, at a lower level of perceptual learning as it relates to speech perception, we attempt to be explicit about exactly the forms of learning being hypothesized.

Within an information theoretic framework, the singular objective is to recognize and absorb predictability in order to increase sensitivity to change. Examples through which auditory processes register and demote first-order statistical properties that are predictable over relatively brief time scales were presented above. With an eye toward second-order statistics, how do listeners discover predictable covariation among stimulus attributes, and which models do and do not explain this learning?

For speech, this process of perceptual organization begins early in life and presumably supports, at least in part, infants' rapid mastery of multiply specified contrasts within their native language environment. To learn about acquisition of sensitivity to correlations among stimulus attributes by adult listeners, Stilp and colleagues (Stilp, Rogers, & Kluender, 2010c; Stilp & Kluender, 2011, 2012, 2016) employed novel complex stimuli that varied across two physically independent acoustic attributes: attack/decay (AD) and spectral shape (SS). A stimulus matrix was generated by crossing AD and SS series for which sounds separated by fixed distance in the stimulus space were approximately equally discriminable.

Stilp and colleagues (Stilp, Rogers, & Kluender, 2010c) presented subsets of stimuli to listeners. Most of the sounds listeners heard lay along a diagonal reflecting a correlation between AD and SS (for half of listeners, $r = 1.0$; for the other half, r = −1.0). Following only 7½ minutes of passive listening, listeners' ability to detect differences between pairs of sounds was characterized by their statistical characteristics and not their acoustic properties. Listeners retained the ability to discriminate sound pairs that obeyed the correlation between the two physical dimensions, but initially demonstrated great difficulty discriminating pairs of sounds that varied in only one dimension (AD or SS) or stimuli that violated the experienced correlation. Only through extended testing did discriminability of unidimensional or orthogonal differences recover. Perception warped to capture the redundancy between acoustic dimensions, only later discovering other variability that was present.

The authors tested three distinct computational models that could "learn" correlations between inputs. These three simple unsupervised-learning neural network models (see Fig. 9) shared similar architectures, but reflected different hypotheses about how sensorineural systems exploit covariance.

First was a Hebbian model (Hebb, 1949; Oja, 1982) in which connection weights adjusted in proportion to the correlation between input and output node activations. Second, an anti-Hebbian (decorrelation) model (Barlow & Földiák 1989) orthogonalized output dimensions by adjusting symmetric inhibition among output nodes proportional to their correlation. Finally, principal components analysis (PCA) was implemented in a third model (Sanger, 1989). For the PCA model, connections to output units adjusted in a Hebbian manner; however, the first output inhibited inputs to the second, effectively capturing the principal component from the input pattern and leaving the second unit to capture residual covariance. This model captured correlation across inputs (like the Hebbian model) and orthogonalized outputs (like the anti-Hebbian model).

The Hebbian model correctly predicted performance immediately following exposure, but owing to lack of inhibitory connections, the model was unable to adjust to capture performance across subsequent test trails. The anti-Hebbian model failed because it substantially expanded the second (orthogonal) dimension, thus predicting outstanding

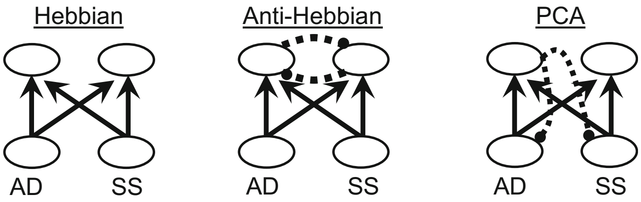

Fig. 9 Three simple network models that encode correlations between inputs. The bottom row is the input layer, and the top row is the output layer. Solid lines are excitatory, and dashed lines are inhibitory

discrimination for sound pairs that were effectively indiscriminable to listeners. The PCA model quickly discovered the first dimension capturing the correlations between AD and SS. With further experience during test trails, the PCA model adjusted apace with listener performance (see also Stilp & Kluender, 2012). Listener performance violated predictions of the Hebbian and anti-Hebbian models, but matched the PCA model quite well.

Data from this and numerous additional experiments by Stilp and colleagues all support the hypothesis that the auditory system rapidly and efficiently captures covariance (redundancy) across the set of complex stimuli. Like the PCA model, listener performance appears initially to capture the principal component of variation in the two-dimensional stimulus space at the expense of the orthogonal component before gradually encoding remaining variance. Both this initial component and the second component (first and second eigenvectors) appear to rapidly become weighted (eigenvalues) in a way that is proportional to the amount of variance accounted for by each dimension (Stilp & Kluender, 2012). Concordance between the model and listener performance is remarkable. Predictions from the PCA model based upon different relative weights on first and second components for different stimulus distributions have been validated with human listeners (Stilp & Kluender, 2016).

The particular PCA model investigated here (Sanger, 1989) is certainly oversimplified and is unlikely to exactly reflect neural learning mechanisms. Because stimuli were normed to equivalent perceptual distances, the perceptual space employed by Stilp and colleagues was linearized in a way that is amenable to a linear model such as PCA. The close correspondence between listener and model performance does suggest that sensorineural processes adapt to reflect experienced covariance so that dimensions of the perceptual space are weighted in a statistically sensible fashion.

While the neural locus or loci responsible for these human perceptual findings remains an open question, multiple neurophysiological studies with animal models are instructive. Most recently, findings by Stilp and colleagues have been replicated in neural response in primary auditory cortex of ferrets upon presentation of sounds that covaried in amplitude modulation and peak frequency of the spectral envelope (Lu, Liu, Dutta, Fritz, & Shamma, 2019). Following exposure to stimuli capturing these correlated attributes, signal-to-noise ratio (SNR) of spike-rate coding decreased orthogonal to the correlation, while remaining intact along the correlation; and, mutual information of spike coding increased only along the correlated dimension. Like results from Stilp and Kluender (2011), these patterns of activity remained intact in the face of variation in a third unrelated dimension.

Circuitry responsible for these responses remains unknown, but some neural observations are suggestive. Successes of the Sanger (1989) connectionist implementation of PCA to predict behavioral results depend on inhibitory circuits from the output layer to input layers. Microcircuitry across layers within cortical columns may convey inhibitory signals in a fashion like that proposed to support predictive coding (Bastos, Usrey, Adams, Mangun, Fries, & Friston, 2012). Less locally, required inhibitory circuitry may be provided within hierarchical auditory cortical regions, which extend from primary auditory cortex (AI) to belt areas to more lateral parabelt regions in a third stage of cortical processing (Kaas & Hackett, 2000). While primary auditory cortex (AI) is responsive to most sounds, responses later in the auditory hierarchy are selective for more complex stimuli, such as band-limited noise and frequency-modulated sweeps in belt areas (Rauschecker, Tian, & Hauser, 1995; Tian & Rauschecker, 2004; Wessinger et al., 2001) and species-specific vocalizations such as human speech in parabelt areas (Chevillet, Riesenhuber, & Rauschecker, 2011).

Redundancy enhances sensitivity to distinctions

From an information-theoretic perspective, systematicities (e.g., covariation in formant properties across talkers, locus equations across phonetic context) are embraced, but their utility is viewed quite differently relative to traditional perspectives. First, the perceptual utility of these systematicities is not to normalize toward some iconic ideal or template (which are as real as unicorns). Second, the real perceptual effect of efficiently coding these redundancies is to increase discriminability of speech sounds from one another. Differences between formant patterns that respect quasi-lawful consequences depending on who is talking are predictable and uninformative with respect to decoding the linguistic message. Consequently, differences that are linguistically significant are enhanced.

The relationship between formant center frequencies and fundamental frequencies (*f*0s) may be especially telling. All things being equal, different vowels are produced with different "intrinsic" *f*0s. Fundamental frequency is a consequence of vocal fold vibration, and in formal considerations of speech production, source properties owing to vocal fold activity (*f*0) can be viewed as largely independent of filter effects (resonances/formants; Fant, 1970). This fact may lead one to expect that the relationship between *f*0 and formant center frequencies should be much more tenuous than the obligatory relationships between $_{\mathrm{ERB}}\mathrm{F}_1$, $_{\mathrm{ERB}}\mathrm{F}_2$, and $_{\mathrm{ERB}}\mathrm{F}_3$—inevitable consequences of vocal-tract length. By contrast, *f*0 is much freer to vary; consider singing as an obvious example. Despite the relatively weak physical relationship between vocal-tract length and vocal-fold mass, talkers appear to enforce this systematicity despite being under no physiologic obligation to do so. Kluender et al. (2013) found that the proportions of variance captured by $_{\mathrm{ERB}}f0$ with mean ERB formant frequencies for each vowel in the Hillenbrand data set have a mean of 0.90.

It is possible that talkers "know" about the correlation between $_{\mathrm{ERB}}F_1$, $_{\mathrm{ERB}}F_2$, and $_{\mathrm{ERB}}F_3$ in vowel sounds? Talkers may reinforce the redundancy between $_{\mathrm{ERB}}F_1$, $_{\mathrm{ERB}}F_2$, and $_{\mathrm{ERB}}F_3$ by producing an $f0$ that respects this correlation. To the extent that the auditory system seizes upon these redundancies, distinctions between phonemically different vowel sounds are enhanced. At the same time, concerns about talker normalization dissolve. Different talkers all produce vowel sounds that share these relational systematicities or redundancies. Listeners discover these redundancies through experience with speech, and encoding of these redundancies serves to enhance discriminability of more informative differences between phonemically different vowel sounds.

The reader will note that, for each of the topics discussed thus far, the emphasis has been upon enhancing detection of that which is not predictable (information), and that sensorineural systems optimize sensitivity to change. Discovering redundancy enhances phonemically significant differences.

Thus, unlike notions of phonetic prototypes, consonants and vowels are revealed much more by what they are not than by how well they approximate some ideal. In this way, our conceptualization is consistent with the formal distinction between discriminative and generative models of classification in machine learning. A discriminative model learns only to tell differences between classes. Generative models learn about the particulars of each class by explicitly modeling the actual distribution of each class using Bayes' theorem. For example, a generative model of character recognition, such as for reading the address on an envelope, would attempt to capture defining characteristics (a la invariant features or prototypes) of each letter (e.g., 'd') across variations such as font and size. By contrast, the discriminative model discovers the ways in which 'd' is distinguished from, for example, 'a,' 'b,' 'c,' 'e,' 'f' across changes in font and size.

In the discipline of pattern classification, discriminative models are preferred over generative models for multiple reasons (Vapnik, 1998), not the least of which is that they typically prove more successful (fewer errors) as the size of training sets (experience) grows larger (Ng & Jordan, 2002). Further discussion of discriminative versus generative models is beyond the scope of this contribution (see, e.g., Vapnik, 1998); however, one can capture the main idea simply by thinking about speech perception with respect to confusion matrices (e.g., G. A. Miller & Nicely, 1955). Correct responses (diagonal) are correct entirely because distinctions from other stimuli (off diagonal) are detected.

This is a departure from notions of phonetic prototypes that may be revealed via a combination of cues, each with varying degrees of reliability and specificity. Instead, correlations embody predictable combinations of attributes in the service of enhancing sensitivity to differences between phonemically different sounds. Through experience, perceptual processes come to register predictable patterns of covariance, and by doing so, become especially sensitive to less predictable acoustic changes that distinguish different consonants and vowels. What matters are distinctions between speech sounds, not identities of consonants and vowels per se. Listeners hear the sounds of a language by virtue of learning how they are distinguished from all other consonants and vowels. This way of conceptualizing phonetic distinctions harkens back at least to Trubetzkoy (1939/1969). Linguists Roman Jakobson and Morris Halle (1971) stated it most clearly in their classic book *Fundamentals of Language*[4]: "All phonemes denote nothing but mere otherness" (p. 22).

Following the preceding approach to addressing problems of objects of perception, rate normalization, lack of invariance, and talker normalization, one may be left asking what happens to phonemes. Conceptualizing speech perception as a process by which phonemes are retrieved from acoustic signals is tradition. Within this tradition, problems of segmentation and lack of invariance arise from the fact that, if phonetic units exist, they are not like typed letters on a page. Some of the most recalcitrant problems in the study of speech perception are the consequence of adopting discrete phonetic segments as a level of perceptual analysis.

While some have questioned the reality of phonemes (e.g., Lotto, 2000; Lotto, & Holt, 2000; Port, 2006), the present approach is agnostic about the existence of phonemes and does not require them. Along the succession of redundancy-reducing operations, encoding may or may not bear resemblance to phonetic units as classically defined. It is not known whether listeners extract phonemes preliminary to recognizing words. Instead, phonemes are unnecessary to an information theoretic approach to speech perception. Adaptive behavior is not informed by defining one sound as A and one as B, but behavior is informed by differentiating functionally useful distinctions between A and B.

Learning to talk

Emphasis upon differences, not similarities to a putative prototype, has benefits when considering young children who are learning to talk. The claim here is that infants learn distinctions

[4] The emphasis upon distinctiveness intended here is not what may be implied within German Idealism popular in the times of Trubetzkoy and Jakobson. For present purposes, the contemporary vernacular interpretation is suggested, as the perspective adopted here is blatantly empiricist.

between sounds, not consonants and vowels as entities per se. Infants can distinguish speech sounds long before they can produce them, and the ways in which they detect differences between sounds become molded to the statistics of their native language sound environment during their first year of life (Werker, Gilbert, Humphrey, & Tees, 1981; Werker & Lalonde, 1988; Werker & Logan, 1985; Werker & Tees, 1983; Werker & Tees, 1984a, 1984b). Information transmission is optimized by maximizing sensitivity to differences—the benefit of consolidating redundant attributes. Emphasizing the ways that sounds are different, versus how they are the same, helps illuminate issues concerning learning how to produce speech sounds.

Owing to the developmental course of supralaryngeal anatomy and control, it is impossible for small developing vocal tracts to produce adult-like sounds of a language (e.g., Kent & Miolo, 1995; Kent & Vorperian, 1995; Vorperian et al., 1999, 2009). Until the vocal tract approaches adult length, mimicking speech sounds of the adult is not an option. Children's resonances are too high, and some vocal-tract configurations (e.g., high back /u/) are physiologically impossible for infants. Different vocal-tract architectures make it fruitless for young children to try to veridically match articulatory or auditory targets. However, it *is* possible for the developing vocal tract to produce sounds that are different in ways similar to how adult speech sounds differ. The child is able to create acoustic contrasts in speech that are proportional to those heard from adult talkers. In perceptual systems that have little or no access to absolute measures of anything, this quality is both attractive and essential.

Discovering word boundaries

Following the claim that a cascade of redundancy-reducing processes delivers keen sensitivity to distinctions among consonant and vowel sounds, might the same kind of processes be engaged when young children discover words? In connected speech, acoustic realization of the beginning and end of one word also overlaps with sounds of preceding and following words. Adult perception, however, is at odds with this acoustic reality. When listening to someone talk, most individual words stand out quite clearly as discrete entities. Listening to someone speak in a different language is often a very different experience. Entire sentences may sound like a single very long word. This is the situation faced by infants.

How children learn words is a complex question requiring sophisticated answers beyond the scope of the present contribution and the expertise of the authors. The reader is encouraged to begin with Cutler's (2012) comprehensive treatment. This being said, the present approach does provide a few insights, especially as they relate to prelexical processes such as finding boundaries between words.

Studies of CSE provide some insights. Higher CSE corresponds closely to the sonority hierarchy with peaks in CSE corresponding most closely to open configurations of the vocal tract, most commonly, vowels (Stilp & Kluender, 2010). Consequently, waxing and waning of CSE signals syllabic structure (Stilp et al., 2010b). Norris and colleagues' (Norris, McQueen, Cutler, & Butterfield, 1997) Possible Word Constraint (PWC) posits that a period of speech with no vowel is unlikely to be a word on its own, and Cutler (2012) explains how language learners might exploit PWC when discovering words. Intervals with no transitions into or out of vowels are intervals with little CSE. Because intervals with little entropy result in decreased changes in neural activity, they naturally become less salient to the infant listener.

Saffran and colleagues (Saffran et al., 1996) demonstrated that infants are sensitive to transitional probabilities between successive sounds within a speech stream. In their studies, they used streams of connected pseudowords, for which the probability of some sequences of consonant-vowels (CVs) was very high (1.0), while probability of other sequences was relatively low (0.33). Infants were sensitive to whether successive sounds share a history of co-occurrence. Here, the interpretation for these findings is that low transitional probabilities become perceptually salient because they correspond to spikes in information because one CV is not predictable from the last. Similar results have been reported for transitional probabilities between words instead of nonsense CV sequences (Pelucchi, Hay, & Saffran, 2009).

Statistics of English support this emphasis on word boundaries, as the ends of most words cannot be identified prior to the onset of the next (Luce, 1986). Infant sensitivity to boundaries is yet another example of using predictability to enhance sensitivity to change, and hence enhance transmission of information. Because this is a principle of perceptual systems most broadly, one expects this use of predictability to apply most generally. Indeed, these patterns of performance extend to infants experiencing tonal sequences (Saffran, Johnson, Aslin, & Newport, 1999), visual shapes (Kirkham et al., 2002), and visual feature combinations (Fiser & Aslin, 2002), and even nonhuman primates (Hauser, Newport, & Aslin, 2001) exhibit this sensitivity to transitional probabilities (for further review, see Saffran & Kirkham, 2018).

The observation that (low) transitional probabilities at junctures between words correspond to peaks in potential information is but a piece of a fuller explanation of how the language learners learn where words begin and end.

General considerations

In the foregoing, the case for an information-theoretic conceptualization of speech perception built progressively from operations that are highly focused in time and in frequency, appropriate to the outmost auditory periphery such as hair cells and the auditory nerve. Next were examples of calibrating to listening environments via first-order statistics across broader frequency and temporal extents, and the first stations along the auditory pathway with sufficient convergences of inputs were identified as MOC and IC.

Following these examples of the auditory system adapting to first-order statistics of the input, second-order statistics were introduced. The statistical technique PCA was presented as an analogy to ways through which correlations across inputs can be extracted. PCA has no a priori assumptions about the correlation structure to be discovered, a tabula rasa of sorts. A set of observations of possibly correlated variables is converted into a set of values of linearly uncorrelated variables called principal components. When multiple observations reveal covariance among variables, a limited number of orthogonal vectors (few relative to the number of variables) can account for a high percentage of the total variance across observations. Consequently, the goal is twofold: discover redundancies and increase efficiency by reducing the dimensionality of the input.

PCA is being used here only as analogy because it is unlikely that real neurons adhere to formal restrictions on how vectors are chosen, and the ways PCA fails as analogy are themselves illuminating. First, PCA is a linear analysis, and it is well known that sensory processes are nonlinear. Second, PCA assumes normally distributed values, and the real world complies with this assumption only to varying extents. A related analysis, independent component analysis (ICA; see, e.g., Hyvärinen & Oja, 2000) permits violations of normality, but this enhancement brings additional computational power and challenges for interpretation. Third, PCA and some versions of ICA require that vectors be ordered from most to least amount of variance accounted for, and neural systems most likely fall short of perfect efficiency. Here, lack of statistical perfection in a biological system is actually desirable. Perfectly efficient coding systems are brittle, and some inefficiency (redundancy) increases the robustness of communication systems against noise (Barlow, 2001). For purposes here, efficiency is gained by discovering correlated (redundant) attributes, but complete efficiency is nether claimed nor desired. What one desires is biologically optimal, not maximal, efficiency.

When perception is construed as a cascade of processes, each stage extracts redundancy across the outputs of earlier processes. To the extent that outputs of prior processes even approach independence, this could imply that seizing upon correlation again would become implausible. The solution to this seeming dead end is that, with every successive reduction of redundancy, information over which processing operates expands in space, frequency, time, and any other dimension of interest. Thus, statistical relationships that hold relatively locally do not constrain correlations at the next coarser grain of processing. As a consequence, the well-attested hierarchical organization of perceptual processing in sensorineural systems is a natural consequence of successive operations of efficient coding within an information theoretic framework.

Conclusions

In this contribution, we first established some first principles that motivate our perspective on speech perception and perception most broadly. There are two substantial consequences of adopting this information-theoretic framework to questions concerning speech perception. First, distractions concerning objects of perception are removed. Second, speech perception can be reframed as a series of processes through which sensitivity to information—that which changes and/or is unpredictable—becomes increasingly sophisticated and shaped by experience.

A simple measure of change in the auditory periphery (CSE) proved to be a remarkable predictor of speech intelligibility. Adopting measures of psychoacoustic change helps to dissolve some traditional concerns about perception across variation in speaking rate that putatively required some process of normalization. Multiple demonstrations were provided that show how reliable spectral characteristics of a listening context are factored out of perception in the service of emphasizing less reliable, more informative characteristics to better inform behavior.

Moving on from first-order statistics, second-order statistics were shown to be powerful descriptors of reliable covariance between acoustic attributes of speech sounds as they are structured by lawful properties of the vocal tract. Examples are vowels across talkers with different vocal-tract lengths and consonants across talkers and vowel contexts. Listeners very quickly learn correlations among stimulus attributes in complex nonspeech sounds, and this has remarkable consequences for discriminability of sounds depending upon whether they respect or violate experienced covariance. Finally, an information theoretic framework makes the life of a young language learner much easier.

We suggest that adopting an information-theoretic efficient coding framework provides a constructive way to address long-standing problems concerning appropriate objects of perception (gestures vs. sounds), rate normalization, variance consequent to articulation, and talker normalization.

Acknowledgements While errors or omissions are the responsibility of the authors, Steven Greenberg and an anonymous reviewer were very helpful to improving this contribution. The authors extend special thanks to John Kingston for his estimable efforts to help us to be more thoughtful, accurate, and accessible to readers. Finally, more than 30 years post-PhD, the first author remains exceedingly grateful for his mentorship from Randy Diehl, who taught him how to think more clearly and to take ideas seriously.

References

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

Ainsworth, W. A. (1974). The influence of precursive sequences on the perception of synthesized vowels. *Language and Speech, 17,* 103–109.

Ainsworth, W. A. (1975). Intrinsic and extrinsic factors in vowel judgments. In G. Fant & M. Tatham (Eds.), Auditory analysis and perception of speech (pp. 103–113). London: Academic Press.

Alexander, J. M., & Kluender, K. R. (2010). Temporal properties of perceptual calibration to local and broad spectral characteristics of a listening context. *Journal of the Acoustical Society of America, 128*(6), 3597–3613.

Anderson, B. L., & Winower, J. (2005). Image segmentation and lightness perception. *Nature, 434*(7029), 79–83.

Antunes, F. M., Nelken, I., Covey, E., & Malmierca, M. S. (2010). Stimulus-specific adaptation in the auditory thalamus of the anesthetized rat. *PLOS ONE.* https://doi.org/10.1371/journal.pone.0014071

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science, 9*(4), 321–324.

Assmann, P. F., & Summerfield, Q. (2004). The perception of speech under adverse conditions. In S. Greenberg, W. A. Ainsworth, A. N. Popper, & R. R. Fay (Eds.), Speech processing in the auditory system Vol. 14 (pp. 231–308). New York: Springer.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review*, *61,* 183–193.

Attneave, F. (1959). Applications of information theory to psychology: A summary of basic concepts, methods, and results. New York: Holt.

Aubanel, V., Cooke, M., Davis, C., & Kim, J. (2018). Temporal factors in cochlea-scaled entropy and intensity-based intelligibility predictions. *The Journal of the Acoustical Society of America, 143*(6), EL443–EL448.

Barlow, H. B. (1959). Sensory mechanisms, the reduction of redundancy, and intelligence. *NPL Symposium on the Mechanization of Thought Process, 10,* 535–539.

Barlow, H. B. (1961). Possible principles underlying the transformations of sensory messages. In W. A. Rosenblith (Ed.), Sensory communication (pp. 53–85). Cambridge: MIT Press.

Barlow, H. B. (1997). The knowledge used in vision and where it comes from. *Philosophical Transactions of the Royal Society of London B, Biological Science, 352*(1358), 1141–1147.

Barlow, H. B. (2001). Redundancy reduction revisited. *Network: Computation in Neural Systems*, 12, 241–253.

Barlow, H. B., & Földiák, P. (1989). Adaptation and decorrelation in the cortex. In R. Durbin, C. Miall, & G. Mitchison (Eds.), The computing neuron (pp. 54–72). New York: Addison-Wesley.

Bastos, A.M., Usrey, W.M., Adams, R.A., Mangun, G.R., Fries, P., & Friston, K.J. (2012). Canonical microcircuits for predictive coding. *Neuron. 76*(4),695–711. https://doi.org/10.1016/j.neuron.2012.10.038

Berkeley, G. (1709/1975). An essay towards a New Theory of Vision. In Philosophical Works Including the Works in Vision, e.d M.R. Ayers. London: J.M. Dent & Sons.

Blumstein, S. E. (1998). The mapping from acoustic structure to the phonetic categories of speech: The invariance problem. *Behavioral and Brain Sciences, 21,* 260.

Boynton, R.M. (1988). Color vision. *Annual Review of Psychology, 39,* 69–100.

Broad, D. J. (1976). Toward defining acoustic phonetic equivalence for vowels. *Phonetica, 33,* 401–424.

Brown, C. R., & Morris, W. E. (1988). Starting with Hume. New York: Continuum International.

Cardozo, B. L. (1967). Ohm's law and masking. *The Journal of the Acoustical Society of America, 42,* 1193.

Cathcart, E. P., & Dawson, S. (1928–1929). Persistence (2). *British Journal of Psychology*, *19,* 343–356.

Champlin, C. A., & McFadden, D. (1989). Reductions in overshoot following intense sound exposures. *The Journal of the Acoustical Society of America, 85,* 2005–2011. https://doi.org/10.1121/1.397853

Chechik, G., Anderson, M. J., Bar-Yosef, O., Young, E. D., Tishby, N., & Nelken, I. (2006). Reduction of information redundancy in the ascending auditory pathway. *Neuron, 51,* 359–368.

Chevillet, M., Riesenhuber, M., & Rauschecker, J.P. (2011). Functional Correlates of the Anterolateral Processing Hierarchy in Human Auditory Cortex. *Journal of Neuroscience. 31*(25), 9345–9352. https://doi.org/10.1523/JNEUROSCI.1448-11.2011

Chiba, T., & Kajiyama, M. (1941). The vowel: Its nature and structure. Tokyo: Tokyo Publishing Co.

Christman, R. J. (1954). Shifts in pitch as a function of prolonged stimulation with pure tones. *American Journal of Psychology, 67,* 484–491.

Clifford, C. W. G., Webster, M. A., Stanley, G. B., Stocker, A. A., Kohn, A., Sharpee, T. O., & Schwartz, O. (2007). Visual adaptation: Neural, psychological and computational aspects. *Vision Research, 47,* 3125–3131.

Cole, R., Yan, Y., Mak, B., Fanty, M., & Bailey, T. (1996). The contribution of consonants versus vowels to word recognition in fluent speech. Paper presented at the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'96), Atlanta.

Cutler, A. (2012). Native listening: Language experience and the recognition of spoken words. Cambridge, MA: MIT Press.

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

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

Delgutte B. (1986). Analysis of French stop consonants with a model of the peripheral auditory system. In J. S. Perkell & D. H. Klatt (Eds.), Invariance and variability of speech processes (pp. 131–177). Hillsdale: Erlbaum.

Delgutte B. (1996). Auditory neural processing of speech. In W. J. Hardcastle & J. Laver (Eds.), The handbook of phonetic sciences (pp. 507–538). Oxford: Blackwell.

Delgutte, B., Hammond, B. M., Kalluri, S., Litvak, L. M., & Cariani, P. A. (1996). Neural encoding of temporal envelope and temporal interactions in speech. In W. Ainsworth & S. Greenberg (Eds.), *Auditory basis of speech perception* (pp. 1–9). European Speech Communication Association.

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

Diehl, R. L. (1986). Coproduction and direct perception of phonetic segments: A critique. *Journal of Phonetics, 14,* 61–66.

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

Diehl, R. L., Kluender, K. R., & Walsh, M. A. (1990). Some auditory bases of speech perception and production. In W. A. Ainsworth (Ed.), Advances in speech, hearing, and language processing. London: JAI Press.

Evans, J., Saffran, J. R., & Robe-Torres, K. (2009). Statistical learning in children with specific language impairments. *Journal of Speech, Language, & Hearing Research*, *52,* 321–335.

Fant, G. (1966). A note on vocal tract size factors and nonuniform F-pattern scalings. *Speech Transmission Laboratory Quarterly Progress and Status Report, 7*(4), 22–30.

Fant, G. (1970). Acoustic theory of speech production with calculations based on X-ray studies of Russian articulations. The Hague: Mouton.

Fiser, J., Aslin, R.N. (2002). Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition. 28*(3),458–67.

Fletcher, H. (1995). Speech and hearing in communication. New York: Krieger. (Original work published 1953)

Fogerty, D., & Kewley-Port, D. (2009). Perceptual contributions of the consonant-vowel boundary to sentence intelligibility. *Journal of the Acoustical Society of America*, *126,* 847–857.

Fogerty, D., Kewley-Port, D. & Humes, L. E. (2012). The relative importance of consonant and vowel segments to the recognition of words and sentences: Effects of age and hearing loss. *Journal of the Acoustical Society of America*, *132,* 1667–1678.

Foster, D.H., Amano, K., & Nascimento, S.M.C. (2006). Color constancy in natural scenes explained by global image statistics. *Visual Neuroscience, 23,* 341–349.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics, 14*(1), 3–28.

Fowler, C. A., Best, C. T., & McRoberts, G. W. (1990). Young infants' perception of liquid coarticulatory influences on following stop consonants. *Perception & Psychophysics*, *48*(6), 59–570.

Frazier J.M., Assgari A.A., & Stilp C.E. (2019) Musical instrument categorization is highly sensitive to spectral properties of earlier sounds. *Attention, Perception, & Psychophysics* (in press)

Frost R., Armstrong, B. C., Siegelman, N., & Christiansen, M. H. (2015). Domain generality versus modality specificity: The paradox of statistical learning. *Trends in Cognitive Science, 19,* 117–125.

Furui, S. (1986). On the role of spectral transition for speech perception. *Journal of the Acoustical Society of America, 80,* 1016–1025.

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., & Dahlgren, N. (1990). DARPA TIMIT acoustic-phonetic continuous speech corpus CDROM (NTIS Order No. PB91-505065). Gaithersburg: National Institute of Standards and Technology.

Gervain, J., Werker, J. F., Black, A., & Geffen, M. N. (2016). The neural correlates of processing scale-invariant environmental sounds. *NeuroImage, 133,* 144–150.

Gervain, J., Werker, J. F., & Geffen, M. N. (2014). Category-specific processing of scale-invariant sounds in infancy. *PLOS ONE*, 9(5), e96278.

Gibson, J. J. (1950). The perception of the visual world. Boston: Houghton Mifflin.

Gibson, J. J. (1966). The senses considered as perceptual systems. Boston: Houghton Mifflin.

Gibson, J. J. (1979). The ecological approach to visual perception. Boston: Houghton Mifflin.

Glasberg, B. R., & Moore, B. C. J. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research, 47,* 103–138.

Gordon, C., Webb, D. L., & Wolpert, S. (1992). One cannot hear the shape of a drum. *Bulletin of the American Mathematical Society, 27,* 134–138

Gottfried, T. L., Miller, J. L., & Payton, P. E. (1990). Effect of speaking rate on the perception of vowels, *Phonetica, 47,* 155–172.

Green, D. M., McKay, M. J., & Licklider, J. C. R. (1959). Detection of a pulsed sinusoid in noise as a function of frequency. *Journal of the Acoustical Society of America, 31,* 1446–1452.

Hauser, M. D., Newport, E. L., & Aslin, R. N. (2001). Segmentation of the speech stream in a non-human primate: Statistical learning in cotton-top tamarins. *Cognition, 78*(3), 53–64.

Hebb, D. O. (1949). Organization of behavior. New York: Wiley.

Hillenbrand, J, Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America, 97,* 3099–3111.

Hillenbrand, J.M., Clark, M.J., and Nearey, T.M. (2001). Effects of consonant environment on vowel formant patt erns. *Journal of the Acoustical Society of America, 109,* 748–763.

Holt, L. L. (1999). *Auditory constraints on speech perception: An examination of spectral contrast* (Doctoral dissertation). University of Wisconsin–Madison.

Holt, L. L., Lotto, A. J., & Kluender, K. R. (2000). Neighboring spectral content influences vowel identification. *Journal of the Acoustical Society of America*, *108,* 710–722.

Houtgast, T. (1972). Psychophysical evidence for lateral inhibition in hearing. *Journal of the Acoustical Society of America*, *51,* 1885–1894.

Hyvärinen, A., & Oja, E. (2000). Independent component analysis: Algorithms and applications. *Neural Networks, 13*(4/5), 411–430.

Iskarous, K., Fowler, C. A., & Whalen, D. H. (2010). Locus equations are an acoustic expression of articulator synergy. *Journal of the Acoustical Society of America*, *128*(4), 2021–2032.

Jakobson, R., & Halle, M. (1971). The fundamentals of language. The Hague: Mouton.

Kaas, J. H., & Hackett, T. A. (2000). Subdivisions of auditory cortex and processing streams in primates. *Proceedings of the National Academy of Science*, *97*(22), 11793–11799.

Kent, R. D. (1979). Isovowel lines for the evaluation of vowel formant structure in speech disorders. *Journal of Speech and Hearing Disorders, 44,* 513–521.

Kent, R. & Miolo, G. (1995). Phonetic abilities in the first year of life. In P. Fletcher & B. MacWhinney (eds), *The Handbook of Child Language*. Blackwell: Oxford.

Kent, R.D. & Vorperian, H.K. (1995). Anatomic development of the craniofacial-oral-laryngeal systems: A review. *Journal of Medical Speech-Language Pathology, 3*(1),145–90.

Keuroghlian, A. S., & Knudsen, E. I. (2007). Adaptive auditory plasticity in developing and adult animals. *Progress in Neurobiology, 82*(3), 109–121.

Kewley-Port, D., Burkle, T. Z., & Lee, J. H. (2007). Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners. *Journal of the Acoustical Society of America, 122,* 2365–2375.

Kiefte, M. (2000). *The perception of spectrally and temporally distorted prevocalic stop consonants* (Unpublished doctoral dissertation). University of Alberta, Edmonton.

Kiefte, M., & Kluender, K. R. (2008). Absorption of reliable spectral characteristics in auditory perception, *Journal of the Acoustical Society of America*, *123,* 366–376.

Kingston, J., & Diehl, R. L. (1994). Phonetic knowledge. *Language*, *70*(3), 419–454.

Kirk, E. C., & Smith, D. W. (2003). Protection from acoustic trauma is not a primary function of the medial olivocochlear efferent system. *Journal of the Association for Research in Otolaryngology, 4,* 445–465.

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition, 83*(2), 35–42.

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

Kluender, K. R., & Alexander, J. M. (2008). Perception of speech sounds. In A. I. Basbaum, A. Kaneko, G. M. Shepard, & G. Westheimer (Eds.), The senses: A comprehensive reference Vol. 3, pp. 829–860). San Diego: Academic Press.

Kluender, K. R., Coady, J. A., & Kiefte, M. (2003). Sensitivity to change in perception of speech. *Speech Communication*, *41*(1), 59–69.

Kluender, K. R., Diehl, R. L., & Killeen, P. R. (1987). Japanese quail can learn phonetic categories. *Science, 237,* 1195–1197.

Kluender, K. R., Diehl, R. L., & Wright, B. A. (1988). Vowel-length differences before voiced and voiceless consonants: An auditory explanation. *Journal of Phonetics, 16*(2), 153–169.

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

Kluender, K. R., & Lotto, A. J. (1999). Virtues and perils of empiricist approaches to speech perception. *Journal of the Acoustical Society of America, 105,* 503–511.

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

Kluender, K. R., Stilp, C. E., & Kiefte, M. (2013). Perception of vowel sounds within a biologically realistic model of efficient coding. In G. S. Morrison & P. F. Assmann (Eds.), Vowel inherent spectral change, modern acoustics and signal processing (pp. 117–151.) Berlin: Springer-Verlag.

Koffka, K. (1935). Principles of gestalt psychology. New York: Hartcourt, Brace.

Krull, V., & Strickland, E. A. (2008). The effect of a precursor on growth of forward masking. *The Journal of the Acoustical Society of America, 123,* 4352–4357.

Kuhl, P. K., & Miller, J. D. (1975). Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants. *Science*, *190*(4209), 69–72.

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

Liberman, A. M., Cooper F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74,* 431–61.

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

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

Lindblom, B. (1963). Spectrographic study of vowel reduction. *The Journal of the Acoustical Society of America*, *35*(11), 1773–1781.

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

Lindblom, B., & Studdert-Kennedy, M. (1967). On the role of formant transitions in vowel recognition. *The Journal of the Acoustical Society of America*, *42,* 830–843.

Lindblom, B., & Sussman, H. M. (2012). Dissecting coarticulation: How locus equations happen. *Journal of Phonetics*, *40*(1), 1–19.

Liu, S. T., Montes-Louride, P., Wang, X., & Sadagopan, S. (2019). Optimal features for auditory categorization. *Nature Communications*. https://doi.org/10.1101/411611

Llanos, F., Jiang, Y., & Kluender, K. R. (2014). Exploiting 2nd-order statistics improves statistical learning of vowels. Poster presented at the 168th Meeting of the Acoustical Society of America, Indianapolis.

Lloyd, R. J. (1890a). Some researches into the nature of the vowel-sound. Liverpool: Turner and Dunnett.

Lloyd, R. J. (1890b). Speech sounds: Their nature and causation (I). *Phonetische Studien*, *3,* 251–278.

Lloyd, R. J. (1891). Speech sounds: Their nature and causation (II-IV). *Phonetische Studien*, *4*, 37–67, 183–214, 275–306.

Lloyd, R. J. (1892). Speech sounds: Their nature and causation (V-VII). *Phonetische Studien*, *5,* 1–32, 129–141, 263–271.

Locke, J. (1690). An essay concerning human understanding. London: Thomas Bassett.

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

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

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

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

Lu, K., Liu, W., Dutta, K., Fritz, J. B., & Shamma, S. A. (2019). Adaptive efficient coding of correlated acoustic properties. *bioRxiv.* https://doi.org/10.1101/548156

Luce, P.A. (1986). A computational analysis of uniqueness points in auditory word recognition. *Perception and Psychophysics. 39*, 155–158.

Malmierca, M. S., Anderson, L. A., & Antunes, F. M. (2015). The cortical modulation of stimulus-specific adaptation in the auditory midbrain and thalamus: A potential neuronal correlate for predictive coding. *Frontiers in Systems Neurosciences, 9,* 9–19.

Malmierca, M. S., Cristaudo, S., Pérez-González, D., & Covey, E. (2009). Stimulus-specific adaptation in the inferior colliculus of the anesthetized rat. *Journal of Neuroscience, 29*(17), 5483–5493.

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

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

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

McFadden, D., & Champlin, C. A. (1990). Reductions in overshoot during aspirin use. *The Journal of the Acoustical Society of America*, 87(6), 2634–2642.

Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America, 27,* 338–352.

Miller, J. D. (1989). Auditory-perceptual interpretation of the vowel. *Journal of the Acoustical Society of America, 85,* 2114–2134.

Miller, J. L. (1981). Effects of speaking rate on segmental distinctions. In P. D. Eimas & J. L. Miller (Eds.), Perspectives on the study of speech (pp. 39–74). Hillsdale: Erlbaum.

Miller, J. L., & Dexter, E. R. (1988). Effects of speaking rate and lexical status on phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, *14,* 369–378.

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

Minifie, F. D. (1973). Speech acoustics. In F. D. Minifie, T. J. Hixon, & F. Williams (Eds.), Normal aspects of speech, hearing, and language (pp. 235–284). Englewood Cliffs: Prentice Hall.

Moore, B. C. J., & Glasberg, B. R. (1983). Suggested formulas for calculating auditory-filter bandwidths and excitation patterns*. Journal of the Acoustical Society of America, 74,* 750–753.

Nassau, K. (1983). The physics and chemistry of color*.* Hoboken: John Wiley & Sons.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America, 85,* 2088–2113.

Nearey, T. M. (2010). A new non-linear regression model for formant trajectories in English monosyllables incorporating dual targets for vowels. *Journal of the Acoustical Society of America, 127,* 2020.

Ng, A. Y., & Jordan, M. I. (2002). *On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes.* In T. Dietterich, S. Becker, & Z. Ghahramani (Eds.), Proceedings of the Advances in Neural Information Processing (NIPS) Conference, 14. Available at https://papers.nips.cc/paper/2020-on-discriminative-vs-generative-classifiers-a-comparison-of-logistic-regression-and-naive-bayes

Nordström, P.-E., & Lindblom, B. (1975). A normalization procedure for vowel formant data. Proceedings of the Seventh International Congress of Phonetic Sciences, Leeds.

Norris, D., McQueen, J. M., Cutler, A., & Butterfield, S. (1997). The possible-word constraint in the segmentation of continuous speech. *Cognitive Psychology, 34*(3), 191–243.

Oja, E. (1982). A simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology, 15,* 267–273.

Okamura, M. (1966). Acoustical studies of Japanese vowels in children: The formant constructions and the developmental process. *Japanese Journal of Otolaryngology, 69,* 1198–1214.

Parker, E. M., Kluender, K. R., & Diehl, R. L. (1986). Trading relations in speech and nonspeech. *Perception & Psychophysics, 39,* 129–142.

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009). Statistical learning in a natural language by 8-month-old infants. *Child Development, 80*(3), 674–685.

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

Port, R. (2006). The graphical basis of phones and phonemes. In M. Munro & O.-S. Bohn (Eds.), Second language speech learning: The role of language experience in speech perception and production (pp. 349–365). Amsterdam: John Benjamins.

Rauschecker, J., Tian, B., & Hauser, M. (1995). Processing of complex sounds in the macaque nonprimary auditory cortex. *Science, 268*(5207), 111–114.

Roverud, E., & Strickland, E. A. (2010). The time course of cochlear gain reduction measured using a more efficient psychophysical technique. *The Journal of the Acoustical Society of America, 128,* 1203–1214.

Roverud E., & Strickland, E. A. (2014). Accounting for nonmonotonic precursor duration effects with gain reduction in the temporal window model. *The Journal of the Acoustical Society of America, 135,* 1321–1334.

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

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274*(5294), 1926–1918.

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition, 70*(1), 27–52.

Saffran, J. R., & Kirkham, N. Z. (2018). Infant statistical learning. *Annual Review of Psychology, 69,* 181–203.

Sanger, T. D. (1989). Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural Networks, 2,* 459–473.

Schouten, J. F. (1940). The residue and the mechanism of hearing. *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen, 43,* 991–999.

Schwartz, O., & Simoncelli, E. P. (2001). Natural signal statistics and sensory gain control. *Nature Neuroscience, 4,* 819–825.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27,* 379–423.

Siegelman, N., Bogaerts, L., Christiansen, M. H., & Frost, R. (2017). Towards a theory of individual differences in statistical learning. *Philosophical Transactions Royal Society of London B Biological Science, 372,* 1711.

Siegelman, N., Bogaerts, L., & Frost, R. (2016). Measuring individual differences in statistical learning: Current pitfalls and possible solutions. *Behavioral Research Methods,* 1–15. Advance online publication. https://doi.org/10.3758/s1342

Simoncelli, E. P. (2003). Vision and the statistics of the visual environment. *Current Opinions in Neurobiology, 13,* 144–149.

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience, 24,* 1193–1215.

Smith, R. L. (1977). Short-term adaptation in single auditory nerve fibers: Some poststimulatory effects. *Journal of Neurophysiology, 40*(5), 1098–1112.

Smith, R. L., & Zwislocki, J. (1975). Short-term adaptation and incremental responses in single auditory-nerve fibers. *Biological Cybernetics, 17*(3),169–182.

Stilp, C. E., Alexander, J. M., Kiefte, M., & Kluender, K. R. (2010a). Auditory color constancy: Calibration to reliable spectral properties across nonspeech context and targets. *Attention, Perception, & Psychophysics, 72,* 470–480.

Stilp, C. E., Kiefte, M., Alexander, J. M., & Kluender, K. R. (2010b). Cochlea-scaled spectral entropy predicts rate-invariant intelligibility of temporally distorted sentences. *Journal of the Acoustical Society of America, 128,* 2112–2126.

Stilp, C. E., Rogers, T. T., & Kluender, K. R. (2010c). Rapid efficient coding of correlated complex auditory properties. *Proceedings of the National Academy of Science, 107*(50), 21914–21919.

Stilp C.E. & Assgari A.A. (2019) Natural signal statistics shift speech sound categorization. Attention, Perception, & Psychophysics (in press)

Stilp, C. E., Kiefte, N., & Kluender, K. R. (2018). Discovering acoustic structure of novel sounds with varying predictability. *Journal of the Acoustical Society of America, 143,* 2460.

Stilp, C. E., & Kluender, K. R. (2010). Cochlea-scaled spectral entropy, not consonants, vowels, or time, best predicts speech intelligibility. *Proceedings of the National Academy of Science, 107*(27), 12387–12392.

Stilp, C. E., & Kluender, K. R. (2011). Non-iromorphism in efficient coding of complex sound properties. *Journal of the Acoustical Society of America, 130*(5), E1352–E1357.

Stilp, C.E., & Kluender, K.R. (2012). Efficient coding and statistically optimal weighting of covariance among acoustic attributes in novel sounds. *PLoS ONE* 7(1), e30845. https://doi.org/10.1371/journal.pone.0030845

Stilp, C. E., & Kluender, K. R. (2016) Stimulus statistics change sounds from near-indiscriminable to hyperdiscriminable. *PLOS ONE, 11*(8), e0161001.

Stilp, C.E., Anderson, P.W., Assgari, A.A., Ellis, G.M., & Zahorik, P. (2016). Speech perception adjusts to reliable spectrotemporal properties in the listening environment. *Hearing Research, 341*, 168–178.

Strickland, E. A. (2001). The relationship between frequency selectivity and overshoot. *Journal of the Acoustical Society of America, 109,* 2062–2073.

Sussman, H. M., Fruchter, D., Hilbert, J., & Sirosh, J. (1998). Linear correlates in the speech signal: The orderly output constraint. *Behavioral and Brain Sciences, 21*(2), 241–259.

Syrdal, A. K., & Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *The Journal of the Acoustical Society of America, 79*(4), 1086–1100.

Tian, B., & Rauschecker, J. P. (2004). Processing of frequency-modulated sounds in the lateral auditory belt cortex of the rhesus monkey. *Journal of Neurophysiology, 92,* 2993–3013.

Trubetzkoy, N. S. (1969). *Principles of phonology* (C. Baltaxe, Trans.). Berkeley: University of California Press. (Original work published in 1939)

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

Vapnik, V. N. (1998). Statistical learning theory. New York: John Wiley & Sons.

Viemeister, N. F. (1980). Adaptation of masking. In G. van den Brink & F. A. Bilsen (Eds.), Psychophysical, physiological and behavioral studies in hearing (pp. 190–198). Delft: Delft University Press.

Viemeister, N. F., & Bacon, S. P. (1982). Forward masking by enhanced components in harmonic complexes. *The Journal of the Acoustical Society of America, 71,* 1502–1507.

Viswanathan, N., Fowler, C. A., & Magnuson, J. S. (2009). A critical examination of the spectral contrast account of compensation for coarticulation. *Psychonomic Bulletin and Review, 16,* 74–79.

Viswanathan, N., Magnuson, J. S., & Fowler, C. A. (2010). Compensation for coarticulation: Disentangling auditory and gestural theories of perception of coarticulatory effects in speech. *Journal of Experimental Psychology: Human Perception and Performance, 36,* 1005–1015.

Viswanathan, N., Magnuson, J. S., & Fowler, C. A. (2013). Similar response patterns do not imply identical origins: An energetic masking account of nonspeech effects in compensation for coarticulation. *Journal of Experimental Psychology: Human Perception and Performance, 39*(4), 1181–1192.

Viswanathan, N., Magnuson, J. S., & Fowler, C. A. (2014). Information for coarticulation: Static signal properties or formant dynamics? *Journal of Experimental Psychology: Human Perception and Performance, 40,* 1228–1236.

von Klitzing, R., & Kohlrausch, A. (1994). Effects of masker level on overshoot in running- and frozen-noise maskers. *Journal of the Acoustical Society of America, 95,* 2192–2201.

Vorperian, H.K., Kent, R.D., Gentry, L.R. & Yandell, B.S. (1999). Magnetic resonance imaging procedures to study the concurrent anatomic development of vocal tract structures: Preliminary results. *International Journal of Pediatric Otorhinolaryngology, 49*(3), 197–206.

Vorperian, H. K., Kent, R. D., Lindstrom, M. J., Kalina, C. M, Gentry, L. R., & Yandell, B. S. (2005). Development of vocal tract length during early childhood: A magnetic resonance imaging study. *The Journal of the Acoustical Society of America, 117,* 338–350.

Vorperian, H. K., Wang, S., Chung, M. K., Schimek, E. M., Durtschi, R. B., Kent, R. D., … Gentry, L. R. (2009). Anatomic development of the oral and pharyngeal portions of the vocal tract: An imaging study. *Journal of the Acoustical Society of America, 125*(3), 1666–1678.

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

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

Werker, J. F., Gilbert, J. H. V., Humphrey, K., & Tees, R. C. (1981). Developmental aspects of cross-language speech perception. *Child Development, 52,* 349–355.

Werker, J. F., & Lalonde, C. E. (1988). Cross-language speech perception: Initial capabilities and developmental change. *Developmental Psychology, 24,* 672–683.

Werker, J. F. & Logan, J. S. (1985). Cross-language evidence for three factors in speech perception. *Perception & Psychophysics, 37,* 35–44.

Werker, J. F., & Tees, R. C. (1983). Developmental changes across childhood in the perception of non-native speech sounds. *Canadian Journal of Psychology, 37,* 278–286.

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

Werker, J. F., & Tees, R. C. (1984b). Phonemic and phonetic factors in adult cross-language speech perception. *Journal of the Acoustical Society of America, 75,* 1866–1878.

Wessinger, C. M., VanMeter, J., Tian, B., Van Lare, J., Pekar, J., & Rauschecker, J. P. (2001). Hierarchical organization of the human auditory cortex revealed by functional magnetic resonance imaging. *Journal of Cognitive Neuroscience, 13*(1), 1–7.

Zwicker, E. (1961). Subdivision of the audible frequency range into critical bands. *The Journal of the Acoustical Society of America, 33*(2), 248–248.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.