

Perception of Vowel Sounds Within a Biologically Realistic Model of Efficient Coding

Keith R. Kluender, Christian E. Stilp and Michael Kieft

Abstract Predicated upon principles of information theory, efficient coding has proven valuable for understanding visual perception. Here, we illustrate how efficient coding provides a powerful explanatory framework for understanding speech perception. This framework dissolves debates about objects of perception, instead focusing on the objective of perception: optimizing information transmission between the environment and perceivers. A simple measure of physiologically significant information is shown to predict intelligibility of variable-rate speech and discriminability of vowel sounds. Reliable covariance between acoustic attributes in complex sounds, both speech and nonspeech, is demonstrated to be amply available in natural sounds and efficiently coded by listeners. An efficient coding framework provides a productive approach to answer questions concerning perception of vowel sounds (including vowel inherent spectral change), perception of speech, and perception most broadly.

Abbreviations

C	Consonant
CV	Consonant–vowel
CVC	Consonant–vowel–consonant

K. R. Kluender (✉)

Department of Speech, Language, and Hearing Sciences, Purdue University, West Lafayette, IN, USA
e-mail: kkluender@purdue.edu

C. E. Stilp

Department of Psychological and Brain Sciences, University of Louisville, Louisville, KY, USA

M. Kieft

School of Human Communication Disorders, Dalhousie University, Halifax, NS, Canada

CSE	Cochlea-scaled spectral entropy
ERB	Equivalent rectangular bandwidth
f0	Fundamental frequency
F1	First formant
F2	Second formant
F3	Third formant
JND	Just noticeable difference
PCA	Principal component analysis
r	Pearson product-moment correlation coefficient
TIMIT	Texas Instruments/Massachusetts Institute of Technology
V	Vowel
VC	Vowel-consonant
VISC	Vowel inherent spectral change

1 Introduction

In this contribution, we offer a perspective on perception of speech, and of vowel sounds in particular, that is motivated by broad principles that apply to sensation and perception across all modalities. Our approach, referred to as ‘efficient coding’ in models of visual perception, is situated within a half-century tradition of information-theoretic approaches that remain fruitful today (e.g., Attneave 1954; Barlow 1961; Fairhall et al. 2001; Simoncelli 2003; Clifford et al. 2007). This information-theoretic perspective dissolves some longstanding debates and distinctions while revealing new insights into processes of speech perception that have been neglected or obscured by more traditional approaches to auditory perception.

The organization of this chapter begins with consideration of speech perception within a broad perspective on perception spanning all modalities. Within this framework, vowel perception can provide both examples of efficient coding and tests of the power of such an approach.

2 Objects of Perception

An enduring distraction for investigators studying speech perception has concerned the extent to which objects of speech perception are defined as articulatory gestures or patterns (e.g., Liberman and Mattingly 1985; Fowler 1986), acoustic patterns, patterns of sensory stimulation (e.g., Diehl and Kluender 1989), or some combination (e.g., Stevens and Blumstein 1981; Nearey 1997). Oftentimes, controversies concerning appropriate objects of perception generated more heat than light. We suggest that debates concerning objects of perception cannot be resolved because the question itself is ill-posed, if not outright misleading.

Unconvinced by arguments exclusively for gestures or acoustics as proper objects of speech perception, Nearey (1997) proposed what he described as a “double-weak” model of speech perception—a judicious combination of both articulatory and acoustic/auditory characterizations of speech perception. In the present approach founded upon information-theoretic principles, we go one step further. We make the claim that there are no *objects* of perception, neither for speech nor for perception in general. Instead, there is an *objective* for perception, which is to maintain adequate agreement between an organism and its world in order to facilitate adaptive behavior.

Within this functional framework, 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 or proximal properties falls short of capturing the essential functional characteristic of perception—the relationship between an organism’s world and its behavior.

If there are no objects *of* perception, how should one think about information *for* perception? Information for perception does not exist in the objects and events in the world, nor does it exist in the brain of the perceiver. Instead, information exists in the relationship between an organism and its world. It may be useful to consider the contrast between information *about* and information *for*. When one discusses objects of perception, it is information *about* that is typically inferred. Implicit in such efforts is the notion that one needs to solve the inverse problem; how does one satisfactorily represent the world within one’s brain? By contrast, if the objective of a successful perceptual system is to maintain adequate agreement between an organism and its world in order to facilitate adaptive behavior, then information *for* successful perception is nothing more or less than information that resides in this relationship (or agreement).

3 Shannon Information Theory

This way of viewing information is consistent with one of the fundamental characteristics of Shannon information theory (Shannon 1948; Weaver 1948). Some readers may be familiar with Fletcher’s pioneering applications of information theory to speech (Fletcher 1953/1995). However, our application here will be more akin to early approaches of vision researchers such as Attneave (1954, 1959) and Barlow (1961), as these efforts continue to be productive in contemporary theories of ‘efficient coding’ (e.g., Barlow 1997, 2001; Simoncelli and Olshausen 2001; Schwartz and Simoncelli 2001; Simoncelli 2003; Stilp et al. 2010c). Developed at Bell Laboratories for practical application to telephone bandwidth, one 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* between organisms and their environments.

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 world. 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. The greater the number of alternatives there are (uncertainty, unpredictability, variability, or entropy), the greater the amount of information that potentially can be transmitted. There is no information when there is no variability. When there is no variability, there is total predictability and hence, no information.

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 et al. 2003). Perceptual systems do not record absolute levels whether loudness, pitch, brightness, or color. Relative change is the coin of the realm for perception, a fact known at least since Ernst Weber in the mid-18th century. Sacrifice of absolute encoding has enormous benefits along the way to optimizing information transmission. For example, biological transducers have impressive dynamic range given their evolution via borrowed parts (e.g., gill arches to middle ear bones); however, this 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 shifts 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 chapter.

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, such as acoustics of concert halls or across trials in an experimental session. 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 context and information for perception more generally, traditional distinctions between sensation, perception, and learning diffuse along a series of processes that operate over broader ranges of time and experience. From 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.

4 Potential Information and Intelligibility

We begin by considering the importance of sensory change at the lowest levels of the auditory system. Stilp and Kluender (2010) recently evaluated the extent to which measures of sensory change, tailored by the cochlea, may serve to explain intelligibility of connected speech. Earlier efforts employed orthodox descriptions of speech signals (strings of consonant and vowel sounds) in studies conducted to evaluate relative contributions of vowel versus consonant portions of the speech stream. When intervals corresponding to consonants or vowels were replaced with noise, vowels appeared to provide more information (contribute more to sentence intelligibility) than consonant sounds (Cole et al. 1996; Kewley-Port et al. 2007; Fogerty and Kewley-Port 2009, but see Owren and Cardillo 2006). Because listeners were better at understanding sentences with consonants replaced than with vowels replaced, a “vowel superiority effect” had been suggested to exist.

Interpreting putative vowel superiority effects is not straightforward. Experiments in which consonants or vowels are replaced by noise rely upon operationally defined temporal boundaries provided by phoneticians to demarcate consonants and vowels in the Texas Instruments/Massachusetts Institute of Technology (TIMIT) database of sentences (Garofolo et al. 1990). While neither the phoneticians who provide these boundaries nor the investigators who use them in experiments are naïve about shortcomings of such demarcations, the fact that no single point in time exists when a consonant ends and a vowel begins in connected speech presents a challenge to interpreting data suggesting vowel superiority. Production of every consonant and vowel overlaps in time (and in the vocal tract) with speech sounds that precede and follow. In addition, acoustic characteristics within vowels provide rich evidence concerning neighboring consonants and *vice versa* (e.g., Liberman et al. 1957; Jenkins et al. 1983; Sussman et al. 1991; Kieft 2000).

A second complication arises from the fact that, at least as delineated in TIMIT, vowel intervals of sentence waveforms are roughly one-third longer than those for consonants. Recently, Lee and Kewley-Port (2009) provided evidence that there may be no intelligibility differences between replacing different combinations of consonant (C) and vowel (V) portions with noise after accounting for duration. Stilp and Kluender (2010) also found that replacement of Vs, CVs, and VCs resulted in closely similar declines in performance, although, they found that relatively greater proportions of Cs could be replaced, relative to other conditions, for performance to decline.

It seems unlikely that most speech intervals are perceptually equivalent in terms of information, and that only duration—not acoustic composition—matters. Stilp and Kluender (2010) tested whether a metric other than consonants versus vowels could better account for intelligibility data. They began by abandoning phonetic designations. Instead of considering consonants, vowels, or their combinations, they tested the degree to which speech intelligibility depends upon amount of potential information, defined psychoacoustically, that is removed and replaced by noise. The simple fundamental principle that perceptual systems respond primarily to change

has the formal consequence of enhancing information transmission. Because there is no new information when events either do not change or are predictable, one can employ relative change as an approximate measure of potential information. Cochlea-scaled spectral entropy (CSE) is a measure of relative spectral change across time, operationalized as the extent to which successive spectral slices differ (i.e., cannot be predicted) from preceding spectral slices.

CSE is quantified as Euclidean distances between adjacent psychoacoustically-scaled spectral slices (Fig. 1). Sentences were RMS-intensity-normalized and divided into 16 ms frames independent of TIMIT segmentations. Slices were passed through 33 filters that capture nonlinear weighting and frequency distribution along the cochlea (Patterson et al. 1982). Filters were spaced one equivalent rectangular bandwidth (ERB) apart up to 8 kHz. ERBs provide a close approximation to tonotopic distribution along the cochlea (Greenwood 1990) and psychophysical auditory filters derived from normal-hearing listeners (Glasberg and Moore 1990). Consequently, each filter corresponds to an equivalent number of inner hair cells and neurons in the auditory nerve.

The ERB scale is roughly logarithmic except in low frequencies where it is more nearly linear. Euclidean distances between adjacent 16 ms slices were calculated across the 33 filter output levels. Distances were then summed in boxcars of either five (80 ms, approximate mean consonant duration) or seven successive

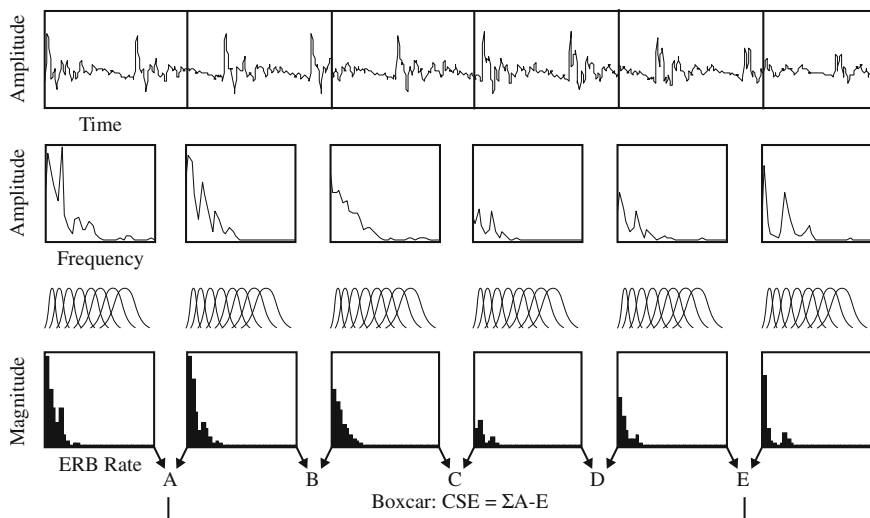


Fig. 1 Calculation of CSE using an 80 ms boxcar. *First row* Original speech waveform, with amplitude plotted as a function of time. Each box corresponds to a 16 ms slice. *Second row* Fourier transform of each 16 ms slice, with amplitude plotted as a function of frequency. *Third row* Auditory filter bank. *Fourth row* Weighted outputs of auditory filter processing with magnitude plotted as a function of ERB rate. Euclidean distances are calculated between each slice, then summed in boxcars

slices (112 ms, approximate mean vowel duration). Cumulative Euclidean distances within a boxcar were taken as measures of spectral entropy and served as a psychoacoustic metric of potential information.

After convolving boxcars of summed distances across entire sentences, entropy measures were sorted into ascending order (Low Entropy condition), descending order (High Entropy), or ascending absolute difference from median boxcar value (Medium Entropy). The boxcar ranked first (lowest, highest, or median CSE) was replaced by speech-shaped noise matched to mean sentence level. Eighty milliseconds before and after selected boxcars were preserved to avoid boxcars overlapping, and the first 80 ms of every sentence was always left intact. The procedure continued iteratively to the next-highest-ranked boxcar, which was replaced only if its content had not already been replaced or preserved.

As expected, replacing longer (112 ms) segments impaired performance more on average than replacing shorter (80 ms) segments. Amount of potential information (CSE), however, plays a much larger role in predicting performance (Fig. 2). Intelligibility closely follows measures of entropy replaced with noise ($r^2 = 0.80, p < 0.01$).

5 Auditory Entropy and the Sonority Hierarchy

There are informative systematicities in the types of speech sounds that have higher and lower CSE (Stilp and Kluender 2010). More consonants than vowels are replaced in low-entropy conditions, and more vowels than consonants are replaced in high-entropy conditions. Although significantly more vowels were replaced with each increase in CSE, proportion of vowels or consonants replaced are not significant predictors of intelligibility ($r^2 = 0.55, n.s.$).

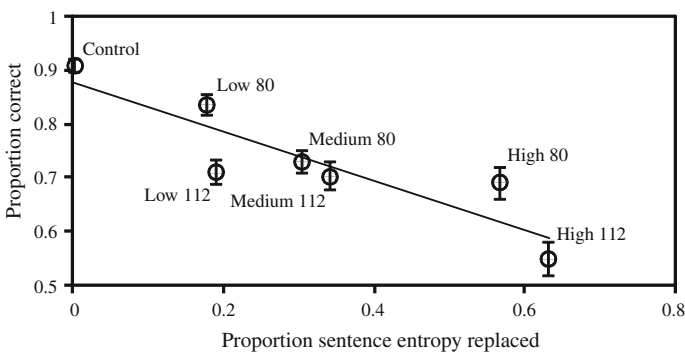


Fig. 2 Results of Stilp and Kluender (2010). Sentence intelligibility (proportion of words correctly identified) is plotted as a function of proportion of sentence entropy (CSE) replaced by noise. Data points are labeled by level of CSE replaced by noise (*Low, Medium, High*) and boxcar duration (80, 112 ms). Potential information is a significant predictor of sentence intelligibility ($r^2 = 0.80, p < 0.01$). *Error bars* represent standard error of the mean

Phonetic compositions of replaced speech segments were analyzed according to vocal tract configuration (vowels) or manner of articulation (consonants). Vowels were subdivided into high (close) versus low (open), front versus central or back, and diphthongs. As one would expect, acoustic segments traditionally labeled as diphthongs were replaced most often in high-entropy conditions and least often in low-entropy conditions. Low vowels are replaced most often in high-entropy conditions and least often in low-entropy conditions, while high vowels show the opposite pattern. There were no systematic differences between front, central, or back vowels.

Consonants were classified as plosives, affricates, closure silence (preceding plosives and affricates), fricatives, laterals/glides, and nasals. Consonants replaced by noise differed substantially across conditions. The most vowel-like consonants—nasals and laterals/glides—were most-often replaced in high-entropy conditions and least-often replaced in low-entropy conditions. The least-vowel-like plosives and affricates showed the complementary pattern. Only modest changes were observed for fricatives and closures across conditions.

This pattern of CSE decreasing from non-high vowels, to high vowels, to laterals/glides and nasals, to fricatives, to affricates and finally plosives closely parallels the sonority hierarchy. The linguistic construct of sonority (or vowel-likeness) is useful for describing phonological systematicity, especially syllable composition; however, to date sonority has been resistant to clear definition in acoustics or articulation (Ohala 1990; Clements 2009).

As an explicit test of the extent to which CSE corresponds to sonority, CSE was measured in VCV recordings from six adult talkers (three female, three male). Twenty American English consonant sounds (/b, d, g, p, t, k, m, n, l, r, w, f, v, θ, ð, s, z, ʃ, tʃ, dʒ/) were flanked by each of three vowels (/a, i, u/), generating 60 VCVs per talker. Recordings were truncated to maintain consistent overall duration (≈ 460 ms). Waveforms were normalized so that amplitudes of the vowels reflected differences in naturally spoken vowels, and onsets and offsets of VCVs (64 ms) were excluded from analysis. Averaged across talkers and consonants, there was more CSE in low-vowel contexts (/aCa/) than in high-vowel contexts (/iCi/, /uCu/; Fig. 3a). Across talkers and all VCVs, laterals and glides had the highest CSE, followed by nasals, fricatives, affricates, and plosives (Fig. 3b). Patterns of CSE for both vowel and consonant analyses follow the sonority hierarchy, corroborating findings from the sentence intelligibility experiment.

Phonetic composition of high- versus low-entropy segments initially may be surprising. One might expect that consonants, created by vocal tract constrictions with correspondingly rapid acoustic changes, would have higher entropy than vowels. However, vowels and vowel-like sounds have greater cochlea-scaled entropy because roughly logarithmic psychoacoustic and physiologic indices weight lower-frequency spectral prominences (formants, F_1 and F_2) more. This explanation depends, however, upon sufficient levels of spectral change in vowel and vowel-like sounds. Multiple contributions to the present volume provide ample testament to the fact that, indeed, vowel formants change substantially over time with few exceptions (Nearey and Assmann 1986), and this vowel inherent

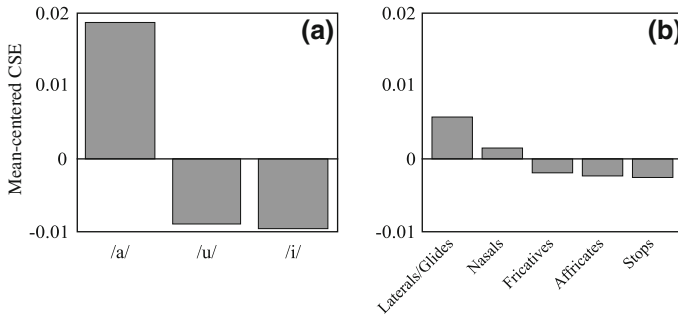


Fig. 3 CSE analyses of VCVs in Stilp and Kluender (2010) corroborating the relationship between CSE and sonority. **a** Measures of vowel CSE averaged across consonants. **b** Measures of consonant CSE averaged across vowels

spectral change (VISIC) is important to listeners (Hillenbrand and Nearey 1999; Bunton and Story 2010).

Similarly, Kiefe and Kluender (2005) found that perception of synthetic monophthongs, relatively poor renditions of real speech, was unrepresentative of normative speech perception. They showed how listeners could rely on putatively unreliable gross acoustic properties to identify monophthongs. Under some conditions, listeners identified monophthongs on the basis of spectral tilt, which is highly vulnerable to transmission channel characteristics such as room acoustics (Klatt 1982). In such cases, listeners responded to stimuli in a manner similar to hearing-impaired listeners (Lindholm et al. 1988; Alexander and Kluender 2008). By contrast, Kiefe and Kluender (2005) found that listeners do not use tilt when provided the opportunity to use changing spectral composition, as formant transitions may serve to enhance perceptibility of formants themselves.

Some correspondence between CSE and VISIC is expected, but there are important differences between the two measures. Because CSE is an accumulation of brief slice-to-slice changes across the full cochlea-scaled spectrum, it is not expected to map perfectly onto simpler characterizations of F_1 and F_2 center frequencies at only two points in time. Nevertheless, vowels displaying inherent spectral change (diphthongs and low but not high vowels) displayed the highest measures of CSE in Stilp and Kluender (2010), so the metric appears to be sensitive to some aspects of VISIC. To evaluate the relationship between CSE and VISIC, we measured cumulative CSE across all vowels spoken by men, women, and child talkers in the database (Hillenbrand et al. 1995) used by Hillenbrand and Nearey (1999). As an index of VISIC, we calculated vector lengths in the ERB-scaled F_1 - F_2 plane for each English vowel (square root of the sum of squared changes in F_1 and F_2 frequencies between 20 and 80 % time points) using formant center frequencies reported by Hillenbrand et al. (1995). Longer vector lengths correspond to more VISIC between these two time points. Cumulative CSE was calculated over the same (20–80 %) intervals. Vector length and CSE share a significant correlation ($r = 0.45$; $p < 0.01$). Given this relationship between VISIC

and CSE, and the extent to which CSE accounts for sentence intelligibility (Stilp and Kluender 2010), VISIC may be important beyond characterization of vowels *per se*.

6 VISIC as Adaptive Dispersion

Our experiments employing noise replacement highlighted the importance not only of vowels, but of changes in spectral composition of vowels for sentence understanding. The more vowel-like a segment of the signal was, the more essential that segment was for intelligibility. This finding explicitly relies on the fact that vowel spectra, and their cochlear consequences, change across time. Further inspection of the nature of these vowel inherent spectral changes reveals that, in addition to increasing psychoacoustic potential information, VISIC may serve to make it easier for listeners to detect differences between vowel sounds.

Different languages use different inventories of vowel sounds, and languages use subsets of vowels that are most easily discriminated from one another. For example, those vowels favored by languages with five vowels are ones that are as acoustically distinct as possible from one another. As a general rule, the set of vowels selected by a language, whether it uses three or ten vowels, is comprised of sounds that tend toward maximal distinctiveness (Liljencrants and Lindblom 1972; Bladon and Lindblom 1981). Lindblom (1986) refers to this fact as ‘adaptive dispersion.’ However, these demonstrations suffer from considering vowel sounds only as relatively static entities by employing steady state formants. To what extent are realistic portrayals of vowels, possessing inherent spectral change, consistent with adaptive dispersion?

To address this question, we begin by plotting formant measurements in cochleotopic (ERB) coordinates for average F_1 and F_2 at 20 % time points drawn from the Hillenbrand et al. (1995) database for western Michigan male talkers (Fig. 4, left). We make a rough division of these F_1/F_2 points into front/back and high/low regions of the vowel space. Next, we capture VISIC by plotting directional vectors (length proportional to spectral change) with the origin being 20 % measurements in F_1/F_2 (Fig. 4, right). In general, formants with starting points closer to one another (in the same F_1/F_2 quadrant, left) tend to disperse such that the direction of spectral change is distinct from that for other vowels with neighboring starting points. This is most obviously true for closely neighboring vowels /i/ and /e/, /e/ and /æ/, and /o/ and /u/. Kinematic dispersion is less substantial in this quadrant-based analysis for /ʌ/ versus neighboring /ɔ/. Neither point vowel /i/ nor /u/ exhibits much VISIC; however, they are quite distant from each other and have no close neighbors (Fig. 4, left). Finally, unlike the relatively stationary /i/ and /u/, spectral composition of the third point vowel /a/ moves toward the center of the vowel space in a way not predicted by dispersion, at least not for F_1 and F_2 alone.

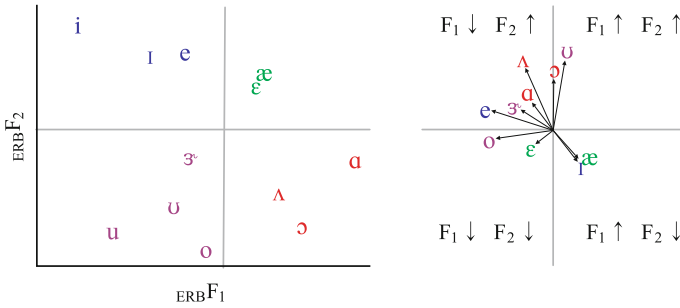


Fig. 4 (Left) Mean F_1/F_2 values of vowels spoken by male talkers in Hillenbrand and Nearey (1999). Phonetic symbols are placed at formant values measured at 20 % of overall vowel duration. Grey lines represent coarse distinctions between *high/low* and *front/back*. (Right) VISC measures in the same vowels. Arrow tails are formant measurements at 20 % of overall duration, and are aligned at the origin of the graph. Arrow heads are formant measurements at 80 % of overall duration, and arrow length is proportional to VISC. Quadrants are labeled according to F_1 and F_2 change over time. Vowels /i/ and /u/ are not included as both display minimal VISC and would lie very close to the origin

Perceptual data support the importance of VISC for distinguishing vowel sounds from one another. For example, Nearey and Assmann (1986) presented listeners three combinations of vowel fragments separated by 10 ms of silence. These were: *natural* (segment of vowel ‘nucleus’ followed by offglide); *repeated* (nucleus followed by itself); and, *reverse* (offglide followed by nucleus). Most telling are cases for which kinematics are reversed. Listeners were more likely to identify reversed /ɪ/ as /e/ and reversed /e/ as /ɪ/, and reversed /ɛ/, /æ/, /o/, and /ʊ/ were among the most likely vowels to be confused with others. When Hillenbrand and Nearey (1999) synthesized vowels with either natural kinematics (VISC) or with flat formants fixed at the frequencies of the least kinematic (quasi ‘steady-state’) portion of vowel sounds, the vowel sounds most likely to be misidentified were /e/, /o/, /ʊ/, /u/, and /ʌ/. Vowels synthesized based upon /e/ were most likely to be confused with /ɪ/ and /ɛ/, /o/ with /ɔ/, /ʊ/ with /ʌ/ and /o/, /u/ with /o/, and /ʌ/ and /ʊ/. Most recently, Bunton and Story (2010) synthesized vowels using both static and time-varying speaker-dependent area functions based on magnetic resonance imaging measurements of the vocal tract (see also Story and Bunton 2012 this volume). Performance was much more consistent for vowel sounds created with time-varying spectra, a pattern of data very similar to that found by Hillenbrand and Nearey (1999). Static vowels synthesized based upon /ɪ/ were most likely to be confused with /e/ and /ɛ/, /e/ with /ɛ/ and /ɪ/, /ɛ/ with /æ/, /ʌ/ with /ʊ/ and /ɑ/, /ɑ/ with /ɔ/, /ɔ/ with /ɑ/ and /o/, /o/ with /ʊ/, and /ʊ/ with /u/ and /o/.

Taken together, VISC serves to make vowels more perceptually distinctive in at least two ways. First, spectral change in itself capitalizes on a fundamental design aspect of sensorineural systems, this being near exclusive sensitivity to change. As demonstrated earlier, the importance of spectral change in vowel sounds is all the more remarkable when one captures this change with respect to cochlear

organization. Second, vowel inherent spectral change is not change for the sake of change itself, however important change is to sensorineural processing. Instead, VISIC is by and large sensible when one conceptualizes the most important property of vowel sounds to be the extent to which distinctions between vowel sounds are enhanced. Vowel inherent spectral change enhances distinctiveness by providing an additional means of distinguishing vowels from one another. In most instances, when vowels begin relatively close in perceptual space, they quickly proceed to increase ‘otherness’ in trajectories of spectral shape change.

7 Information and Rate ‘Normalization’

When one adopts relative change as a metric for perceptual information, some other traditional measures, beyond linguistic constructs such as consonants and vowels, are left behind. For example, relative change in the cochlea-scaled spectrum does more than simply morph the frequency axis. When relative change is employed, units corresponding to absolute frequency and intensity are discarded. To the extent that change is a fundamentally unitless measure, absolute time also ought not matter within limits. In Stilp and Kluender’s (2010) experiment above, absolute signal duration replaced by noise did not explain performance. 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 predicted on the basis of amount of relative change more or less independent of the time course over which that information is accrued.

Stilp et al. (2010b) conducted a series of experiments in which measures of relative change were used to predict sentence intelligibility across wide variation in rate of speech. They synthesized 115 seven-syllable sentences from the Hearing In Noise Test (HINT; Nilsson et al. 1994) at three different speaking rates: slow, medium, and fast (2.5, 5.0, and 10.0 syllables per second, respectively). Next, they time-reversed equal-duration segments (20, 40, 80, and 160 ms) at the nearest zero crossings for every sentence (see e.g., Saberi and Perrott 1999). As seen in Fig. 5 (right), listener performance across conditions is very well predicted based upon proportion of the utterance distorted (reversal duration) and not absolute duration (left).

CSE functions peak at roughly two-thirds of mean syllable duration (64 ms for fast sentences, 128 ms for medium, and 256 ms for slow) 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 (Fig. 6, right). For English VCVs, the identity of the second vowel is 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

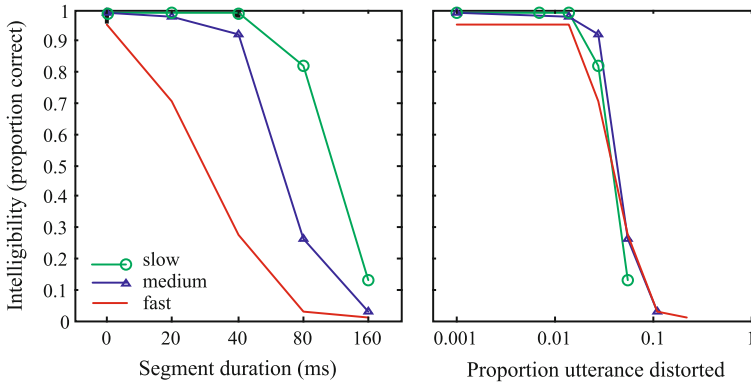


Fig. 5 Results from Stilp et al. (2010b). (Left) Intelligibility of sentences at a wide range of speaking rates when fixed-duration segments were temporally reversed. Performance declines faster (slower) for sentences at faster (slower) rates relative to medium-rate speech. (Right) Data converge to a common function when plotted using the relative measure of proportion of utterance distorted (segment duration divided by mean sentence duration)

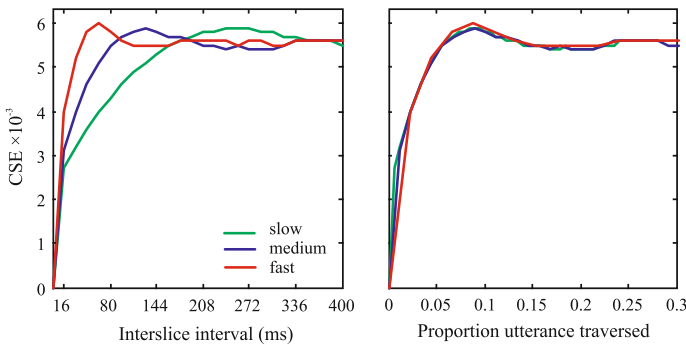


Fig. 6 CSE analyses of variable-rate sentences from Stilp et al. (2010b). (Left) CSE measured in 16 ms slices with increasing interslice intervals. Fast-rate speech peaks first (64 ms), followed by medium (128 ms), then slow (256 ms). Distances regress to the mean spectral distance between any two slices 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

information conveyed by spectral change accounts for a substantial proportion of variance in listener performance across all rate conditions ($r^2 = 0.89$; $p < 0.001$).

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 materials. There have been substantial efforts to better understand how listeners normalize across speaking rate when identifying individual consonants (e.g., Miller and Liberman 1979; Miller 1981), vowels (e.g., Ainsworth 1972, 1974; Gottfried et al. 1990), or words (e.g., Miller and Dexter

1988), and all of these efforts have concentrated upon absolute physical changes in frequency and time. 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.

8 When Acoustics are Predictable

To this point, discussion has concerned relatively local effects of context, as we have addressed changes in frequency on the order of hertz and time over the course of milliseconds. Over the past decade, multiple studies have been conducted to investigate processes by which the auditory system compensates for reliable spectral and temporal characteristics of a sound source under different listening conditions. Here, attention is drawn to the full spectrum across much longer intervals. It has become quite clear that the auditory system calibrates to different listening environments by perceptually compensating for or absorbing reliable, redundant spectral features of the acoustic input (Watkins 1991; Watkins and Makin 1994; Kiefte and Kluender 2008; Alexander and Kluender 2010). The advantage of such processes is that they optimize sensitivity to new (unpredictable) information.

Kiefte and Kluender (2008) conducted experiments designed to assess relative contributions of spectrally global (spectral tilt) versus local (spectral peak) characteristics of a listening context. They varied both spectral tilt and center frequency of F_2 to generate a matrix of steady-state vowel sounds that perceptually varied from /u/ to /i/. Listeners identified these sounds following filtered natural or time-reversed precursor sentences. When either set of precursor sentences was 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, as listeners predominantly used the less predictable spectral property that was not consistent within the precursor.

Stilp and colleagues (Stilp et al. 2010a) demonstrated that such compensation appears to be relatively automatic and naïve to the source of sounds. They found that listeners compensate for reliable spectral properties of a listening context when identifying musical instruments. Perception of tenor saxophone versus French horn adjusts following a filtered passage of speech or of other musical instruments (Schubert string quintet). This adjustment for spectral characteristics of a listening context is closely analogous to visual color constancy through which colors appear relative constant despite widely varying reflectances under different types of illumination (Stilp et al. 2010a).

9 Efficient Coding of Second-Order Statistics

It is clear that the auditory system absorbs predictability within a listening context, and the result is increased sensitivity to information which, by definition, is something that is not predictable. Much predictability in sounds is not about listening contexts, but instead is inherent in the ways sounds are structured. Natural sounds are complex and are typically changing along multiple acoustic dimensions. For sounds created by real structures including musical instruments and vocal tracts, changes in different acoustic dimensions cohere in accordance with physical laws governing sound-producing sources. For example, articulatory maneuvers that produce consonant and vowel sounds give rise to multiple acoustic attributes. This redundancy across attributes contributes to robust speech perception despite substantial signal degradation (Assmann and Summerfield 2004; Kluender and Kiefte 2006; Kluender and Alexander 2007). For example, because the relationship between formant frequency and amplitude is determined entirely by the physics of speech production (Fant 1956; Stevens 1998), it is to be expected that listeners are relatively insensitive to variations in one property when the other provides more reliable information regarding vowel identity. In support of this, Kiefte et al. (2010) have shown that listeners are indeed insensitive to variations in formant amplitude—a relatively unstable acoustic property across different listening contexts (e.g., Kiefte and Kluender 2005, 2008). They further show that any evidence for an important role for formant amplitude in vowel perception (e.g., Aaltonen 1985; Hedrick and Nábělek 2004) can be explained entirely by peripheral auditory effects such as simultaneous masking.

Such redundancy in speech epitomizes the general fact that objects and events in the world have structure. Attneave (1954) emphasizes how information received by the visual system is 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 “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 “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 covariance 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 a way that absorbs 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: (1) makes it easier to detect newly appearing associations resulting from new causal factors in the environment; (2) provides a role for recurrent collaterals, which are a conspicuous feature of cortical neurons; and, (3) could account for part of the effects of experience during cortical development. To this list, 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 correlated activity across synapses. Finally, there is 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).

There are multiple recent findings concerning the ways through which the visual perceptual system exploits redundancies among optical attributes. For example, models that capture edge co-occurrence in natural images precisely predict observer performance in a contour grouping task (Geisler et al. 2001). Perhaps the most impressive instances of efficient coding are perceptual “metamers” which are composite stimuli that cannot be discriminated even though individual properties can be discriminated when presented in isolation (Backus 2002). For example, perception of visual slant is cued by both binocular disparity and texture gradient among other cues (Hillis et al. 2002). In nature, these two cues are highly correlated, and when altered in an experimental setting, observers are incapable of discriminating either binocular disparity or texture gradient independent of the other cue.

10 Second-Order Statistics in Speech

It is well-attested that all contrasts between speech sounds are multiply specified. For example, the distinction between medial /b/ and /p/ includes at least sixteen different acoustic differences (Lisker 1978). No single attribute is, in itself, both necessary and sufficient to support perception of /b/ or /p/, which instead relies upon combinations of attributes. Kluender and colleagues (Kluender and Lotto 1999; Kluender and Kiefte 2006; Kluender and Alexander 2007) 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.

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 unique among perceptual achievements. However, Kluender et al. (1987) demonstrated that birds could learn the mapping for /d/ versus /b/ and /g/.

In order to account for some of the variability for place of articulation across vowel contexts, Sussman and colleagues (e.g., Sussman et al. 1998) reintroduced the idea of locus equations (Delattre et al. 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 ($r = 0.75 - 0.96$).

Kieft (2000) similarly measured formant transitions for prevocalic plosives consonants /b/, /d/, and /g/ followed by each of ten Western Canadian English vowels /i, I, e, ε, æ, a, Λ, o, u, u/ preceding syllable-final /k/, /tʃ/ and /l/. These CVCs were spoken by five male and six female adults. Correlations between F_2 values measured at the end of the second glottal pulse after onset of periodicity and F_2 60 ms later were $r = 0.98$ for /b/, $r = 0.87$ for /d/, and $r = 0.95$ for /g/. These relational patterns are consistent with Sussman and colleagues' earlier observations, and coefficients are highly similar when ERB-scaled values are used.

We asked whether vowels, like consonants, can be similarly characterized by reliable relationships between F_2 values as a function of time. Using the same data set (Kieft 2000), we analyzed ERB formant trajectories at the same time points for the ten vowels following consonants /b/, /d/, and /g/. 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 $r = 0.82$. 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 is unknown at present. Very recently, Nearey (2012 this volume) demonstrated that, at least for cases tested thus far (Hillenbrand et al. 2001), it is possible to decompose CVC syllables into locus constituents (CV and VC) and kinematic representations of vowels (VISC). Success thus far suggests that useful relational properties for Cs and Vs are separable and available to listeners.

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 our 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.

11 Learning Correlated Attributes

Later in this contribution, we will present other examples of naturally occurring covariance relationships between acoustic attributes of vowel sounds as they are created by talkers and their vocal tracts. First, one should ask whether and how listeners detect and exploit redundancy between stimulus attributes. 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 more about acquisition of sensitivity to correlations among stimulus attributes by adult listeners, Stilp et al. (2010c) designed novel complex stimuli that varied across two physically independent acoustic attributes: attack/decay (AD) and spectral shape (SS). SS was varied via summation of two instrument endpoints (French horn, tenor saxophone) in different proportions. In principle, AD and SS are relatively independent both perceptually and in early neural encoding (Caclin et al. 2006). Physically complex attributes were chosen with the expectation that more complex attributes should be more plastic relative to elemental properties such as frequency, which serves as a primitive dimension in the tonotopically organized auditory system.

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. Two stimulus subsets were selected from this matrix, each rotated 90° from the other, capturing near-perfect correlations between the two acoustic cues (Fig. 7a: $r_{AD,SS} = \pm 0.97$). Half of forty listeners completed a discrimination task with pairs of stimuli drawn from the distribution shown, and the other 20 listeners discriminated pairs drawn from the rotated subset. Across three blocks of 144 trials within a single experimental session, listeners discriminated sound pairs of three types: (1) consistent with the correlation (Consistent condition; blue); (2) orthogonal to experienced correlation (Orthogonal condition; red); and (3) differing by equal magnitude along only one stimulus dimension (Single-cue condition; green). Test sounds were presented in two-alternative forced choice (AXB) trials using 18 pairs of equally distant stimuli [three steps of AD (one pair), SS (one pair), or AD and SS (15 Consistent plus one Orthogonal pair).]

No feedback was provided. Performance across the three blocks of trials changed in highly informative ways (Fig. 7b). Discrimination of Consistent pairs that respect the correlation began well and remained relatively high throughout the experiment; however, performance on Orthogonal trials was significantly inferior early in testing. Discrimination of stimulus pairs that differed only in a single dimension was near chance in the first block and slowly improved across successive blocks. Perceptual performance rapidly became attuned to the correlation

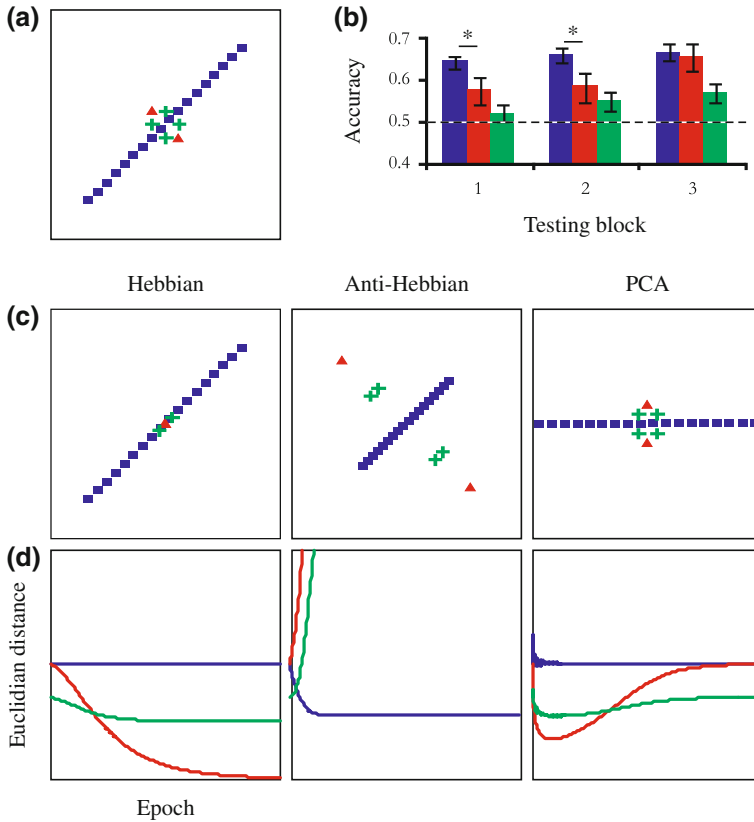


Fig. 7 Stimuli, performance, and modeling results from Stilp et al. (2010c). **a** Listeners discriminated sounds with a strong correlation between AD and SS ($r = \pm 0.97$). Consistent: *blue squares*, Orthogonal: *red triangles*, Single-cue: *green plus signs*. **b** Listener performance. *Dashed line* indicates chance performance, and error bars represent standard error of the mean. Asterisks indicate significant contrasts of interest assessed by paired-sample *t*-tests following Bonferroni correction for multiple comparisons ($p < 0.025$). **c** Model representations of stimuli at the end of simulation. Simulations of the Hebbian network are shown in the *left* column, anti-Hebbian network in the *center* column, and PCA network in the *right* column. Representations are plotted in first-output-unit-by-second-output-unit space. **d** Measures of Euclidean distance between test points throughout the simulation for each model

among stimulus attributes during the very first block of trials, and discrimination of sounds that did not share this covariance was initially impaired. Following successive blocks of trials, performance on Orthogonal test pairs improved to levels comparable to Consistent test pairs.

We employed three simple unsupervised-learning neural network models with similar architectures, but reflecting different hypotheses about how sensorineural systems exploit covariance, to assess how each accounts for listener data. First is a Hebbian model (Hebb 1949; Oja 1982) in which connection weights adjust in

proportion to the correlation between input and output node activations. Second, an anti-Hebbian (decorrelation) model (Barlow and Földiák 1989; Clifford et al. 2007) that orthogonalizes output dimensions by adjusting symmetric inhibition among output nodes proportional to their correlation. Finally, principal component analysis (PCA) was implemented in a third model (Sanger 1989). Connections to output units adjust in a Hebbian manner; however, the first output inhibits inputs to the second, effectively capturing the principal component from the input pattern and leaving the second unit to capture residual covariance. This model captures correlation across inputs (like the Hebbian model) and orthogonalizes outputs (like the anti-Hebbian model). Final solution spaces for each of the three models are shown in Fig. 7c.

As a model analog of perceptual discriminability, Euclidean distances between output activations for each pair of test sounds were computed at each pass through the 18 stimulus pairs. The Hebbian model, owing to the robust correlation in the input, assimilated Orthogonal and Single-cue stimuli to the correlation (Fig. 7d, left). Thus, this model predicts that performance on Orthogonal cues should collapse and never recover. In the anti-Hebbian model, symmetric inhibition between outputs grew in proportion to their pairwise correlation until activity of the output units was uncorrelated. As a result, the Consistent dimension was compressed, and the Orthogonal dimension strongly expanded. Thus, this model predicts that Orthogonal cues should grow more and more discriminable over time (Fig. 7d, middle). The PCA model quickly discovered the first component (the Consistent dimension) so that distances between Orthogonal and Single-cue pairs initially decreased (Fig. 7d, right). With further experience on the same materials, the PCA model gradually captured the modest variance not explained by the first component, progressively increasing distances between Orthogonal pairs to original relative distances. Listener performance violated predictions of the Hebbian and anti-Hebbian models, but matched the PCA model quite well. Continued exposure with the same highly-structured stimulus set quickly eroded, then gradually recaptured, distinctions along the Orthogonal dimension.

Data from this and numerous additional experiments by Stilp and colleagues (Stilp et al. 2010c; Stilp and Kluender 2012) 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 to initially capture the principal component of variation in the two-dimensional stimulus space at the expense of the orthogonal component, and only gradually comes to encode remaining variance. Both this initial component and the second component appear to rapidly become weighted in a way that is proportional to the amount of variance accounted for by each dimension.

The particular PCA model investigated here (Sanger 1989) is certainly oversimplified and is unlikely to exactly reflect neural learning mechanisms. It bears note that, because stimuli were normed to equivalent perceptual distances (JNDs), the perceptual space 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. Small amounts of experience with highly-correlated items provide evidence that stimuli align along a single dimension, so discriminability of differences along orthogonal components is reduced.

Brief experience with correlation between two acoustic attributes may illuminate how extended experience with natural covariance among many attributes contributes to categorical perception. Studies of categorical perception employ highly familiar complex stimuli that vary along multiple dimensions. One criterion of categorical perception—poor within-category discrimination—may arise from efficient coding of covariance structure in a high-dimensional feature space. To the extent that correlations between stimulus attributes are quite strong and there is reduction in dimensionality, one would predict that discrimination of stimulus differences that do not respect those correlations should be relatively poor. To make such a claim, however, requires extending existing efforts to more complex covariance structures. To date, these investigations have been limited to a single covariance structure with no competing correlations available to listeners. Ongoing studies include two separate correlation structures within a three-dimensional perceptual space. Our expectation is that, when separate patterns of redundancies inherent to respective correlations are learned, variability not predicted within either correlation will be enhanced.

12 Second-Order Statistics in Vowels

Given that correlations between a priori independent acoustic attributes can be learned quickly and efficiently, we now turn to speech sounds with which listeners have a lifetime of experience. We will explore ways through which predictability among acoustic attributes (patterns of fundamental and formant frequencies) may reveal important insights into how vowels are perceived. To the extent that this approach is correct, classic concerns about ‘talker normalization’ may be dissolved in a way related to that for ‘rate normalization’.

We begin with some simple facts concerning the relationship between patterns of formants for different vowel sounds across different talkers. Spectra of vowel sounds include peaks (formants) corresponding to resonances in the vocal tract. Center frequencies of these peaks depend upon two physical properties of vocal tracts. First, formant frequencies depend upon the shape of the vocal tract. Vowels vary mostly in how high or low (close, open) and how forward (front, back) the tongue body is in the oral tract. In addition, some vowels are produced with rounded lips (e.g., /u/ as in “boot”) or with different fundamental frequencies among other variations. 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_1 s. 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 /a/, 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 in English can be depicted roughly by relative frequencies of F_1 and F_2 with the exception of /ɜ/.

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 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 phonemically the same, such as /æ/ produced by men, women and children, differ greatly in acoustic properties according to vocal tract length. This variation across talkers is so extreme that clear renditions of any given vowel overlap considerably with different vowels by talkers with vocal tracts of different lengths.

Nearey (1989) carefully reviewed both this challenge for vowel perception and a host of potential solutions to the problem of talker-dependent overlap. Following Ainsworth (1975), Nearey divides solutions into two types: intrinsic and extrinsic. Extrinsic models recommend that perception adjusts or normalizes following development of a frame of reference based upon formant frequencies across a talker's entire vowel system (e.g., Ladefoged and Broadbent 1957; Ladefoged 1967; Gerstman 1968; Nordström and Lindblom 1975; Nearey 1978). Inclusion of extrinsic adjustments can contribute to performance of vowel pattern recognition models (Nearey 1989), but extrinsic models will not be discussed further here.

By contrast, intrinsic models assume that differences between talkers can be more elegantly accommodated if one adopts transformations that reveal underlying commonality. Confusion between vowels across talkers is suggested to ameliorate, if not disappear if some relational measure is adopted across an appropriately transformed vowel space. Intrinsic models have a very long history, extending back to Lloyd (1890a, b, 1891, 1892, cf. 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., Chiba and Kajiyama 1941; Okamura 1966; Minifie 1973; Broad 1976; Kent 1979; Miller 1989).

Much of talker-dependent differences in vowel sounds, or at least those accounted for by vocal-tract length, decrease following two operations. First, formant center frequencies are converted from a linear scale to more psycho-acoustically realistic scales such as logarithmic, Koenig, mel, or Bark. Second, to capture systematicities across talkers, Miller (1989) and Nearey (1989) employed measures of $\log(F_2/F_1)$ and $\log(F_3/F_2)$. When one projects lines corresponding to these ratios across the vowel space, much of the variance between talkers for each vowel is captured. For both Miller and Nearey, ratios that capture this relationship

serve to normalize talker-dependent renditions of vowel sounds in a way that permits matching within a pattern recognition model.

Consistent with efforts related above, we adopt the more contemporary ERB scale (Moore and Glasberg 1983). We employed principal component analysis to measure the amount of shared covariance for ‘steady state’ $ERBF_1$, $ERBF_2$, and $ERBF_3$ for each of the twelve vowels spoken by 139 men, women, and children and reported by Hillenbrand and colleagues (Hillenbrand et al. 1995). These proportions of variance accounted for by covariance between cochleotopic indices (ERB; see Fig. 8) of spectral peaks are provided in the first column of Table 1. The relationship between ERB F_1 , F_2 , and F_3 captures over three fourths of the substantial variability across men, women, and child talkers.

13 Redundancy Enhances Sensitivity to Phonemic Distinctions

Our efficient coding approach embraces these systematicities, but departs from previous efforts in two important ways. First, we do not claim that the perceptual utility of these systematicities across talkers is to normalize toward some iconic ideal or template. Instead, these systematicities describe redundancy inherent to formant relationships for each phonemic vowel across talkers. Second, the real perceptual effect of efficiently coding these redundancies is to increase discriminability of each vowel from any other vowel across talkers. Differences between formant patterns that respect quasi-lawful consequences of vocal tract length are

Fig. 8 Mean $ERBF_1/ERBF_2$ of vowels in the Hillenbrand et al. (1995) database measured at vowel midpoint (50 % of overall duration). Measures are averaged across men (*lower-left* point on each line), women (*center*), and children (*upper-right*)

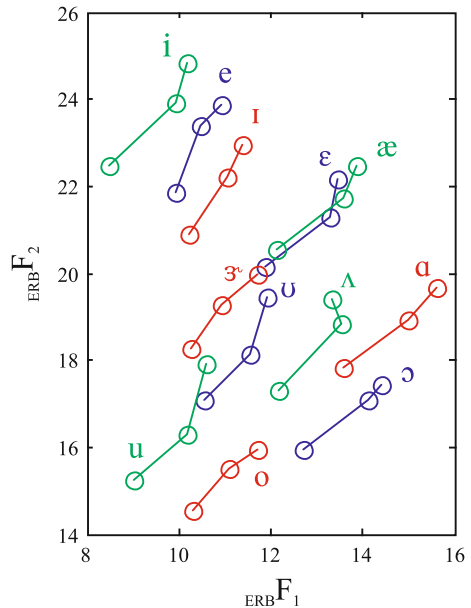


Table 1 Proportions of variance accounted for across talkers (men, women, boys, girls) for American English vowels in the Hillenbrand et al. (1995) database

Vowel	F_1, F_2, F_3	f_0 , mean F_{1-3}	f_0, F_1, F_2, F_3
æ	0.78	0.94	0.79
ɑ	0.76	0.90	0.75
ɔ	0.77	0.90	0.76
ɛ	0.79	0.92	0.79
e	0.76	0.93	0.78
ɜ	0.83	0.90	0.79
ɪ	0.85	0.94	0.83
i	0.82	0.93	0.81
o	0.75	0.86	0.69
ʊ	0.80	0.87	0.74
ʌ	0.80	0.93	0.80
u	0.78	0.85	0.69
Mean	0.79	0.90	0.77

Only vowels with measures of f_0 and all formant frequencies were analyzed. All calculations are done in ERB frequency. Columns list proportion of variance explained by the principal component (eigenvalue of the principal component divided by the sum of all eigenvalues)

predictable and uninformative with respect to decoding the linguistic message. As a consequence, detection of differences that are linguistically significant, for example ‘bet’ versus ‘bat’, is enhanced. Such an approach is, of course, consistent with patterns of perception for vowel sounds. While perception of systematic acoustic changes between vowels may not always be as compellingly categorical as for consonants, it certainly is true that detection of differences between vowel sounds is heightened when stimuli are examples of two different vowels (e.g., /i/ vs. /ɪ/) than examples of the same vowel (e.g., Lotto et al. 1998).

The relationship between formant center frequencies and fundamental frequencies (f_0 s) may be especially telling. Fundamental frequency is a consequence of vocal fold vibration, and rate of vibration depends upon both the mass and stiffness of vocal folds during phonation. While talkers have substantial control over relative stiffness/laxness, vocal fold mass is a different matter. Vocal fold mass, like vocal tract length, is given by anatomy. All things being equal, larger people generally have larger vocal folds, but there are notable exceptions (Nearey 1989). This relationship is complicated by the fact that post-pubescent males have much heftier vocal folds as a secondary sex characteristic. This change in vocal fold mass consequent to puberty results in dramatically lower f_0 beyond expectations for body size alone. 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).

These facts may lead one to expect that the relationship between f_0 and formant center frequencies should be much more tenuous than the relationships between ${}_{\text{ERB}}F_1$, ${}_{\text{ERB}}F_2$, and ${}_{\text{ERB}}F_3$. Relationships between vocal tract resonances are 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, especially given mass-enhancing consequences of male development, $_{\text{ERB}}f_0$ correlates well with mean ERB formant values. Talkers appear to enforce this systematicity despite being under no physiologic obligation to do so. Proportion of variance captured by $_{\text{ERB}}f_0$ with mean ERB formant frequencies for the same vowels (Hillenbrand et al. 1995) are listed in the second column of Table 1, with a mean of 0.90. Finally, proportions of variance accounted for by covariance between cochleotopic indices of F_1 , F_2 , F_3 , and f_0 are presented in the third column, with a mean across vowels of 0.77.

Here, we suggest that significant relationships between ERB f_0 and ERB formant center frequencies suggest that talkers ‘know’ about the correlation between $_{\text{ERB}}F_1$, $_{\text{ERB}}F_2$, and $_{\text{ERB}}F_3$. Further, talkers adjust their f_0 when producing vowels in a way that reinforces the redundancy between $_{\text{ERB}}F_1$, $_{\text{ERB}}F_2$, and $_{\text{ERB}}F_3$ by producing an f_0 that respects this correlation. To the extent that the auditory system seizes upon redundancies, the obligatory relationship between $_{\text{ERB}}F_1$, $_{\text{ERB}}F_2$, and $_{\text{ERB}}F_3$ together with relatively volitional adjustment of $_{\text{ERB}}f_0$, 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.

Although some experiments have failed to find a substantial role of f_0 for identification of vowels in quiet when all other acoustic properties were available to listeners (Katz and Assmann 2001), other studies have suggested that listeners are sensitive to coordination of f_0 and formant frequencies. When Assmann and Nearey (2007) allowed listeners to adjust either f_0 or formant center frequencies (maintained at constant $\log F_{1,2,3}$ intervals), participants made adjustments that produced combinations of f_0 and formant frequencies that match the covariance pattern observed in acoustic measurements of natural vowels. Following this study using listeners’ adjustment of stimulus attributes, Assmann and Nearey (2008) employed a relatively large-scale identification task using 11 different vowel sounds varying in five or six equal-proportion (log-spaced) steps of formant frequency values and three to six log-spaced steps of f_0 . Performance declined systematically when spectral envelopes were shifted upward or downward from formant frequencies that are typical for a given f_0 , suggesting that perceptual processes engaged in vowel perception are sensitive to covariation of f_0 and formant frequencies in natural speech. For both of these studies (Assmann and Nearey 2007, 2008), the authors report changes in performance across all vowels tested, and it would be useful to know the extent to which, if any, listeners are sensitive to vowel-specific relationships between ERB f_0 and ERB formant patterns.

Finally, one recent finding (Katseff et al. 2010) provides further evidence that talkers are implicitly sensitive to relational systematicities among different acoustic attributes of vowel sounds. There is a long history of studying changes in

vocal output when talkers receive acoustically altered auditory feedback. Through the use of clever signal processing methods, some recent studies have investigated if and how listeners adjust their productions when receiving near-simultaneous feedback of their own voice with only F_1 adjusted away from its original peak center frequency (Houde and Jordan 2002; Purcell and Munhall 2006, 2008). The authors of these studies report how talkers adjust their F_1 frequency (higher, lower) in the direction that would compensate for the altered rendition (lower, higher F_1) of their speech.

Katseff and colleagues extended this work by also measuring whether talkers adjusted other formants, specifically F_2 , following perturbed auditory feedback of F_1 , and whether production of F_1 adjusted in response to perturbation of F_2 feedback. They employed CVC words with the vowels / ϵ / and / Λ /. Talkers were very unsystematic in their productions of / Λ / in the presence of altered-formant feedback, but productions of / ϵ / (depicted as a line in Fig. 9) were systematic and readily interpretable. Consistent with previous studies, when auditory feedback displaced F_1 higher than originally spoken (dashed arrow), talkers decreased F_1 in production a compensatory fashion; the same effect was observed for altered F_2 feedback. However, talkers' adjustments of the unaltered formant are especially telling. When F_1 was altered upward in frequency, talkers increased F_2 in frequency, and when F_2 was altered upward, talkers increased the center frequency of F_1 .

If, as we propose, phonemically different vowels are defined by relational properties between formant frequencies (a line in ${}_{\text{ERB}}F_1$ - ${}_{\text{ERB}}F_2$ space), there are two ways to restore this relationship when the center frequency of one formant is altered. First, talkers can and do compensate by changing the frequency of the adjusted formant in the direction opposite the alteration. Second, talkers adjust the center frequency of the nonaltered formant in the same direction as the perturbation (net effect depicted in bold arrow), thus better preserving the desired

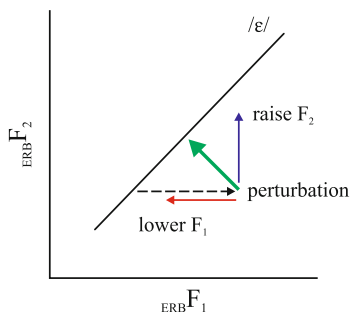


Fig. 9 Schematized results of the altered-feedback experiment by Katseff et al. (2010). As in Fig. 8, / ϵ / is depicted as a line in ${}_{\text{ERB}}F_1$ / ${}_{\text{ERB}}F_2$ space. Listeners hear their own vocalizations of / ϵ / with F_1 shifted upward (dashed line). Listeners compensate for this alteration by simultaneously lowering F_1 and raising F_2 (solid arrows). These compensatory shifts have the net effect of taking the shortest path possible to get back to line depicting / ϵ / (bold arrow)

relational properties across both formants. This is another instance for which one can interpret talker's productions in terms of detecting and reinforcing redundancies in ways that enhance discriminability of more informative differences between phonemically different vowel sounds.

14 Discriminative, Not Generative, Models

The reader will note that, for each of the topics discussed thus far, the emphasis has been upon enhancing detection of change. Our information-theoretic premise is that sensorineural systems optimize sensitivity to change. Discovering redundancy between formant peaks and f_0 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 discriminative, versus generative, models of classification. For example, a generative model of character recognition, such as for reading the address on an envelope, would attempt to capture defining characteristics (*à la* prototype) 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 'a', 'b', 'c', 'e', 'f', etc. across changes in font and size. In the discipline of pattern classification, discriminative models are greatly 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 and 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., Miller and Nicely 1955). Correct responses (diagonal) are correct to the extent that distinctions from other stimuli (off diagonal) are detected.

Contrasts between sounds, not commonalities, are emphasized in speech perception. 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 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) and persists in contemporary contrastive phonology, especially at University of Toronto (e.g., Dresher and Rice 1994). Linguists Roman Jakobson and Morris Halle stated this most starkly in their classic book *Fundamentals of Language* (1971): "All phonemes denote nothing but mere otherness" (p. 22).

15 Learning to Talk

Here, it may be worthwhile to briefly consider implications for young children who are learning to talk. Our claim is that infants learn distinctions between sounds, not consonants and vowels as entities *per se*. Infants can distinguish speech sounds long before they can produce them, as 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. Information transmission is optimized by maximizing sensitivity to differences; this is 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 and Miolo 1995; Vorperian et al. 2005). The infant vocal tract begins more as a single tube not unlike that of a chimpanzee. While this configuration facilitates simultaneous drinking and breathing, it impedes production of many speech sounds. The larynx begins too high with a vocal tract too short, and laryngeal and articulatory structures undergo drastic restructuring across the first 6 years (Vorperian et al. 1999, 2005). What is a neotalker to do?

Mimicking speech sounds of the adult is not an option. Resonances are too high and some vocal-tract configurations (e.g., high back /u/) are physiologically impossible. However, it is possible for the developing vocal tract to produce sounds that are different in ways similar to how adult speech sounds differ. Different vocal tract architectures make it fruitless for young children to try to make a veridical match to articulatory or auditory targets. However, the child is able to preserve acoustic contrasts in speech 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.

16 Conclusions

In this contribution, we first established some first principles that motivate our perspective on speech perception and perception most broadly. We adopted an information-theoretic framework that has a long and productive history in the study of vision and is now more commonly described by the contemporary moniker ‘efficient coding.’ There are two substantial consequences of adopting this information-theoretic framework to questions concerning speech perception. First, distractions concerning objects of perception (gestures versus sounds) are removed. Second, we reframe speech perception as a series of processes through which sensitivity to information—that which changes and/or is unpredictable—becomes increasingly sophisticated and shaped by experience.

We showed how a simple measure of change in the auditory periphery (CSE) proved to be a remarkable predictor of speech intelligibility. Moreover, intelligibility is shown to be critically dependent upon spectral change in vowels and vowel-like speech sounds. Further, we illuminated the ways through which vowel inherent spectral change (VISC) serves to enhance distinctions between vowel sounds, especially those for which relatively static compositions are acoustically or auditorally proximate.

Next, we demonstrated how 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. Providing further evidence that perceptual processes operate in ways that factor out predictability in order to emphasize spectral change or information, we reviewed studies that show how reliable spectral characteristics of a listening context are factored out of perception entirely.

We then considered how listeners may efficiently code reliable covariance between acoustic attributes of speech sounds as they are structured by lawful properties of the vocal tract. We provided evidence that 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. We then re-examined well-known relational characteristics among spectral peaks in vowels as a function of talker differences. In this case, efficient coding of these predictable relationships serves both to dissolve concerns about talker normalization and to enhance distinctiveness between renditions of phonemically different vowel sounds.

We suggest that adopting an efficient coding framework provides a productive way to address questions concerning perception of vowel sounds, perception of speech, and perception most broadly.

Acknowledgments We wish to thank Ray Kent, Peter Assmann, and Catherine Rogers for helpful insights from previous drafts of this chapter. Funding has been provided by NIDCD (first and second authors) and SSHRC (third author).

References

- Aaltonen, O.: The effect of relative amplitude levels of F2 and F3 on the categorization of synthetic vowels. *J. Phon.* **13**, 1–9 (1985)
- Ainsworth, W.A.: Duration as a cue in the recognition of synthetic vowels. *J. Acoust. Soc. Am.* **51**, 648–651 (1972). doi:[10.1121/1.1912889](https://doi.org/10.1121/1.1912889)
- Ainsworth, W.A.: The influence of precursive sequences on the perception of synthesized vowels. *Lang. Speech* **17**, 103–109 (1974). doi:[10.1177/002383097401700201](https://doi.org/10.1177/002383097401700201)
- Ainsworth, W.A.: Intrinsic and extrinsic factors in vowel judgments. In: Fant, G., Tatham, M. (eds.) *Auditory Analysis and Perception of Speech*, pp. 103–113. Academic, London (1975)
- Alexander, J.M., Kluender, K.R.: Spectral tilt change in stop consonant perception. *J. Acoust. Soc. Am.* **123**, 386–396 (2008). doi:[10.1121/1.2817617](https://doi.org/10.1121/1.2817617)

- Alexander, J.M., Kluender, K.R.: Temporal properties of perceptual calibration to local and broad spectral characteristics of a listening context. *J. Acoust. Soc. Am.* **128**(6), 3597–3613 (2010). doi:[10.1121/1.3500693](https://doi.org/10.1121/1.3500693)
- Assmann, P.F., Nearey, T.M.: Relationship between fundamental and formant frequencies in voice preference. *J. Acoust. Soc. Am.* **122**, 35–43 (2007). doi:[10.1121/1.2719045](https://doi.org/10.1121/1.2719045)
- Assmann, P.F., Nearey, T.M.: Identification of frequency-shifted vowels. *J. Acoust. Soc. Am.* **124**, 3203–3212 (2008). doi:[10.1121/1.2980456](https://doi.org/10.1121/1.2980456)
- Assmann, P.F., Summerfield, Q.: The perception of speech under adverse conditions. In: Greenberg, S., Ainsworth, W.A., Popper, A.N., Fay, R.R. (eds.) *Speech Processing in the Auditory System*, vol. 14, pp. 231–308. Springer, New York (2004). doi:[10.1007/b97399](https://doi.org/10.1007/b97399)
- Attneave, F.: Some informational aspects of visual perception. *Psychol. Rev.* **61**, 183–193 (1954). doi:[10.1037/h0054663](https://doi.org/10.1037/h0054663)
- Attneave, F.: *Applications of Information Theory to Psychology: A summary of Basic Concepts, Methods, and Results*. Henry Holt and Company, Inc., New York (1959)
- Backus, B.T.: Perceptual metamers in stereoscopic vision. In: Dieterich, T.G., Becker, S., Ghahramani, Z. (eds.) *Advances in Neural Information Processing Systems*, vol. 14. MIT Press, Cambridge (2002)
- Barlow, H.B.: Sensory mechanisms, the reduction of redundancy, and intelligence. *NPL Symp. Mech. Thought Process.* **10**, 535–539 (1959)
- Barlow, H.B.: Possible principles underlying the transformations of sensory messages. In: Rosenblith, W.A. (ed.) *Sensory Communication*, pp. 53–85. MIT Press, Wiley, Cambridge, New York (1961)
- Barlow, H.B.: The knowledge used in vision and where it comes from. *Philos. Trans. Roy. Soc. Lond. B, Biol. Sci.* **352**(1358), 1141–1147 (1997). doi:[10.1098/rstb.1997.0097](https://doi.org/10.1098/rstb.1997.0097)
- Barlow, H.B.: Redundancy reduction revisited. *Netw. Comput. Neural Syst.* **12**, 241–253 (2001). doi:[10.1080/net.12.3.241.253](https://doi.org/10.1080/net.12.3.241.253)
- Barlow, H.B., Földiák, P.: Adaptation and decorrelation in the cortex. In: Durbin, R., Miall, C., Mitchison, G. (eds.) *The Computing Neuron*, pp. 54–72. Addison-Wesley, New York (1989)
- Bladon, R.A.W., Lindblom, B.: Modeling the judgment of vowel quality differences. *J. Acoust. Soc. Am.* **69**, 1414–1422 (1981). doi:[10.1121/1.385824](https://doi.org/10.1121/1.385824)
- Blumstein, S.E.: The mapping from acoustic structure to the phonetic categories of speech: The invariance problem. *Behav. Brain Sci.* **21**, 260 (1998). doi:[10.1017/S0140525X98221170](https://doi.org/10.1017/S0140525X98221170)
- Broad, D.J.: Toward defining acoustic phonetic equivalence for vowels. *Phonetica* **33**, 401–424 (1976)
- Bunton, K., Story, B.H.: Identification of synthetic vowels based on a time-varying model of the vocal tract area function. *J. Acoust. Soc. Am.* **127**, 146–152 (2010). doi:[10.1121/1.3313921](https://doi.org/10.1121/1.3313921)
- Caclin, A., Brattico, E., Tervaniemi Näätänen, R., Morlet, D., Giard, M.-H., McAdams, S.: Separate neural processing of timbre dimensions in auditory sensory memory. *J. Cogn. Neurosci.* **18**, 1959–1972 (2006). doi:[10.1162/jocn.2006.18.12.1959](https://doi.org/10.1162/jocn.2006.18.12.1959)
- Chechik, G., Anderson, M.J., Bar-Yosef, O., Young, E.D., Tishby, N., Nelken, I.: Reduction of information redundancy in the ascending auditory pathway. *Neuron* **51**, 359–368 (2006). doi:[10.1016/j.neuron.2006.06.030](https://doi.org/10.1016/j.neuron.2006.06.030)
- Chiba, T., Kajiyama, M.: *The Vowel: Its Nature and Structure*. Tokyo Publishing Co., Tokyo (1941)
- Clements, G.N.: Does sonority have a phonetic basis? In: Raimy, E., Cairns, C. (eds.) *Contemporary Views on Architecture and Representations in Phonological Theory*, pp. 165–175. MIT Press, Cambridge (2009)
- Clifford, C.W.G., et al.: Visual adaptation: neural, psychological and computational aspects. *Vision. Res.* **47**, 3125–3131 (2007). doi:[10.1016/j.visres.2007.08.023](https://doi.org/10.1016/j.visres.2007.08.023)
- Cole, R., Yan, Y., Mak, B., Fenty, M., Bailey, T.: The contribution of consonants versus vowels to word recognition in fluent speech. In: *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'96)*, pp. 853–856, Atlanta, GA, (1996)
- Delattre, F.C., Liberman, A.M., Cooper, F.S.: Acoustic loci and transitional cues for consonants. *J. Acoust. Soc. Am.* **27**, 769–773 (1955). doi:[10.1121/1.1908024](https://doi.org/10.1121/1.1908024)

- Diehl, R.L., Kluender, K.R.: On the objects of speech perception. *Ecol. Psychol.* **1**, 121–144 (1989). doi:[10.1207/s15326969eco0102_2](https://doi.org/10.1207/s15326969eco0102_2)
- Dresher, B.E., Rice, K.: Complexity in phonological representations. *Toronto Working Papers in Linguistics*, vol. 12, pp. i–iv (1994)
- Fairhall, A.L., Lewen, G.D., Bialek, W., de Ruyter van Steveninck, R.R.: Efficiency and ambiguity in an adaptive neural code. *Nature* **412**, 787–792 (2001). doi:[10.1038/35090500](https://doi.org/10.1038/35090500)
- Fant, C.G.M.: On the predictability of formant levels and spectrum envelopes from formant frequencies. In: Halle, M. (ed.) *For Roman Jakobson: Essays on the Occasion of His Sixtieth Birthday*, pp. 109–120. Mouton, The Hague (1956)
- Fant, G.: *Acoustic Theory of Speech Production with Calculations Based on X-Ray Studies of Russian Articulations*. Mouton, The Hague (1970)
- Fletcher, H.: *Speech and Hearing in Communication*. Krieger, New York, (1953/1995)
- Fogerty, D., Kewley-Port, D.: Perceptual contributions of the consonant-vowel boundary to sentence intelligibility. *J. Acoust. Soc. Am.* **126**, 847–857 (2009). doi:[10.1121/1.3159302](https://doi.org/10.1121/1.3159302)
- Fowler, C.A.: An event approach to the study of speech perception from a direct-realist perspective. *J. Phon.* **14**, 3–28 (1986)
- Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N.: *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM*. National Institute of Standards and Technology, NTIS Order No. PB91–505065 (1990)
- Geisler, W.S., Perry, J.S., Super, B.J., Gallogly, D.P.: Edge co-occurrence in natural images predicts contour grouping performance. *Vision. Res.* **41**, 711–724 (2001). doi:[10.1016/S0042-6989\(00\)00277-7](https://doi.org/10.1016/S0042-6989(00)00277-7)
- Gerstman, L.: Classification of self-normalized vowels. *IEEE Trans. Audio Electroacoust.* **16**, 78–80 (1968). doi:[10.1109/TAU.1968.1161953](https://doi.org/10.1109/TAU.1968.1161953)
- Glasberg, B.R., Moore, B.C.J.: Deviation of auditory filter shapes from notched-noise data. *Hear. Res.* **47**, 103–138 (1990). doi:[10.1016/0378-5955\(90\)90170-T](https://doi.org/10.1016/0378-5955(90)90170-T)
- Gottfried, T.L., Miller, J.L., Payton, P.E.: Effect of speaking rate on the perception of vowels. *Phonetica* **47**, 155–172 (1990). doi:[10.1159/000261860](https://doi.org/10.1159/000261860)
- Greenwood, D.D.: A cochlear frequency-position function for several species—29 years later. *J. Acoust. Soc. Am.* **87**, 2592–2605 (1990). doi:[10.1121/1.399052](https://doi.org/10.1121/1.399052)
- Hebb, D.O.: *Organization of Behavior*. Wiley, New York (1949)
- Hedrick, M.S., Nábělek, A.K.: Effect of F2 intensity on identification of /u/ in degraded listening conditions. *J. Speech Lang. Hear. Res.* **47**, 1012–1021 (2004). doi:[10.1044/1092-4388\(2004\)075](https://doi.org/10.1044/1092-4388(2004)075)
- Hillenbrand, J.M., Nearey, T.M.: Identification of resynthesized /hVd/ utterances: effects of formant contour. *J. Acoust. Soc. Am.* **105**, 3509–3523 (1999). doi:[10.1121/1.424676](https://doi.org/10.1121/1.424676)
- Hillenbrand, J., Getty, L.A., Clark, M.J., Wheeler, K.: Acoustic characteristics of American English vowels. *J. Acoust. Soc. Am.* **97**, 3099–3111 (1995). doi:[10.1121/1.411872](https://doi.org/10.1121/1.411872)
- Hillis, J.M., Ernst, M.O., Banks, M.S., Landy, M.S.: Combining sensory information: Mandatory fusion within, but not between, senses. *Science* **298**, 1627–1630 (2002). doi:[10.1126/science.1075396](https://doi.org/10.1126/science.1075396)
- Houde, J.F., Jordan, M.I.: Sensorimotor adaptation of speech i: compensation and adaptation. *J. Speech Lang. Hear. Res.* **45**, 295–310 (2002). doi:[10.1044/1092-4388\(2002\)023](https://doi.org/10.1044/1092-4388(2002)023)
- Jakobson, R., Halle, M.: *The Fundamentals of Language*. Mouton, The Hague (1971)
- Jenkins, J.J., Strange, W., Edman, T.R.: Identification of vowels in ‘vowelless’ syllables. *Percept. Psychophys.* **34**, 441–450 (1983). doi:[10.3758/BF03203059](https://doi.org/10.3758/BF03203059)
- Katseff, S., Johnson, K., House, J.: Auditory feedback shifts in one formant cause multi-formant compensation (A). *J. Acoust. Soc. Am.* **127**, 1955 (2010). doi:[10.1121/1.3384960](https://doi.org/10.1121/1.3384960)
- Katz, W.F., Assmann, P.F.: Identification of children’s and adults’ vowels: Intrinsic fundamental frequency, fundamental frequency dynamics, and presence of voicing. *J. Phon.* **29**, 23–51 (2001). doi:[10.1006/jpho.2000.0135](https://doi.org/10.1006/jpho.2000.0135)
- Kent, R.D.: Iso vowel lines for the evaluation of vowel formant structure in speech disorders. *J. Speech Hear. Disord.* **44**, 513–521 (1979)

- Kent, R.D., Miolo, G.: Phonetic abilities in the first year of life. In: Fletcher, P., MacWhinney, B. (eds.) *Handbook of Child Language*, pp. 303–334. Blackwell, London (1995)
- Kewley-Port, D., Burkle, T.Z., Lee, J.H.: Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners. *J. Acoust. Soc. Am.* **122**, 2365–2375 (2007). doi:[10.1121/1.2773986](https://doi.org/10.1121/1.2773986)
- Kiefte, M., Kluender, K.R.: The relative importance of spectral tilt in monophthongs and diphthongs. *J. Acoust. Soc. Am.* **117**, 1395–1404 (2005). doi:[10.1121/1.1861158](https://doi.org/10.1121/1.1861158)
- Kiefte, M., Kluender, K.R.: Absorption of reliable spectral characteristics in auditory perception. *J. Acoust. Soc. Am.* **123**, 366–376 (2008). doi:[10.1121/1.2804951](https://doi.org/10.1121/1.2804951)
- Kiefte, M.: The perception of spectrally and temporally distorted prevocalic stop consonants. unpublished doctoral dissertation, University of Alberta (2000)
- Kiefte, M., Enright, T., Marshall, L.: The role of formant amplitude in the perception of /i/ and /u/. *J. Acoust. Soc. Am.* **127**, 2611–2621 (2010). doi:[10.1121/1.3353124](https://doi.org/10.1121/1.3353124)
- Klatt, D.H.: Prediction of perceived phonetic distance from critical band spectra: a first step. In: *Proceedings of ICASSP*, pp. 1278–1281 (1982)
- Kluender, K.R., Alexander, J.M.: Perception of speech sounds. In: Dallos, P., Oertel, D. (eds.) *The Senses: A Comprehensive Reference*, vol. 3, pp. 829–860. Academic, San Diego (2007)
- Kluender, K.R., Kiefte, M.: Speech perception within a biologically-realistic information-theoretic framework. In: Gernsbacher, M.A., Traxler, M. (eds.) *Handbook of Psycholinguistics*, pp. 153–199. Elsevier, London (2006)
- Kluender, K.R., Lotto, A.J.: Virtues and perils of empiricist approaches to speech perception. *J. Acoust. Soc. Am.* **105**, 503–511 (1999). doi:[10.1121/1.424587](https://doi.org/10.1121/1.424587)
- Kluender, K.R., Diehl, R.L., Killeen, P.R.: Japanese quail can learn phonetic categories. *Science* **237**, 1195–1197 (1987). doi:[10.1126/science.3629235](https://doi.org/10.1126/science.3629235)
- Kluender, K.R., Coady, J.A., Kiefte, M.: Sensitivity to change in perception of speech. *Speech Commun.* **41**(1), 59–69 (2003). doi:[10.1016/S0167-6393\(02\)00093-6](https://doi.org/10.1016/S0167-6393(02)00093-6)
- Ladefoged, P.: *Three Areas of Experimental Phonetics*. Oxford University Press, London (1967)
- Ladefoged, P., Broadbent, D.: Information conveyed by vowels. *J. Acoust. Soc. Am.* **29**, 98–104 (1957). doi:[10.1121/1.1908694](https://doi.org/10.1121/1.1908694)
- Lee, J.H., Kewley-Port, D.: Intelligibility of interrupted sentences at subsegmental levels in young normal-hearing and elderly hearing-impaired listeners. *J. Acoust. Soc. Am.* **125**, 1153–1163 (2009). doi:[10.1121/1.3021304](https://doi.org/10.1121/1.3021304)
- Lieberman, A.M., Mattingly, I.G.: The motor theory of speech perception revised. *Cognition* **21**, 1–36 (1985). doi:[10.1016/0010-0277\(85\)90021-6](https://doi.org/10.1016/0010-0277(85)90021-6)
- Lieberman, A.M., Harris, K.S., Hoffman, H.S., Griffith, B.C.: The discrimination of speech sounds within and across phoneme boundaries. *J. Exp. Psychol.* **54**, 358–368 (1957). doi:[10.1037/h0044417](https://doi.org/10.1037/h0044417)
- Liljencrants, J., Lindblom, B.: Numerical simulation of vowel quality systems: the role of perceptual contrast. *Language* **48**(4), 839–862 (1972). doi:[10.2307/411991](https://doi.org/10.2307/411991)
- Lindblom, B.: Phonetic universals in vowel systems. In: Ohala, J.J., Jaeger, J.J. (eds.) *Experimental Phonology*, pp. 13–44. Academic, Orlando (1986)
- Lindholm, J.M., Dorman, M., Taylor, B.E., Hannley, M.T.: Stimulus factors influencing the identification of voiced stop consonants by normal-hearing and hearing impaired adults. *J. Acoust. Soc. Am.* **83**, 1608–1614 (1988). doi:[10.1121/1.395915](https://doi.org/10.1121/1.395915)
- Lisker, L.: Rapid versus ravid: a catalogue of acoustical features that may cue the distinction. *Haskins Laboratories Status Report on Speech Research*, SR-54, pp. 127–132 (1978)
- Lloyd, R.J.: *Some Researches into the Nature of the Vowel-Sound*. Turner and Dunnett, Liverpool (1890a)
- Lloyd, R.J.: Speech sounds: their nature and causation (II-IV). *Phonetische Studien* **4**, 37–67, 183–214, 275–306 (1891)
- Lloyd, R.J.: Speech sounds: their nature and causation (V-VII). *Phonetische Studien* **5**, 1–32, 129–141, 263–271 (1892b)
- Lloyd, R.J.: Speech sounds: their nature and causation (I). *Phonetische Studien* **3**, 251–278 (1890b)

- Lotto, A.J., Kluender, K.R., Holt, L.L.: Depolarizing the perceptual magnet effect. *J. Acoust. Soc. Am.* **103**, 3648–3655 (1998). doi:[10.1121/1.423087](https://doi.org/10.1121/1.423087)
- Miller, J.L.: Effects of speaking rate on segmental distinctions. In: Eimas, P.D., Miller, J.L. (eds.) *Perspectives on the Study of Speech*, pp. 39–74. Erlbaum Associates, New Jersey (1981)
- Miller, J.D.: Auditory-perceptual interpretation of the vowel. *J. Acoust. Soc. Am.* **85**, 2114–2134 (1989). doi:[10.1121/1.397862](https://doi.org/10.1121/1.397862)
- Miller, J.L., Dexter, E.R.: Effects of speaking rate and lexical status on phonetic perception. *J. Exp. Psychol. Hum. Percept. Perform.* **14**, 369–378 (1988). doi:[10.1037/0096-1523.14.3.369](https://doi.org/10.1037/0096-1523.14.3.369)
- Miller, J.L., Liberman, A.M.: Some effects of later-occurring information on the perception of stop-consonant and semivowel. *Percept. Psychophys.* **25**, 457–465 (1979). doi:[10.3758/BF03213823](https://doi.org/10.3758/BF03213823)
- Miller, G.A., Nicely, P.E.: An analysis of perceptual confusions among some English consonants. *J. Acoust. Soc. Am.* **27**, 338–352 (1955). doi:[10.1121/1.1907526](https://doi.org/10.1121/1.1907526). [Erratum: (1955) 27, 339. doi:[10.1121/1.1907983](https://doi.org/10.1121/1.1907983)]
- Minifie, F.D.: Speech acoustics. In: Minifie, F.D., Hixon, T.J., Williams, F. (eds.) *Normal Aspects of Speech, Hearing, and Language*, pp. 235–284. Prentice-Hall, Englewood Cliffs (1973)
- Moore, B.C.J., Glasberg, B.R.: Suggested formulas for calculating auditory-filter bandwidths and excitation patterns. *J. Acoust. Soc. Am.* **74**, 750–753 (1983). doi:[10.1121/1.389861](https://doi.org/10.1121/1.389861)
- Nearey, T.M.: *Phonetic Feature Systems for Vowels*. Indiana University Linguistics Club, Bloomington (1978)
- Nearey, T.M.: Static, dynamic, and relational properties in vowel perception. *J. Acoust. Soc. Am.* **85**, 2088–2113 (1989). doi:[10.1121/1.397861](https://doi.org/10.1121/1.397861)
- Nearey, T.M.: Speech perception as pattern recognition. *J. Acoust. Soc. Am.* **101**, 3241–3254 (1997). doi:[10.1121/1.418290](https://doi.org/10.1121/1.418290)
- Nearey, T.M., Assmann, P.: Modeling the role of inherent spectral change in vowel identification. *J. Acoust. Soc. Am.* **80**, 1297–1308 (1986). doi:[10.1121/1.394433](https://doi.org/10.1121/1.394433)
- Nearey, T.M.: Vowel inherent spectral change in the vowels of North American English. In: Morrison, G.S., Assmann, P.F. (Eds.) *Vowel Inherent Spectral Change* (ch. 4). Springer, Heidelberg (2012)
- Ng, A.Y., Jordan, M.I.: On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes. In: Dietterich, T., Becker, S., Ghahramani, Z. (Eds.) *Advances in Neural Information Processing (NIPS)*, vol. 14, MIT Press, Cambridge (2002)
- Nilsson, M., Soli, S., Sullivan, J.: Development of the Hearing In Noise Test for the measurement of speech reception thresholds in quiet and in noise. *J. Acoust. Soc. Am.* **95**, 1085–1099 (1994). doi:[10.1121/1.408469](https://doi.org/10.1121/1.408469)
- Nordström, P.-E., Lindblom, B.: A normalization procedure for vowel formant data. In: *Proceedings of the 7th International Congress of Phonetic Sciences*, Leeds, England (1975)
- Nordström, P.-E.: Attempts to simulate female and infant vocal tracts from male area functions. *Speech Transmission Laboratory Quarterly Progress and Status Report (KTH, Stockholm)*, pp. 2–3, 20–33, (1975)
- Ohala, J.J.: There is no interface between phonology and phonetics: a personal view. *J. Phon.* **18**, 153–171 (1990)
- Oja, E.: A simplified neuron model as a principal component analyzer. *J. Math. Biol.* **15**, 267–273 (1982). doi:[10.1007/BF00275687](https://doi.org/10.1007/BF00275687)
- Okamura, M.: Shouni boin no nenrei teki henka ni kansuru kenkyuu: Sound Spectrograph niyuru formant kouzou to boin no bunka no kentou [Acoustical studies of Japanese vowels in children: The formant constructions and the developmental process]. *Nippon Jibiinkoka Gakkai Kaiho [Japan. J. Otolaryngol.]* **69**, 1198–1214 (1966). doi:[10.3950/jibiinkoka.69.6_1198](https://doi.org/10.3950/jibiinkoka.69.6_1198)
- Owren, M.J., Cardillo, G.C.: The relative roles of vowels and consonants in discriminating talker identity versus word meaning. *J. Acoust. Soc. Am.* **119**, 1727–1739 (2006). doi:[10.1121/1.2161431](https://doi.org/10.1121/1.2161431)

- Patterson, R.D., Nimmo-Smith, I., Weber, D.L., Milroy, R.: The deterioration of hearing with age: Frequency selectivity, the critical ratio, the audiogram, and speech threshold. *J. Acoust. Soc. Am.* **72**, 1788–1803 (1982). doi:[10.1121/1.388652](https://doi.org/10.1121/1.388652)
- Purcell, D.W., Munhall, K.G.: Weighting of auditory feedback across the English vowel space. In: *Proceedings of the 8th International Seminar on Speech Production* (2008)
- Purcell, D.W., Munhall, K.G.: Adaptive control of vowel formant frequency: evidence from real-time formant manipulation. *J. Acoust. Soc. Am.* **120**, 966–977 (2006). doi:[10.1121/1.2217714](https://doi.org/10.1121/1.2217714)
- Sanger, T.D.: Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural Netw.* **2**, 459–473 (1989). doi:[10.1016/0893-6080\(89\)90044-0](https://doi.org/10.1016/0893-6080(89)90044-0)
- Saberi, K., Perrott, D.R.: Cognitive restoration of reversed speech. *Nature* **398**, 760 (1999). doi:[10.1038/19652](https://doi.org/10.1038/19652)
- Schwartz, O., Simoncelli, E.P.: Natural signal statistics and sensory gain control. *Nat. Neurosci.* **4**, 819–825 (2001). doi:[10.1038/90526](https://doi.org/10.1038/90526)
- Shannon, C.E.: A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948)
- Simoncelli, E.P.: Vision and the statistics of the visual environment. *Curr. Opinions Neurobiol.* **13**, 144–149 (2003). doi:[10.1016/S0959-4388\(03\)00047-3](https://doi.org/10.1016/S0959-4388(03)00047-3)
- Simoncelli, E.P., Olshausen, B.A.: Natural image statistics and neural representation. *Annu. Rev. Neurosci.* **24**, 1193–1215 (2001). doi:[10.1146/annurev.neuro.24.1.1193](https://doi.org/10.1146/annurev.neuro.24.1.1193)
- Stevens, K.N.: *Acoustic Phonetics*. MIT, Cambridge (1998)
- Stevens, K.N., Blumstein, S.E.: The search for invariant acoustic correlates of phonetic features. In: Eimas, P.D., Miller, J.L. (eds.) *Perspectives in the Study of Speech*. Erlbaum, Hillsdale (1981)
- Stilp, C.E., Kluender, K.R.: Cochlea-scaled spectral entropy, not consonants, vowels, or time, best predicts speech intelligibility. *Proc. Natl. Acad. Sci.* **107**(27), 12387–12392 (2010). doi:[10.1073/pnas.0913625107](https://doi.org/10.1073/pnas.0913625107)
- Stilp, C.E., Kluender, K.R.: Efficient coding and statistically optimal weighting of covariance among acoustic attributes in novel sounds. *PLoS ONE* **7**(1), e30845 (2012). doi:[10.1371/journal.pone.0030845](https://doi.org/10.1371/journal.pone.0030845)
- Stilp, C.E., Alexander, J.M., Kiefte, M., Kluender, K.R.: Auditory color constancy: calibration to reliable spectral properties across nonspeech context and targets. *Atten. Percept. Psychophys.* **72**, 470–480 (2010a). doi:[10.3758/APP.72.2.470](https://doi.org/10.3758/APP.72.2.470)
- Stilp, C.E., Kiefte, M., Alexander, J.M., Kluender, K.R.: Cochlea-scaled spectral entropy predicts rate-invariant intelligibility of temporally distorted sentences. *J. Acoust. Soc. Am.* **128**, 2112–2126 (2010b). doi:[10.1121/1.3483719](https://doi.org/10.1121/1.3483719)
- Stilp, C.E., Rogers, T.T., Kluender, K.R.: Rapid efficient coding of correlated complex auditory properties. *Proc. Natl. Acad. Sci.* **107**(50), 21914–21919 (2010c). doi:[10.1073/pnas.1009020107](https://doi.org/10.1073/pnas.1009020107)
- Story, B.H., Bunton, K.: Simulation and identification of vowels based on a time-varying model of the vocal tract area function. In: Morrison G.S., Assmann P.F. (Eds.) *Vowel Inherent Spectral Change* (ch. 7), Springer, Heidelberg (2012)
- Sussman, H.M., McCaffrey, H.A., Matthews, S.A.: An investigation of locus equations as a source of relational invariance for stop place categorization. *J. Acoust. Soc. Am.* **90**, 1309–1325 (1991). doi:[10.1121/1.401923](https://doi.org/10.1121/1.401923)
- Sussman, H.M., Fruchter, D., Hilbert, J., Sirosh, J.: Linear correlates in the speech signal: the orderly output constraint. *Behav. Brain Sci.* **21**(2), 241–259 (1998). doi:[10.1017/S0140525X98001174](https://doi.org/10.1017/S0140525X98001174)
- Trubetzkoy, N.S.: *Principles of Phonology* (C. Baltaxe, Translator) University of California Press, Berkeley. (Original work published in 1939) (1969)
- Vapnik, V.N.: *Statistical Learning Theory*. Wiley, New York (1998)
- Vorperian, H.K., Kent, R.D., Lindstrom, M.J., Kalina, C.M., Gentry, L.R., Yandell, B.S.: Development of vocal tract length during early childhood: a magnetic resonance imaging study. *J. Acoust. Soc. Am.* **117**, 338–350 (2005). doi:[10.1121/1.1835958](https://doi.org/10.1121/1.1835958)

- Vorperian, H.K., Kent, R.D., Gentry, L.R., Yandell, B.S.: Magnetic resonance imaging procedures to study the concurrent anatomic development of vocal tract structures: preliminary results. *Int. J. Pediatr. Otorhinolaryngol.* **49**, 197–206 (1999). doi:[10.1016/S0165-5876\(99\)00208-6](https://doi.org/10.1016/S0165-5876(99)00208-6)
- Watkins, A.J.: Central, auditory mechanisms of perceptual compensation for spectral-envelope distortion. *J. Acoust. Soc. Am.* **90**, 2942–2955 (1991). doi:[10.1121/1.401769](https://doi.org/10.1121/1.401769)
- Watkins, A.J., Makin, S.J.: Perceptual compensation for speaker differences and for spectral-envelope distortion. *J. Acoust. Soc. Am.* **96**, 1263–1282 (1994). doi:[10.1121/1.410275](https://doi.org/10.1121/1.410275)
- Weiner, N.: *Cybernetics*. Wiley, New York (1948)