1.59 Perception of Speech Sounds

K R Kluender and J M Alexander, University of Wisconsin-Madison, Madison, WI, USA

Glossary

adaptation The decrease in the firing rate of a nerve fiber in response to constant stimulation. Sometimes, adaptation is referred to as fatigue which is really a misnomer if one conceptualizes the process as a simple mechanism by which sensitivity to change across a population of nerve fibers is enhanced.

affricate consonants Speech sounds produced by a complete constriction of the vocal tract that is released into a very small opening such as that for a fricative.

allophonic variation Acoustic or phonetic variants of the same phoneme; discriminably different speech sounds that are identified as the same phoneme.

alveolar stops Stop consonants /t/ and /d/ which are produced by forming a complete constriction of the vocal tract at the alveolar ridge of the hard palate.

categorical perception The phenomenon in which stimuli that are made to vary continuously along one or more dimensions are perceived

discretely, rather than continuously, as distinct classes. The hallmark of categorical perception is the finding that in discrimination tasks, two stimuli that have been given the same label in an identification task are discriminated at levels near chance, while two equidistant stimuli that have been given different labels are discriminated with almost perfect accuracy.

co-articulation Articulatory overlap of adjacent speech sounds which results in a blending of acoustic features across successive speech units.

color constancy The phenomenon that the perceived color of an object stays the same despite the fact that the actual wavelength of reflected light changes with changes in the ambient lighting.

contrast effects Refers to the perceptual enhancement of a stimulus attribute when paired in time and/or space with a stimulus attribute that is to a varying extent opposite or contrastive.

dental stops Stop consonants produced by forming a complete constriction of the vocal tract behind the teeth (like the fricative th in English) as in Hindi [d̪].

dynamic range The physical range over which the biological sensors operate effectively. The difference between the stimulus levels near threshold and those where the physiological response begins to saturate.

enhancement effect A decrease in threshold, or enhanced sensitivity, for a tone in a complex when preceded by a complex in which the tone is missing.

entropy Unpredictability, randomness; entropy is directly proportional to potential information and inversely proportional to redundancy (see Shannon information theory).

formants Peaks in the speech spectrum. Frequency locations of the formants, especially the first three, are important for identifying different speech sounds and depend on the vocal tract resonances, which vary with vocal tract length.

formant transitions Monotonically increasing or decreasing changes in the frequency locations of the formants, or vocal tract resonances, associated with movements of the articulators from one speech sound to another. Some important features of formant transitions that aid in the identification of speech sounds include their onset and offset frequencies and their rate of change.

fricative consonants Speech sounds characterized by a nearly complete constriction of the vocal tract that produces noise as turbulence is created by air passing through the small opening.

fundamental frequency The lowest frequency of a periodic signal. Distance in frequency between successive harmonics in voiced speech sounds corresponding to the rate of vocal fold vibration. The harmonic spectrum consists of the fundamental frequency (first harmonic) with higher harmonics at multiples of the fundamental.

information-theoretic See Shannon information theory.

inverse projection problem The fact that it is impossible to project a given physical stimulus (e.g., auditory or visual) back to a determinate source because multiple sources can produce the exact same physical output.

lack of invariance (problem of variability) Refers to the fact that no one acoustic or articulatory feature is both necessary and sufficient to identify a given speech sound. The lack of a simple one-to-one correspondence between attributes of speech sounds and linguistic units such as phonemes is a problem only when classifying speech sounds, but not when perceiving them.

lateral inhibition When high neural activity at one frequency region causes a reduction in the neural activity at an adjacent frequency region. In this way, the edge between regions of contrast (high and low activity) is enhanced.

lexicon The morphemes and words comprising an individual's vocabulary.

liquid Sound category for /l/ and /r/ sounds, which are produced with vocal tract constrictions that are less than those for fricative and affricates but more than those for vowels.

manner of articulation A way of classifying consonant sounds that describes how the constrictions in the vocal tract are made when they are produced.

McGurk effect Phenomenon in which the auditory perception of a syllable is influenced by the simultaneous visual presentation of a conflicting syllable. This effect is dependent on auditory stimuli that are easily confused and visual stimuli (visemes) that are easily distinguished.

morpheme The smallest unit of language that carries meaning.

nasal consonants Speech sounds characterized by a complete constriction of the oral tract accompanied with a coupling of the nasal tract through which air flow is directed.

necker cubes An optical illusion in which a two-dimensional cube is drawn in a way that causes ambiguity when it is perceived as a three-dimensional object. Two faces of the cube appear to be in front, but not at the same time. One's perception oscillates between the two equally plausible interpretations.

perceptual constancy Functional equivalence of discriminably different physical representations of the same object.

perceptrons Single-layer artificial neural network in which multiple weighted inputs are fed into a single binary output.

phonemes Abstract linguistic categories for speech sounds that are functionally equivalent, but not necessarily acoustically equivalent.

phonetic categories See phonemes.

phonetic segments Divisions of the acoustic speech signal corresponding to consonants and vowels.

phonotactic regularities Sequences of speech sounds that tend to co-occur in a language.

pink noise Noise with an energy distribution that decreases at a rate of −3 dB per octave as frequency increases, or a halving of energy with a doubling of frequency.

place of articulation A way of classifying consonant sounds that describes where in the vocal tract the constriction is made to produce them.

retroflex stops Stop consonants produced by forming a complete constriction of the vocal tract at the hard palate (like the liquid r in English) as in Hindi [d].

Shannon information theory Claude Shannon's theory of communication in which information is represented as a mathematical entity that is agnostic with respect to meaning. The potential amount of information transmitted, or entropy, is related to the number of logarithmic units (e.g., bits) needed to uniquely code the communicated message. Fewer bits are needed when the context and a receiver's prior experience make certain elements of the message redundant.

spectral contrast Contrast effect (see definition) for the frequency composition of a sound. For example, a sound with a mid-frequency peak will be perceived as relatively higher in frequency when preceded by a sound with a lower-frequency peak and will be perceived as relatively lower in frequency when preceded by a sound with a higher-frequency peak.

spectral density The distribution of energy as a function of frequency.

stop consonants Speech sounds characterized by a complete constriction of the vocal tract that is held momentarily and then released in a more or less abrupt manner.

suppression The decrease in the physiological response to a particular frequency region when another, more intense, sound at a nearby frequency region is present.

trading relations Because humans have abundant experience hearing speech sounds that covary along multiple attributes, a change in one attribute toward one percept can be traded or offset by a change in another attribute toward the opposite percept.

voiced Speech sounds created with the vocal folds together and in vibration.

voiceless Speech sounds created with the vocal folds apart, hence not in vibration.

1.59.1 Introduction

During the second half of the twentieth century, research concerning speech perception stood relatively distinct from the study of audition and other modalities of high-level perception such as vision. Contemporary research, however, is beginning to bridge this traditional divide. Fundamental principles that govern all perception, some known for more than a century, are shaping our understanding of perception of speech as well as other familiar sounds.

The study of speech perception traditionally consisted of attempting to explain how listeners perceive the spoken acoustic signal as a sequence of consonants and vowels, collectively referred to as phonetic segments or units. When one describes speech sounds in this way, brackets are used to surround phonetic symbols such as [y] (as in yes) and [o] (as in oh). By contrast, phonemes are abstract linguistic units that roughly correspond to letters in written language, and are transcribed enclosed by slashes (/y/ and /o/.) Morphemes are the smallest meaningful units of language, roughly corresponding to words (e.g., dog,

taste, as well as dis- and -ful) with phonemes being the smallest units that can change the meaning of a morpheme (e.g., /yo/ vs. /go/) (Trubetzkoy, N. S., 1939/1969).

Conceptualizing speech perception as a process by which phonemes are retrieved from acoustic signals is the traditional approach. Within this tradition, research in speech perception often has been focused on problems concerning segmentation and lack of invariance. The problem of segmentation refers to the fact that, if phonetic units exist, they are not like typed letters on a page. Instead, they overlap extensively in time, much like cursive handwriting. The problem of lack of invariance (or, problem of variability) is related to the segmentation problem. Because speech sounds are produced such that articulations for one consonant or vowel overlaps with production of preceding ones, and vice versa, every consonant and vowel produced in fluent connected speech is dramatically colored by its neighbors. Some of the most recalcitrant problems in the study of speech perception are the consequence of adopting discrete phonetic segments as a level of perceptual analysis. However, phonetic segments may be neither discrete nor real.

No experimental evidence clearly demonstrates that either phonetic segments or phonemes are real outside of linguistic theory (e.g., Lotto, A. J., 2000), and the intuitive appeal of phonetic segments and phonemes may arise principally from experience with alphabetic writing systems (Port, R. F., in press). One ought not be sanguine about whether speech perception really is about recognizing consonants and vowels per se. It is not known whether or not listeners extract phonemes preliminary to recognizing words. There may or may not be some place in the brain where phonemes reside independent of the words they comprise.

Either morphemes or words may be the first units of language that stand more or less on their own accord. It is possible, even likely, that speech perception is a series of nondiscrete processes along the way from waveforms to words. In this chapter, speech perception will be described as a continuum of processes operating on the acoustic signal with varying levels of sophistication. The consistent theme will be common principles that define how these processes work.

Before explaining perception of speech, two preliminary topics need to be addressed. First, understanding perception always requires understanding the ecology within which perceptual systems operate. If one wishes to understand speech perception, at least some knowledge of speech production proves helpful. Second, because perception of speech adheres to the same principles that govern perception of other environmental objects and events, it will be helpful to briefly review some broad principles that govern perception most broadly.

1.59.2 Sounds Created with Ears in Mind

Speech and music are somewhat distinct from most other environmental sounds because, much like visual art such as paintings and sculpture, speech and music are created with perceivers in mind. Across hundreds of generations of humans, language users have found ways to use their lungs, larynx, and mouths to produce different sounds that can transmit a wealth of information to one another.

1.59.2.1 Speech Production

There are three basic components to production of speech: respiratory (lungs), phonatory (vocal chords), and articulatory (vocal tract). First is the respiratory system through which the diaphragm pushes air out of the lungs, through the trachea, and then the larynx. At the larynx, air must pass through the two vocal folds which are made up of muscle tissue that can be adjusted to vary how freely air passes through the opening between them. In some cases, such as voiceless sounds like [p] and [s], vocal folds are apart; they do not restrict airflow and do not vibrate. For voiced speech sounds like [b] and [z], vocal folds are closer together, and the pressure of airflow causes them to vibrate at the fundamental frequency (f_0). Owing to variations in the pressure of airflow from the lungs and muscular control of vocal fold tension, talkers can vary the fundamental frequency of voiced sounds. If one were to consider the nature of speech at this point, it could be depicted as a spectrum of energy spread across frequency and concentrated at the fundamental frequency and at multiples of the fundamental (i.e., a harmonic spectrum) with decreasing energy at each successive multiple as seen in Figure 1(a).

The area above the larynx, the oral tract and the nasal tract combined, is referred to as the vocal tract. The nearly continuously varying configuration of the vocal tract is responsible for shaping the spectrum

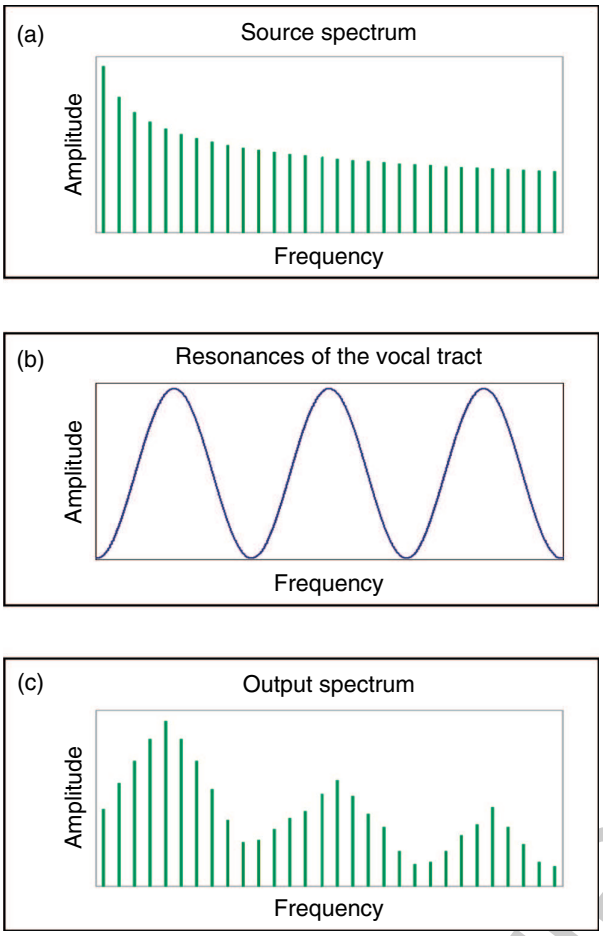

Figure 1 Harmonic spectrum of the laryngeal source (a) before passing through filtering properties (b) of the vocal tract resulting in the output spectrum with formant peaks (c).

differently for different speech sounds. There are many ways the jaw, lips, tongue body, tongue tip, velum (soft palate), and other vocal tract structures can be manipulated to shape the sounds that emanate from the mouth and nose. Widening and narrowing of points along the vocal tract selectively attenuate some frequencies while making others more prominent. Figure 1(b) illustrates the filtering effects of the vocal tract for the vowel [a] as in father. Figure 1(c) portrays the net result from passing the glottal source (a) through the vocal tract (b).

Peaks in the resultant spectrum are referred to as formants, described by number, lowest to highest (F_1, F_2, F_3,. . .). Only the first three formants are depicted in Figure 1, and for the most part, speech sounds can be identified on the basis of energy in the region of these lowest three formants. However, additional formants exist with lower amplitudes at higher frequencies (F_4, F_5, F_6, etc.), and are relatively more prominent in the speech of children.

Airflow can be channeled, constricted, or obstructed to produce different vowel and consonant sounds. Vowels are made with a relatively open unoccluded vocal tract. In terms of articulation, vowels vary mostly in how high or low and how forward (front) or 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 modestly different fundamental frequencies among other variations.

In part because early linguists had greater access to their own vocal tracts than to sophisticated audio analysis, it is a general convention that speech sounds are described in terms of the articulations necessary to produce them. In addition to variation in tongue height, frontness/backness, and lip rounding as descriptions for vowels, consonants also are described by articulatory characteristics. For example, consonants are described in terms of the manner in which constrictions are introduced along the vocal tract. Stop consonants or plosives such as [b], [p], [d], [t], [g], and [k] include complete constriction such that no air may pass through. Nasal consonants such as [n], [m], and [ŋ] (as in sing) are like [b], [d], and [g], respectively, with complete constriction at some point in the oral tract, but air is allowed to escape through the nasal tract because the velum is lowered. Fricative consonants are caused by nearly complete obstruction of the vocal tract, with a noisy sound being produced by turbulence of airflow passing through a very small opening. Some examples of English fricatives are [s], [z], [š] (as in ash), and [ž] (as in azure.) Affricate consonants are produced by a combination of complete occlusion (like a stop) followed by nearly complete occlusion (like a fricative.) Examples of affricates in English are [č] (as in chug) and [j] (as in ǰug.) The least constricted consonants in English are laterals and semivowels, such as [l], [w], [r], and [y].

Consonants also are described in terms of whether the vocal folds are close and vibrating, voiced, or further apart and not vibrating, voiceless. Thus, sounds such as [b], [d], [g], [z], [ž], [j], [l], [w], [r], and [y] are voiced. And, [p], [t], [k], [s], [š], and [č] are voiceless. Finally, consonants are described on the basis of place of articulation. Constrictions can be placed along a number of places in the oral tract. In English, the three major places of articulation are bilabial (lips, [p], [b], [m]), alveolar (alveolar ridge behind teeth, [t], [d], [n]), and velar (soft palate or velum, [k], [g], [ŋ].)

1.59.2.2 Linguistic Sound Systems

The above description does not exhaust all the distinctions among the 40 or so sounds used in English, and it is a vast underdescription of variation among languages more generally. Owing primarily to unique characteristics of supralaryngeal anatomy (Lieberman, P., 1984), the adult human possesses sound-producing abilities unrivaled among other organisms. This capacity is revealed in a grand assortment of over 850 different speech sounds used contrastively by the more than 5000 distinct languages used around the world (Maddieson, I., 1984). There are more than 550 consonants and 300 vowels (including diphthongs such as [e^y] and [o^y], as in bay and boy). Such capacity dwarfs that of other animals, being more than an order of magnitude larger than the largest reported inventory of nonhuman primate calls (Cleveland, J. and Snowdon, C. T., 1982).

In contrast to this diversity in potential speech sounds, systematic inspection reveals that collections of consonants and vowels used by individual languages are anything but random. The vast majority of speech sounds are relatively rare, while a handful are extremely common. For example, all known languages have stop consonants. The overwhelming majority of languages have three places of articulation for stop consonants, typically the three described above for English. Over 80% of languages include a distinction in voicing (e.g., [p] vs. [b], [s] vs. [z]). Diversity does not imply randomness.

The structure of vowel systems is as much or more orderly than that for consonants. There is a fair amount of variety in the particular number of vowels used by languages. Some languages use as few as three vowels while others use as many as 24. English uses about 15 depending upon dialect. However, the most common number of vowels used by languages is only five, and other numbers of vowels appear to be relatively favored. Especially for the five- to nine-vowel systems that predominate across languages, particular sets of vowels are typically used. Figure 2 displays some of the more common three, five, and seven vowel systems. Although there are relatively fewer languages that use more than seven vowels, there remains a good deal of commonality among systems with the same number of vowels.

What are the forces acting upon languages that encourage the selection of some sounds and groups of sounds over others? Although the number of possible speech sounds is prodigious, one guiding factor explaining regularities is how easy some sounds are to produce either in isolation or in sequence with others. The role of articulatory ease is perhaps best evidenced by the fact that languages tend to use articulatorily simpler consonants before incorporating more complex consonants (Lindblom, B. and Maddieson, I., 1988).

Consistent with the close tethering of speech production and speech perception, languages have come to use sets of speech sounds that enhance perceptual effectiveness. Talkers expend effort for communicative robustness. For example, the tense high vowels [i] and [u] (as in beet and boot) require more effort to produce relative to their lax counterparts [I] and [U] (as in bit and book). The vowels [i] and [u], however, are acoustically more distinct, not only from each other, but also from other possible vowel sounds. And, across languages, these tense vowels [i] and [u] occur five times more frequently than lax vowels [I] and [U].

Common examples of talkers molding their utterances to the needs of the listener include instances in which conditions for communication are not optimal. Talkers speak more clearly to young or non-native listeners for whom distinctions are not obvious. When environments are noisy or reverberant, talkers strive to produce contrasts that are maximally distinctive. Most generally, speech sound repertoires of languages have developed over generations of individuals toward greater communicative effectiveness. Generations of talkers have come to capitalize upon auditory predispositions of listeners.

An obvious way that a language community achieves such robustness is by developing an inventory of phonemes so as to optimize distinctiveness acoustically and auditorily. Inspection of regularities in vowel sounds used by languages, such as those shown in Figure 2 provides some of the most illuminating examples of auditory processes operating as a driving force. Different languages use different sets of vowel sounds, and languages use subsets of vowels that are most easily discriminated from one another. In particular, those vowels favored for languages with five vowels are vowels that are as acoustically distant 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 (Bladon, R. A. W. and Lindblom, B., 1981; Liljencrantz, J. and Lindblom, B., 1972).

There is another way that differences between speech sounds are perceptually more dependable. Every consonant and vowel is defined by multiple acoustic properties. For example, the distinction between [b] and [p] as in rabid and rapid includes

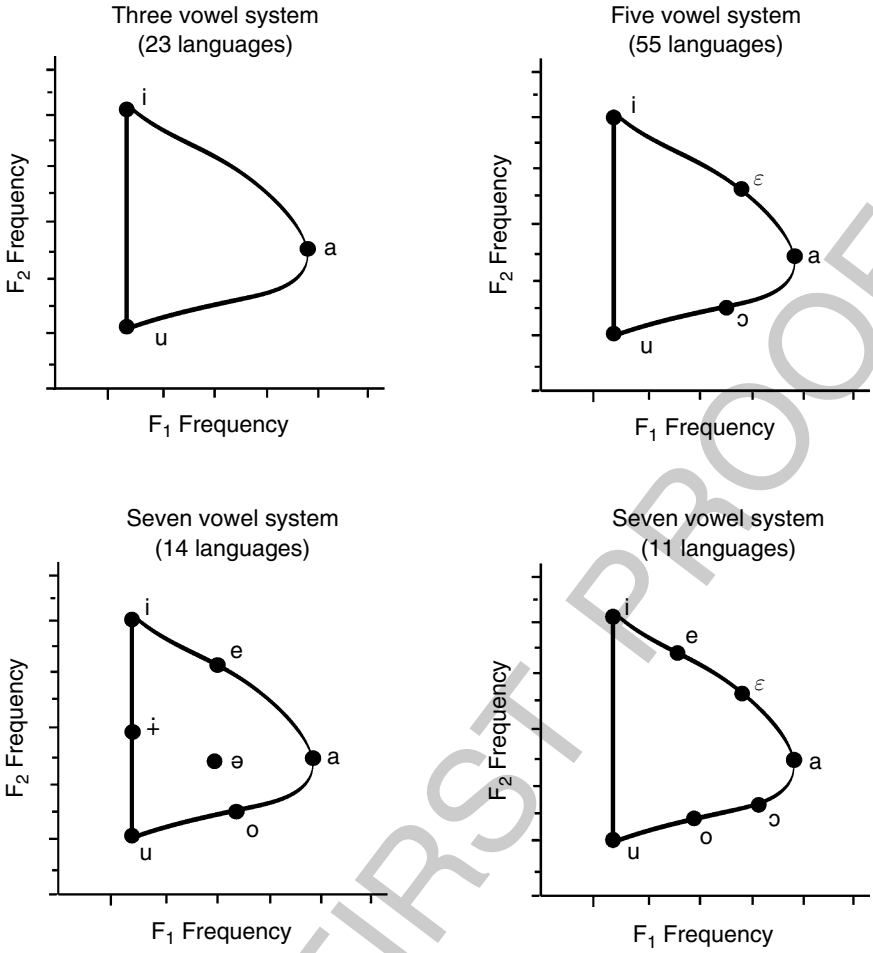

Figure 2 Common three-, five-, and seven-vowel systems used around the world as depicted in frequencies of first and second formants. Note that languages tend to use sets of vowels that are as acoustically distinct as possible within the range of vocal tract limits (solid lines.)

at least 16 different acoustic differences (Lisker, L., 1978). Further, no single acoustic attribute from these 16 is necessary to signal the distinction. Much of this redundancy in speech sounds results from the complexity of speech signal, a consequence of the fact that the structures that produce speech, the larynx and vocal tract, also are complex. Talkers morph their vocal tracts into widely varying shapes with different lengths. The surfaces inside vocal tracts vary considerably from tooth enamel to fleshy lips and soft palate, resulting in wildly varying absorption and reflection properties throughout the length of the vocal tract. Talkers produce multiple acoustic cues in a fashion that makes differences between speech sounds more perceptible (Kingston, J. and Diehl, R. L., 1994; Kluender, K. R., 1994) resulting in yet greater redundancy in the acoustic signal.

Languages have developed to be robust signaling systems, and distinctions between speech sounds do not rely upon the ability to make fine-grained discriminations bordering on thresholds of auditory systems. In this way, perception of speech bears little likeness to classic psychophysical studies of absolute thresholds and just-noticeable differences. Hearing the differences between vowels like [æ] (as in bat) and [ɛ] (as in bet), on the basis of gross spectral differences, shares little in common with psychoacoustic studies that demonstrate humans' abilities to detect changes as small as 1 Hz from a 1000 Hz sinusoid tone.

Because languages use distinctions that maximize acoustic differences and exploit auditory capacities, several observations follow. First, human infants are quite proficient at discriminating differences between speech sounds from a very early age. Three decades of studies document the impressive abilities of human infants, some less than one week old, to discriminate a wide variety of consonants and vowels

from across many languages (see, e.g., Jusczyk, P. W., 1981). A plethora of positive findings indicate that infants have the discriminative capacity necessary for most or all of the speech distinctions they will need to use in their language. Languages tend to use distinctions that are relatively easy to detect, and infant auditory abilities appear to be quite well developed. By 3 months of age, the human auditory system is nearly adult-like in absolute sensitivity and frequency-resolving power within the frequency range of most speech sounds (e.g., Olsho, L. W.et *al.*, 1988; 1987; Werner, L. and Gillenwater, J., 1990; Werner, L., 00386).

A second observation, perhaps more telling than the first, is that discrimination of speech contrasts by nonhuman animals appears to be generally quite good. There have been a fair number of demonstrations that animals can distinguish human speech sounds with facility (see, e.g., Kluender, K. R. *et al.*, 2005). Differences between vowel sounds present no apparent difficulty for nonhuman animals. Several species have been shown to distinguish voiced from voiceless stop consonants, or to distinguish consonants on the basis of place of articulation (even beyond the basic three places described for English above). Such findings are expected on the basis of languages tending to use distinctions that are relatively easy to discriminate.

1.59.3 Some Fundamentals of Perception

To mostly ill effect, for many years much of speech perception was studied in relative isolation from the study of perception more generally. In part, this state of affairs was encouraged by the focus of language researchers (linguists and psycholinguists) seeking to know more about elemental aspects of language use. Consistent with focus on apparently unique characteristics of human language, early speech researchers were encouraged to believe that perception of speech might be as unique as language itself. For this and other historical reasons, research in speech perception was often naïve to developments in related areas of perception.

1.59.3.1 The Inverse Problem

An enduring distraction for investigators studying speech perception has concerned the extent to which articulatory gestures (e.g., Fowler, C. A., 1986; Liberman, A. M. and Mattingly, I. G., 1985), acoustic patterns, patterns of sensory stimulation (e.g., Diehl, R. L. and Kluender, K. R., 1989), or some combination (e.g., Nearey, T. M., 1997; Stevens, K. N. and Blumstein, S. E., 1981) serve as proper objects of speech perception. Controversies concerning appropriate objects of perception generated more heat than light. However, debates concerning objects of perception cannot be resolved because the question itself is ill-posed, if not outright misleading. There are no objects of perception, neither for speech nor for perception in general. There is an objective for perception, which is to maintain adequate agreement between an organism and its world in order to facilitate adaptive behavior. Success with this objective does not require objects of perception.

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 may lead to inspection of real-world objects and events, patterns of light or sound pressure waves, transduction properties, or neural responses. By and large, however, viewing perception with a focus toward either distal or proximal properties falls short of capturing the essential functional characteristic of perception – the relationship between an organism's environment and its actions.

Much work in perception has been concerned, in one way or another, with addressing the inverse problem. The inverse problem emerges from the simple fact that information available to sensory transducers (eyes, ears, etc.) is inadequate to authentically reconstruct a unique distal state of affairs. In vision, for any two-dimensional projection, there are an infinite number of possible three-dimensional objects that could give rise to exactly the same two-dimensional (2D) retinal image (Figure 3). In audition, for any sound-pressure wave, there are an infinite number of sound producing events that could give rise to that waveform. One telling example of the difficulty of mapping from acoustics to a sound-producing event is the case of attempting to solve the inverse from waveform to simpler 2D surfaces (e.g., the shape of a drum.). Mathematicians have formally proved that even this relatively simple translation from waveform to plane geometry is impossible (Gordon, C. *et al.*, 1992).

Because multiple sound sources yield the same waveform, waveforms can never be more complex than characteristics of physical sources. Researchers within the field of speech perception have long been

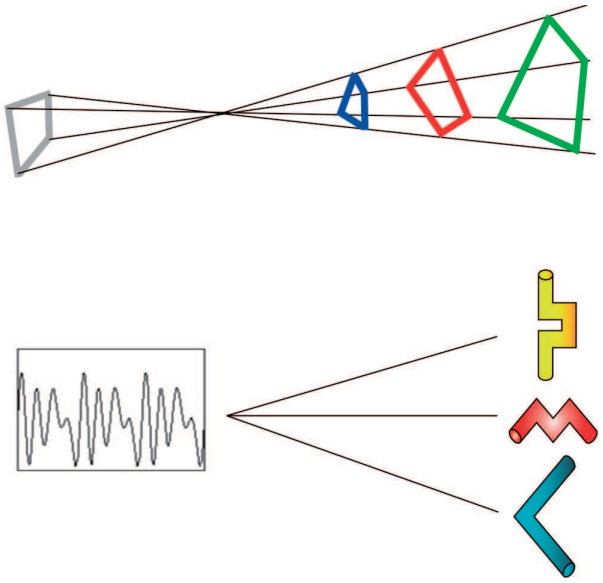

Figure 3 An infinite number of external three-dimensional objects give rise to the same two-dimensional retinal image (top, left). An infinite number of sound producing sources (characterized here as resonator shapes) give rise to the same waveform available to the ear.

familiar with appeals to perception via articulatory gestures as a simplifying construct, and there have been a series of efforts to extract gestures in order to facilitate machine speech recognition, albeit with very limited success. What physics demands, however, is that depiction of speech in terms of articulatory gestures gives only the illusion of simplicity. Because scientists are much better at measuring details of sounds than they are at measuring details of articulator activity, articulatory gestures appear simpler only because they are defined more abstractly and are measured with less precision. Because multiple resonator configurations can give rise to the same waveform, the acoustic waveform available to listeners is always less variable than articulation.

For all of the discussion that follows regarding specific issues concerning speech perception, speech typically will be described as sounds. This is not because sounds are legitimate objects of perception, but rather because, along the chain of events from creating patterns of sound pressure to encoding these patterns in some collection of neural firings to eliciting behavior, waveforms are public, easily measurable, and simpler than alternatives.

1.59.3.2 Why Perception Seems Veridical

If perceiving the true state of the world is impossible, one might ask why phenomenal experience is not fuzzy and uncertain. To effectively guide behavior and not leave the organism pondering multiple possibilities, all that is required is that the perceptual system comes to the same adaptive output every time it receives the same functional input. It is this deterministic nature of perception that prevents paralysis within a sea of alternatives. Phenomenal experience of certain reality does not depend upon authentic rendering of the world. Instead, phenomenal experience of a clear and certain world is the consequence of perceptual systems reliably arriving at deterministically unique outputs. It is this reliability that encourages certainty (Hume, D., 1748/1963), but reliability is not validity.

On rare occasions, perceptual systems do not converge upon a unique output and are left oscillating between equally fitting outputs when sensory inputs are not singly determinate (usually in response to impoverished stimuli.) Many readers are familiar with bistability when viewing Necker cubes. One such auditory experience is encountered when listening to a repeating synthesized syllable intermediate between [da] and [ta] or any other pair of similar speech sounds. When two perceptual outputs fit the input equally well, phenomenal experience oscillates between each percept (Tuller, B. *et al.*, 1994).

1.59.3.3 Information for Perception

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 head 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. 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).

This way of viewing information as a relationship is consistent with one of the fundamental characteristics of Shannon's information theory (Shannon, C. E., 1948; Weiner, N., 1948). Some readers may be familiar with Fletcher's pioneering applications of information theory to speech (Fletcher, H., 1953/1995). However,

the application here will be more akin to the approach of Attneave, F. (1954; 1959) and Barlow, H. B. (1961) for vision, an approach that remains highly productive (e.g., Barlow, H. B., 1997; 2001; Simoncelli, E. P. and Olshausen, B. A., 2001; Schwartz, O. and Simoncelli, E. P., 2001). One important point 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 it does not portray any essential characteristics about either transmitters or receivers. Within this information-theoretic sense, perceptual information exists in the relationship between organisms and their environments. This is the objective of perception.

Within a sea of alternative perceptual endpoints, agreement between the organism and environment aims to establish the alternative that gives rise to the most adaptive behavior. Information is transmitted when uncertainty is reduced and agreement is achieved between organism and environment. The greater the number of alternatives (uncertainty, unpredictability, variability, or entropy) there are, the greater the amount of information that potentially can be transmitted (ssee Figure 4(a)). There is no information when there is no variability. When there is no variability, there is total predictability and hence, no information transmitted. There is much that stays the same in the world from time to time and place to place, but there is no information in stasis. Uncertainty is reduced consequent to the perceiver's current experience (context) as well as past experiences with the environment (learning).

Although the amount of theoretical potential information transmitted is maximized at maximum entropy (total unpredictability or randomness), it is not advantageous for biological systems to shift their dynamic range as far as possible toward this maximum. In natural environments, this would result in diminishing returns if the system adjusts to register the last bits of near-random energy flux. Instead, biological systems should maximize the efficiency with which they capture information relative to the distribution of energy flux in real environments. The best estimate of statistics of natural environments is $1/f$ (pink) noise (Figure 4(b)). This simple power law with a negative exponent (f^1)is scale-invariant, and it is a ubiquitous characteristic across many systems from radioactive decay to fluid dynamics, biological systems, and astronomy. As one would expect, spectral density of fluctuations in acoustic power of music and speech varies as $1/f$ (Voss, R. F. and Clarke, J.,

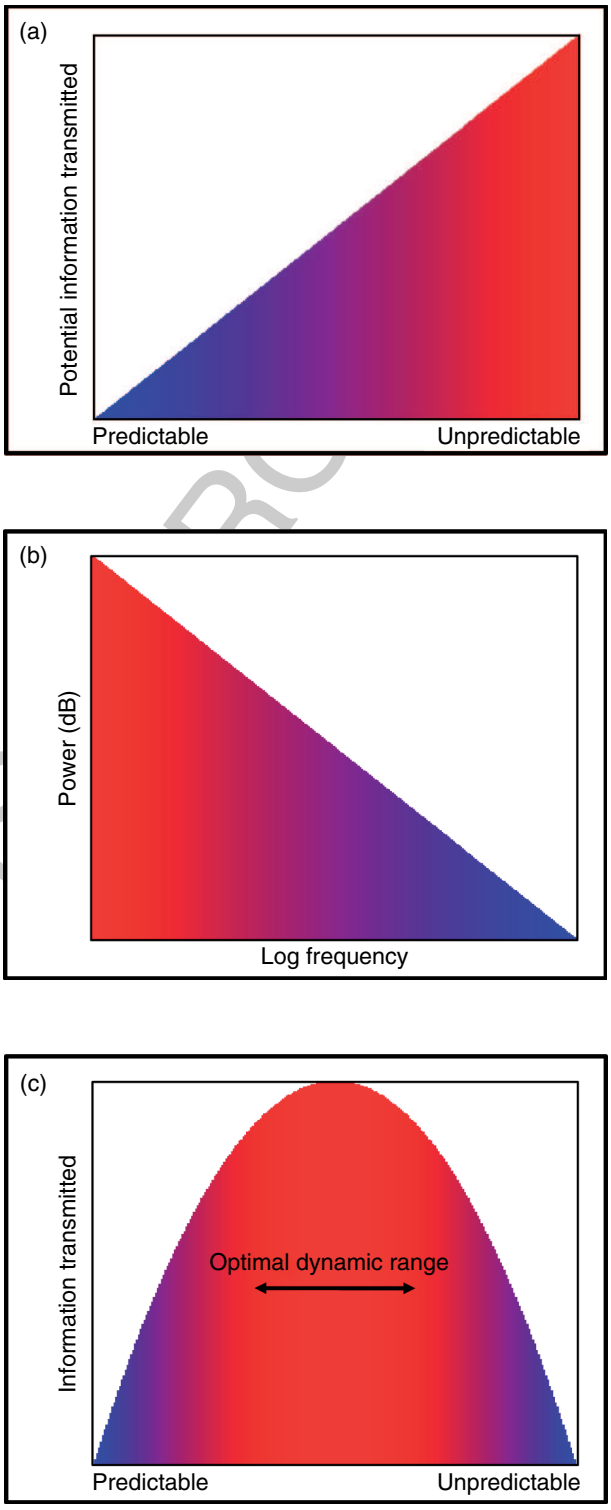

Figure 4 The greater the number of alternatives (uncertainty, unpredictability, variability, or entropy) there are, the greater the amount of information that potentially can be transmitted (a). There is no new information in what stays the same or is predictable. Relative power of energy flux in natural environments approximates 1/*f* (b). Information transmission optimized relative to energy flux in the environment (c). An ideal sensori-neural system should center dynamic range about this maximum.

1975; 1978.) Efficient information transmission for sensori-neural systems with limited dynamic range may be depicted best as the product of the positive exponential growth in information and the negative exponential of $1/f$. This yields the quadratic function shown in Figure 4(c) describing optimal transmission of information relative to energy flux in the environment.

1.59.3.4 Sensory Systems Respond to Change (and Little Else)

Given these facts about information, it is true and fortunate that sensori-neural systems operate as they do. Sensori-neural systems respond only to change relative to what is predictable or does not change. 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-eighteenth century. This has been demonstrated perceptually in every sensory domain. Humans have a remarkable ability to make fine discriminations, or relative judgments, about frequency and intensity. The number of discriminations than can be made numbers in the hundreds or thousands before full dynamic range is exhausted. Yet, most humans are capable of reliably categorizing, or making absolute judgments about only a relatively small number of stimuli regardless of physical dimension (e.g., Gardner, W.R. and Hake, H. W. 1951; Miller, G. A., 1956).

Sacrifice of absolute encoding has enormous benefits along the way to optimizing information transmission. Although biological sensors have impressive dynamic range given their evolution via borrowed parts (e.g., gill arches to middle ear bones), this dynamic range is always a fraction of the physical range of absolute levels available from the environment and essential to organisms' survival. This is true whether one is considering optical luminance or acoustic pressure. The beauty of sensory systems is that, by responding to relative change, a limited dynamic range shifts upward and downward to optimize the amount of change that can be detected in the environment at a given moment.

The simplest way that sensory systems adjust dynamic range to optimize sensitivity is via processes of adaptation. Following nothing, even a subtle sensory stimulus can trigger a strong sensation. However, when a level of sensory input is sustained over time, constant stimulation loses impact. This sort of sensory attenuation due to adaptation is ubiquitous, and has been documented in vision (Riggs, L. A. *et al.*, 1953), audition (Hood, J. D., 1950), taste (Urbantschitsch, 1876, cf. Abrahams, *et. al.*, 1937), touch (Hoagland, H., 1933), and smell (Zwaardemaker, H., 1895, cf. Engen, 1982). It is not uncommon to find instances where the terms adaptation and fatigue are used interchangeably. However, this equivocation is inappropriate. Consider, for example, visual dark adaptation whereby sensitivity to light is increased. Adaptation is a process whereby dynamic range adjusts, upward or downward, to maximize sensitivity to change. There are increasingly sophisticated mechanisms supporting sensitivity to change with ascending levels of processing, and several will be discussed in this chapter. Most important for now is the fundamental principle that perception of any object or event is always relative – critically dependent on its context.

1.59.4 Contrast and Low-Level Speech Perception

1.59.4.1 Contrast in General

Because it is only change that is perceived, perception at any particular time or place always depends on temporally or spatially adjacent information. Many instances of sensitivity to change are revealed through demonstration of contrast. For example, as depicted in Figure 5, a gray region appears darker against a white background and lighter next to a black background (see, e.g., Anderson, B. L. and Winawer, J., 2005). While examples abound for all modalities, sensory contrast has been most amply demonstrated in studies of visual perception, with contrast being a fundamental process that reveals edges of objects and surfaces.

Contrast effects are ubiquitous, and of course, they exist for audition (Cathcart, E. P. and Dawson, S, 1928/1929; Christman, R. J., 1954). Forms of auditory contrast are important for several aspects of speech perception. Over the past few years, multiple studies have provided evidence that simple processes of spectral contrast contribute to solving one of the most, if not the most, difficult questions concerning speech perception, co-articulated speech. Co-articulation is the spatial and temporal overlap of adjacent articulatory activities, and it is reflected in the acoustic signal by severe context dependence. Acoustic information specifying one speech sound varies substantially depending on surrounding sounds.

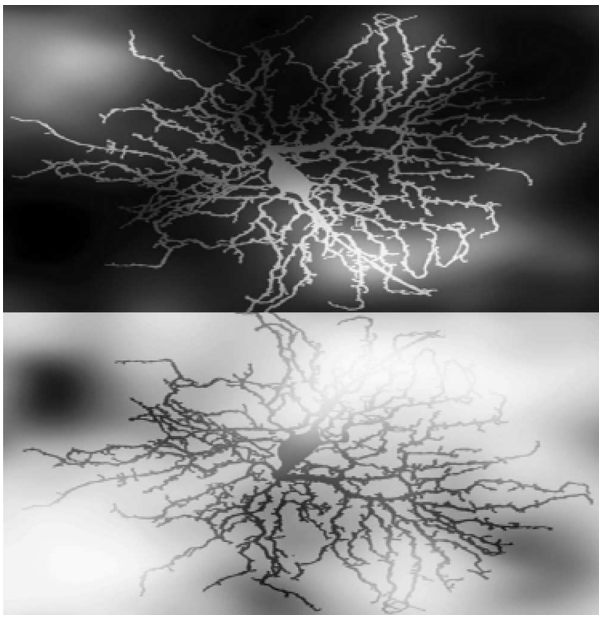

Figure 5 Contrast effects are ubiquitous in perception. For this example of lightness contrast, absolute lightness/darkness of these mirror-image neurons is identical.

1.59.4.2 Contrast and Perception of Co-articulated Speech

The problem for speech perception is how listeners hear a speech sound such as [d] when acoustic characteristics change dramatically depending upon sounds that precede and follow (e.g., vowels [e] versus [o]) (see Figure 6). Co-articulation presents a major challenge to automatic speech recognition (ASR) systems, which largely identify speech sounds on the basis of matching stored templates. Instead of storing a single template for [d], multiple templates must be stored for [d] following all other possible speech sounds, and each of those templates must be stored multiply for every instance of [d] preceding all other possible speech sounds. For ASR, this strategy using a geometrically expanding set of templates can be made to work so long as one has sufficient memory and sufficient processing speed to sort through templates. Not surprisingly, progress in ASR over decades is closely correlated with speed of microprocessors and price of memory (Lippman, R. P., 1996).

There is a consistent pattern to co-articulation that suggests a simpler solution. Adjacent sounds always assimilate toward the spectral characteristics of one another. Owing to mass and inertia of articulators (as well as planning), articulatory movements are compromises between where articulators have been and where they are headed. Because the acoustic signal directly reflects these articulatory facts, the

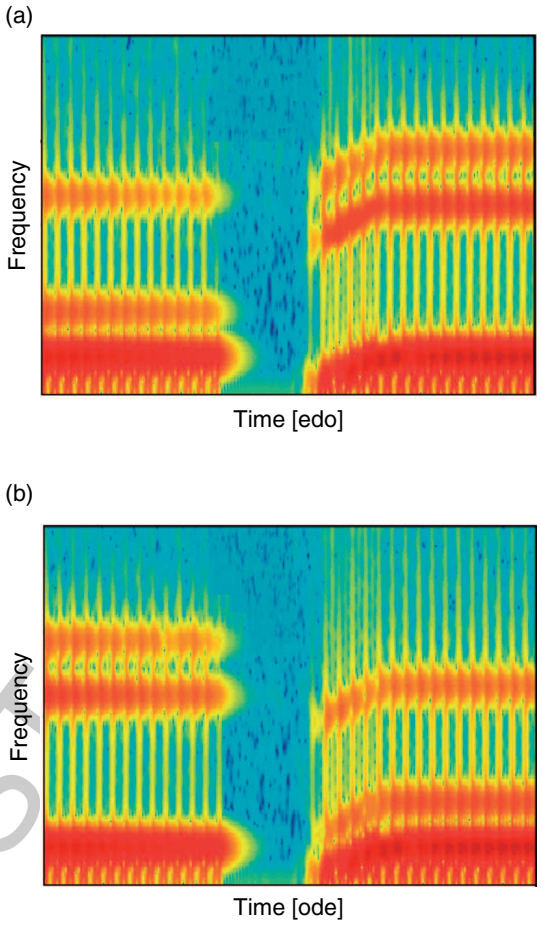

Figure 6 Schematic spectrograms of [edo] (top) and [ode] (bottom.) Note that acoustic properties of [d] depend upon characteristics of preceding and following vowel sounds.

frequency spectrum assimilates in the same fashion that speech articulation assimilates.

Lindblom, B. E. F. (1963) provided some of the best early evidence concerning how context systematically influences speech production. He reported that the frequency of the second formant (F_2) was higher in the productions of [dId] (did) and [dUd] (dud) than for the vowels [I] and [U] in isolation, and that F_2 was lower for vowels in [bIb] and [bUb]. In both contexts, F_2 frequency approached that of flanking consonants, which are higher for [d] than for [b]. In a subsequent study, Lindblom, B. E. F and Studdert-Kennedy, M. (1967) demonstrated that perception of co-articulated vowels is complementary to these facts of articulation. Listeners reported hearing /I/ (higher F_2) more often in [wVw] (lower F_2) context, and /U/ more often in [jVj] (higher F_2) context. Consonant context affected vowel perception in a manner complementary to the assimilative effects of co-articulation. Lindblom, B. E. F. and Studdert-Kennedy, M. (1967) wrote: "It is worth reiterating... that mechanisms of perceptual analysis

whose operations contribute to *enhancing contrast* in the above-mentioned sense are precisely the type of mechanisms that seem well suited to their purpose given the fact that the slurred and sluggish manner in which human speech sound stimuli are often generated tends to reduce rather than sharpen contrast. (p. 842, italics added)"

One of the most thoroughly investigated cases for perceptual context dependence concerns the realization of [d] and [g] as a function of preceding liquid (Mann, V. A., 1980) or fricative (Mann, V. A. and Repp, B. H., 1981). Perception of /d/ as contrasted with perception of /g/, can be largely signaled by the onset frequency and trajectory of the third formant (F_3). In the context of a following [a], a higher F_3 onset encourages perception of /da/ while a lower F_3 onset results in perception of /ga/. Onset frequency of the F_3 transition varies as a function of the preceding consonant in connected speech. For example, F_3-onset frequency for [da] is higher following [al] in [alda] than when following [ar] in [arda]. The offset frequency of F_3 is higher for [al] owing to a more forward place of articulation, and is lower for [ar].

Perception of /da/ and /ga/ has been shown to be complementary to the facts of production much as it is for CVCs. Listeners are more likely to report hearing /da/ (high F_3) when preceded by the syllable [ar] (low F_3), and hearing /ga/ (low F_3) when preceded by [al] (high F_3) (Mann, V. A., 1980; Lotto, A. J. and Kluender, K. R., 1998). In subsequent studies, the effect has been found for speakers of Japanese who cannot distinguish [l] and [r] (Mann, V. A., 1986), for prelinguistic infants (Fowler, C. A. *et al.*, 1990), and for avian subjects (Lotto, A. J. *et al.*, 1997.) The same pattern of findings has been replicated for perception of /d/ and /g/ following fricatives [s] and [ʃ] such that listeners are more likely to report hearing /d/ (high F_3) following [ʃ] (lower-frequency noise) and hearing /g/ (low F_3) following [s] (higher-frequency noise) (Mann, V. A. and Repp, B. H., 1981).

Co-articulation *per se* can be dissociated from its acoustic consequences by combining synthetic speech targets with nonspeech flanking energy that captures minimal essential spectral aspects of speech. Lotto, A. J. and Kluender, K. R. (1998) replaced [al] and [ar] precursors with nothing more than constant-frequency sinusoids set to the offset frequencies of F_3 for [al] and [ar] syllables. Perception of following [da-ga] shifted just as it did following full-spectrum [al] and [ar] (Figure 7).

Holt, L. L., *et al.*, (2000) replicated the Lindblom and Studdert-Kennedy findings (1967) with CVCs using the vowels [ɛ] and [ʌ] flanked by stop consonants [b] and [d]. They replaced flanking [b] and [d] with FM glides that tracked the center frequency of only F_2 for [b] or [d]. Again, the pattern of results for flanking nonspeech FM glides mimicked that for full-spectrum [b] and [d] syllable-initial and syllable-final transitions. Based upon the results for VCCVs (Lotto, A. J. and Kluender, K. R., 1998) and these results for CVCs, one can conclude that much of perceptual accommodation for co-articulation is not restricted to speech-like signals. All of the findings are consistent with spectral contrast, whereby the spectral composition of context serves to diminish or enhance the perceptual efficacy of spectral components for adjacent sounds.

In keeping with typical usage, the term contrast has been used in a largely descriptive way thus far. There are a large number of experimental precedents for spectral contrast – often called auditory enhancement, and these precedents provide more specific hypotheses. Summerfield, Q. *et al.*, (1984) established the existence

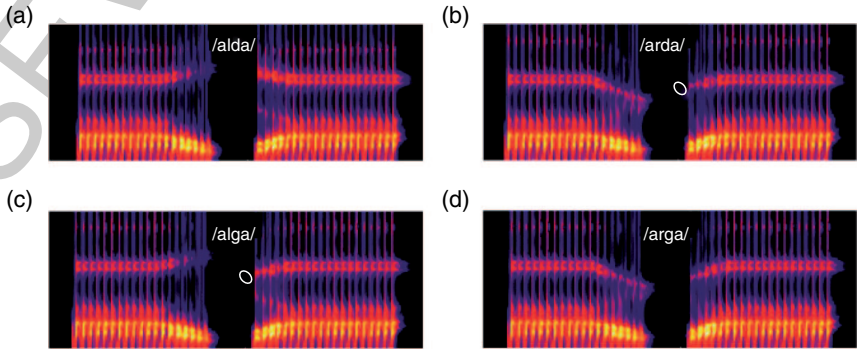

Figure 7 Due to co-articulation, acoustic properties of any speech sound become more similar to the properties of sounds preceding and following. This assimilation is a property of all fluent connected speech. Here, acoustic characteristics of [d] (e.g., F_3, marked by ellipse) following [r] are very similar to those of [g] following [l]. Listeners hear the same consonant vowel (CV) as /d/ following [ar] and as /g/ following [al].

of an aftereffect in vowel perception. When a uniform harmonic spectrum was preceded by a spectrum that was complementary to a particular vowel with troughs replacing peaks and vice versa, listeners reported hearing a vowel during presentation of the uniform spectrum. The vowel percept (for the uniform spectrum) was appropriate for a spectrum with peaks at frequencies where there were troughs in the preceding spectrum.

Summerfield, Q. *et al.* (1984) noted that perceiving vowel sounds in uniform spectra (following appropriate complementary spectral patterns) has a well-known precedent in psychoacoustics. This oft-reported finding is that, if just one member of a set of harmonics of equal amplitude is omitted from a harmonic series and is reintroduced, then it stands out perceptually against the background of the preexisting harmonics (Schouten, J. F., 1940; Green, D. M. *et al.,* 1959; Cardozo, B. L., 1967; Viemeister, N. F., 1980; Houtgast, T., 1972). Viemeister (1980), for example, demonstrated that the threshold for detecting a tone in a harmonic complex is 10–12 dB lower when the incomplete harmonic complex (missing the target tone) precedes the tone as compared to when the onset of the inharmonic complex is coincident with that for the target tone. This was referred to as an enhancement effect. Viemeister (1980) then examined a number of properties of this effect, finding that the complex need not be harmonic and that noise maskers or band-pass noise signals also served to enhance the detection of the tone. He also found the effect over a wide range of intensities for maskers and targets.

Summerfield and colleagues (1984; 1987) suggested that their demonstration of vowel aftereffects may be rooted in peripheral sensory adaptation. One could suggest that neurons adapt, and the prominence of the added harmonic is due to the fact that neurons tuned to its frequency were not adapted prior to its onset. Alternatively, some researchers (e.g., Houtgast, T., 1974; Moore, B. C. J. and Glasberg, B. R., 1983) have suggested that rapid adaptation serves mostly to enhance onsets selectively, with suppression being a process through which differences in level of adjacent spectral regions in complex spectra (e.g., formants in speech signals) are preserved and/or enhanced.

Viemeister, N. F. and Bacon, S. P. (1982) showed that, not only was an enhanced target tone more detectable, the tone also served as a more effective masker of a following tone. They suggested that suppression must be included in an adaptation scenario to place it in closer accord to this finding. Different frequency components of a signal serve to suppress one another, and two-tone suppression has been cast as an instance of lateral inhibition in hearing (Houtgast, T., 1972). Investigators have argued that suppression helps to provide sharp tuning (e.g., Wightman, F. L. *et al.*, 1977; Festen, J. M. and Plomp, R., 1981), and with respect to speech perception, Houtgast, T. (1974) has argued that this process serves to sharpen the neural projection of a vowel spectrum in a fashion that enhances spectral peaks.

Many neurophysiological observations bear upon enhancement effects. In particular, a number of neurophysiological studies of auditory nerve (AN) recordings (e.g., Smith, R. L. and Zwislocki, J. J., 1971; Smith, R. L., 1979; Smith, R. L. *et al.*,, 1985) strongly imply a role for peripheral adaptation. Delgutte and colleagues (Delgutte, B., 1980; 1986; 1996; Delgutte, B. *et al.*, 1996; Delgutte, B. and Kiang, N. Y. S., 1984) have established the case for a much broader role of peripheral adaptation in perception of speech. Delgutte notes that peaks in AN discharge rate correspond to spectro-temporal regions that are rich in phonetic information, and that adaptation increases the resolution with which onsets are represented. He also notes neurophysiological evidence that "adaptation enhances *spectral contrast* between successive speech segments" (Delgutte, B. *et al.*, 1996, p. 3, italics added). This enhancement arises because a fiber adapted by stimulus components close to its CF is relatively less responsive to subsequent energy at that frequency, while stimulus components not present immediately prior are encoded by fibers that are unadapted – essentially the same process offered by psychoacousticians but now grounded in physiology. In addition, Delgutte notes that adaptation takes place on many timescales, and is sustained longer with increasing level in the auditory system.

Inspired by the vowel aftereffect studies by Summerfield and his colleagues (1984; 1987), Coady, J. A. *et al.,* (2003) sought to make clearer the connections between experiments using very simple nonspeech flanking stimuli (e.g., FM glides) and Summerfield's studies using rich spectra that were complementary to those for vowel sounds. Although sine waves and FM glides have often been used as nonspeech proxies for formants, such sounds have limited resemblance to speech formants. While it is true that spectrograms illustrate formants as bands of energy, and formant transitions as bands of energy traversing frequency, such descriptions can be misleading. For example, if fundamental frequency (f_0) is constant, individual harmonics of the fundamental do not change frequency at all, and all that changes are relative amplitudes of

harmonics. Individual frequency components of the speech spectrum change frequency no more than f_0 changes.

Coady, J.A and colleagues (2003) used VCV sequences for which the initial vowel ([e] or [o]) affects perception of the following consonant (/ba/ or /da/). In addition to creating synthetic vowels [e] and [o], they created spectral complements of these vowels [∼e] and [∼o] by creating troughs where formants occurred for [e] and [o]. These precursor vowel-complements altered perception in a fashion opposite that of normal (noncomplement) vowels because troughs in energy increased excitability within a frequency range for which excitability was attenuated when a frequency prominence (formant) was present in a normal vowel. In addition, they demonstrated that these perceptual effects relied substantially upon spectral characteristics of onsets.

It appears that the same underlying processes account for effects of both very simple nonspeech precursors and spectrally rich vowel-like complements. Although more complex or domain-limited theories of speech perception have been proposed to explain perception of co-articulated speech, the above patterns of perception with both simple and complex stimuli suggest that spectral contrast is an important part of the explanation for perceptual accommodation of co-articulated speech.

1.59.4.3 Broader Spectral and Temporal Effects

Contributions of spectral contrast to perception of co-articulated speech are narrowly focused in both time and frequency. Processes through which the auditory system optimizes detection of spectral change operate over durations on the order of less than 0.5 s, and spectral components of interest are relatively local (e.g., formants.) In keeping with the fundamental principle that, in the interest of maximizing transmission of new information, perceptual systems respond primarily to change, long-term signal characteristics that do not change should also alter perception in similar ways.

For vision, perceivers maintain color constancy in the face of changes in level or spectral composition of illumination, respectively. The visual system adjusts for the spectral composition of illumination (e.g., sunlight versus tungsten or fluorescent lighting), while maintaining relatively consistent perception of color under widely varying viewing conditions. Analogous challenges arise for hearing. When auditory experiments are conducted in the laboratory, experimenters typically endeavor to maintain consistent response across all frequencies through the use of high-quality audio equipment and headphones. However, in real-world listening environments, the spectrum is virtually always colored by characteristics of the listening environment. Energy at some frequencies is reinforced by reflective properties of surfaces, while energy at other frequencies is dampened by absorption properties of materials and shapes of objects in the environment. For hearing to be most effective, listeners must adapt to reliable spectral characteristics in order to be maximally sensitive to the most informative characteristics of sounds.

Kiefte, M. and Kluender, K. R. (in press) used vowel sounds to examine how auditory systems may adapt to predictable (redundant) spectral characteristics of the acoustic context in order to be more sensitive to information-bearing characteristics of sounds. Simple vowel sounds are useful in this application because it is known that listeners use both spectrally narrow (formant peaks) and broad properties (gross spectral tilt) to perceive vowel sounds. For example, low frequency F1 is heard as /u/ (as in 'boot') when accompanied by a low frequency F2, and as /i/ (as in 'beet') when accompanied by higher frequency F_2. In addition, gross spectral tilt, the relative balance between low- and high-frequency energy, is quite different for these vowels. The vowel [u] has more low- than high-frequency energy, resulting in a gross spectral tilt of rapidly declining energy as a function of increasing frequency. In contrast, [i] has relatively more high-frequency energy, and energy decreases much more gradually with increasing frequency. When listening to isolated vowel sounds, listeners use a combination of both formant frequencies and gross spectral tilt to identify vowels (Kiefte, M. and Kluender, K. R., 2005).

Kiefte and Kluender created a matrix of vowel stimuli that varied perceptually from /u/ to /i/ in two dimensions: center frequency of F_2 and gross spectral tilt. Along one dimension, center frequency of F_2 varied from low ([u]) to high ([i]). Along the second dimension, gross spectral tilt varied in the same stepwise fashion, from a spectral tilt characteristic of [u] to one characteristic of [i]. Listeners identified this matrix of vowel sounds preceded by a synthesized rendition of the sentence "You will now hear the vowel. . ." When long-term spectral tilt of context sentences was altered to match that of the following vowel, listeners relied virtually exclusively upon the frequency of F_2 when identifying /i/ and /u/. This pattern of performance indicates that tilt, as a predictable spectral property of

acoustic context, was effectively canceled out of perception (Figure 8).

Encoding of predictable characteristics such as spectral tilt requires only the most abstract characterization of the spectrum, with little spectral detail. Changes in spectral tilt can be due to changes in acoustic properties of the physical environment, or particular to speech, changes in emotional state or speaker identity (Klatt, D. H., 1982). Neither of these properties change rapidly and both are relatively stable across time relative to the rapid spectral changes found in speech. Kiefte and Kluender then tested whether perception also compensates for local spectral properties when these properties are stable across time.

Using the same sentence context, sentences were processed with a single-pole filter which corresponded exactly to the frequency and bandwidth of F_2 of the target, yielding intact sentences with an additional constant-frequency spectral peak added throughout. When presented with a stimulus in which F_2 was an unchanging acoustic property of the context, listeners relied largely upon global spectral characteristics (tilt) for identification of the target vowel. Effects of preceding context are not restricted to gross spectral properties. Perceptual cancellation of predictable spectral characteristics also occurs for local, relatively narrowband spectral characteristics of the acoustic context.

Perceptual cancellation of predictable acoustic context does not depend upon preceding context being speech (Holt, L. L., 2005; Stilp, C. and Kluender, K. R., 2006), nor does it depend upon the context being identical trial to trial (Kiefte, M. and Kluender, K. R., in press). Listener performance provides evidence that the auditory system is quite adept at factoring out predictable characteristics of a listening context, and is consequently more sensitive to informative changes in spectral composition across time.

Underlying mechanisms by which the auditory system calibrates for characteristics of acoustic context have not yet been extensively investigated, and are not yet understood. Some primary auditory cortex (AI) neurons encode spectral shape with respect to both broad and narrow complex spectral shapes (Barbour, D. L. and Wang, X., 2003), and neurons in AI are sensitive to the relative probabilities of pure tones of different frequencies in an extended sequence of tones (Ulanovsky, N. *et al.*, 2003). Calibrating to acoustic context in the service of enhancing sensitivity to change would have been efficacious since the very first auditory systems, even before the advent of

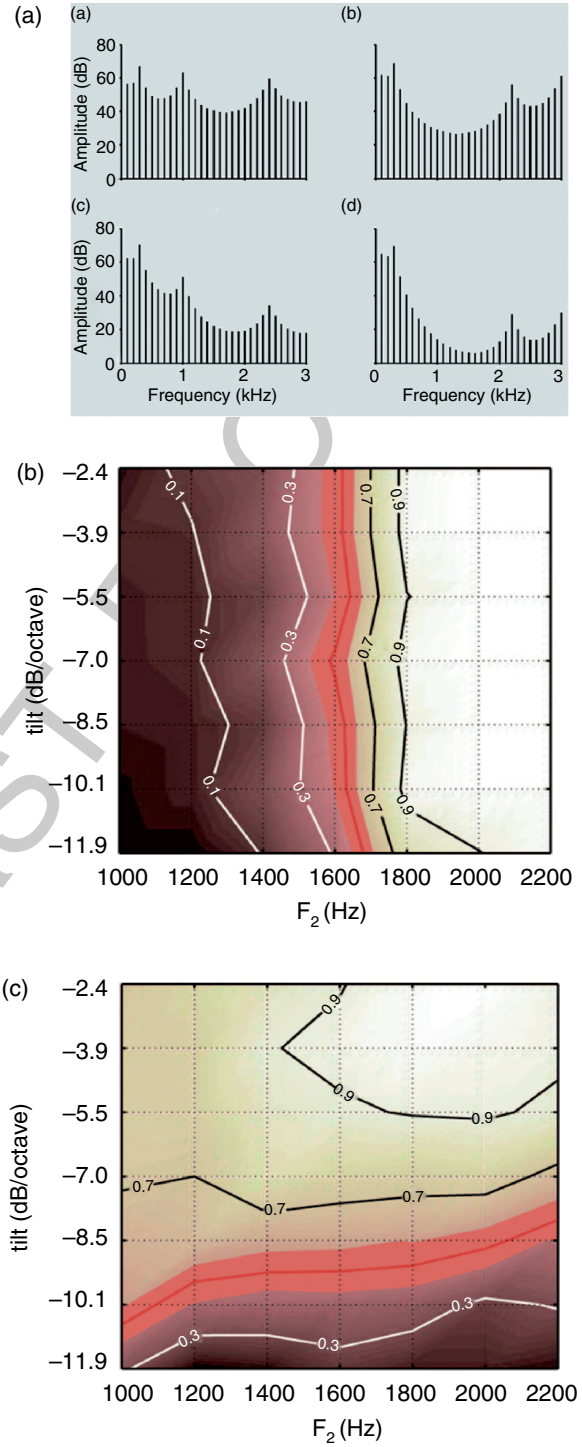

Figure 8 Spectra of vowel sounds [u] and [i] vary both in center frequency of F_2 and overall spectral tilt, and listeners use both acoustic properties to identify these vowels. Kiefte, M. and Kluender, K. R. (in press) created sounds that varied in each dimension independently (a). When listeners heard these vowel sounds following a precursor that was filtered to have the same spectral tilt as a target vowel, listeners used F_2 frequency exclusively when identifying the vowels (b). The complementary pattern of results, exclusive use of tilt, obtains when precursors include a constant-frequency spectral peak matching F_2 of the target sound (c).

neocortex. It is possible, or perhaps likely, that brainstem processes play an important role.

Projections from superior olive to outer hair cells, collectively called the medial olivocochlear (MOC) efferent system have been hypothesized to provide adjustments of basilar membrane tuning to improve resolution of signals against background noise. Kirk, E. C. and Smith, D. W. (2003), for example, hypothesized the MOC evolved as a mechanism for unmasking biologically significant acoustic stimuli. A few synapses further from the cochlea, inferior colliculus (IC) neurons adapt to structural regularities in natural sounds in a fashion that increases information transmission (Escabi, M. A. *et al.*, 2003), and stimulus specific adaptation much like that found by Ulanovsky, N. *et al.* (2003) also has been demonstrated in IC (Pérez-González, D. *et al.*, 2005).

1.59.5 Maximizing Transmission of Speech Information with Multiple Dimensions

1.59.5.1 Speech Perception Uses Multiple Sources of Information

A signature property of speech perception is its extreme resilience in the face of dramatic signal degradation. For example, listeners understand speech at signal-to-noise ratios less than 0 dB, and they understand speech either when all energy is removed above 1500 Hz or when all energy is removed below 1500 Hz. Listeners can understand speech when the only information available is fluctuations in amplitude of eight or so bands of noise across frequency (Shannon, R. V. *et al.*, 1995), and some listeners can understand speech consisting of little more than sine waves that track the center frequencies of formants (Remez, R. E. *et al.*, 1981). In large part, these as well as other demonstrations of perceptual resilience can be explained by the fact that listeners can rely upon experience with speech that far exceeds experience with any other type of sounds. This power of experience rests upon the high degree of redundancy within the speech signal.

Experience with redundancy is among the reasons why performance of listeners detecting tones in quiet or detecting differences in pitch between two tones are rather poor predictors of the ability of the same listeners to understand speech. Listeners who suffer significant hearing loss can, nonetheless, manage to understand speech until the level of impairment becomes severe or there is substantial background noise competing with the speech signal. Profoundly deaf people who have received cochlear prosthetics (electrodes implanted into the cochlea) and who consequently receive extremely degraded signal via as few as one–four electrode contacts, nevertheless, can often understand speech, sometimes sufficiently well to talk on the telephone.

Redundancy does not distinguish speech from other objects and events in the world. For example, Attneave, F. (1954) notes that information received by the visual system is redundant with sensory events that are highly interdependent in both space and time, and this is simply because "the world as we know it is lawful." (p. 183). Because multiple attributes are used when perceiving speech, the presence of one attribute can compensate for the absence of another, as increasing the magnitude of one source of information serves to compensate for decrease of another. There have been many demonstrations of these trading relations (e.g., Repp, B. H., 1982).

Using multiple stimulus attributes is common to perception across modalities. For example, multiple monocular and binocular cues contribute to visual perception of distance. While individual neurons rarely provide high fidelity, populations of neurons acting in concert robustly encode information even when signals are substantially degraded. Implicit to population encoding is the fact that relationships between activities of multiple neurons conspire for effective perception. It is the correlation of activity across neurons (i.e., redundancy) that makes sensorineural systems robust. Exploiting correlations among multiple attributes in speech perception provides another example of maximizing the performance of perceptual systems by extracting predictability in the service of emphasizing change. The quintessential example of combining multiple acoustic attributes is categorical perception (Figure 9).

Although rarely recognized as being so (Kluender, K. R., 1988; 1994), perceptual constancy and categorization share a great deal in common. A classic definition of categorization is that it permits treating discriminably different examples as functionally equivalent. A virtue of categorization typically is presented as efficiently directing responses to functionally equivalent objects or events. Similarly, perceptual constancy maintains when discriminably different exposures (varying with size, orientation, etc.) are treated as equivalent. For example, the apparent size of an object remains the same even when brought nearer or farther from the perceiver. And, perceived shape stays the same across rotations. The simple observation that nonhuman

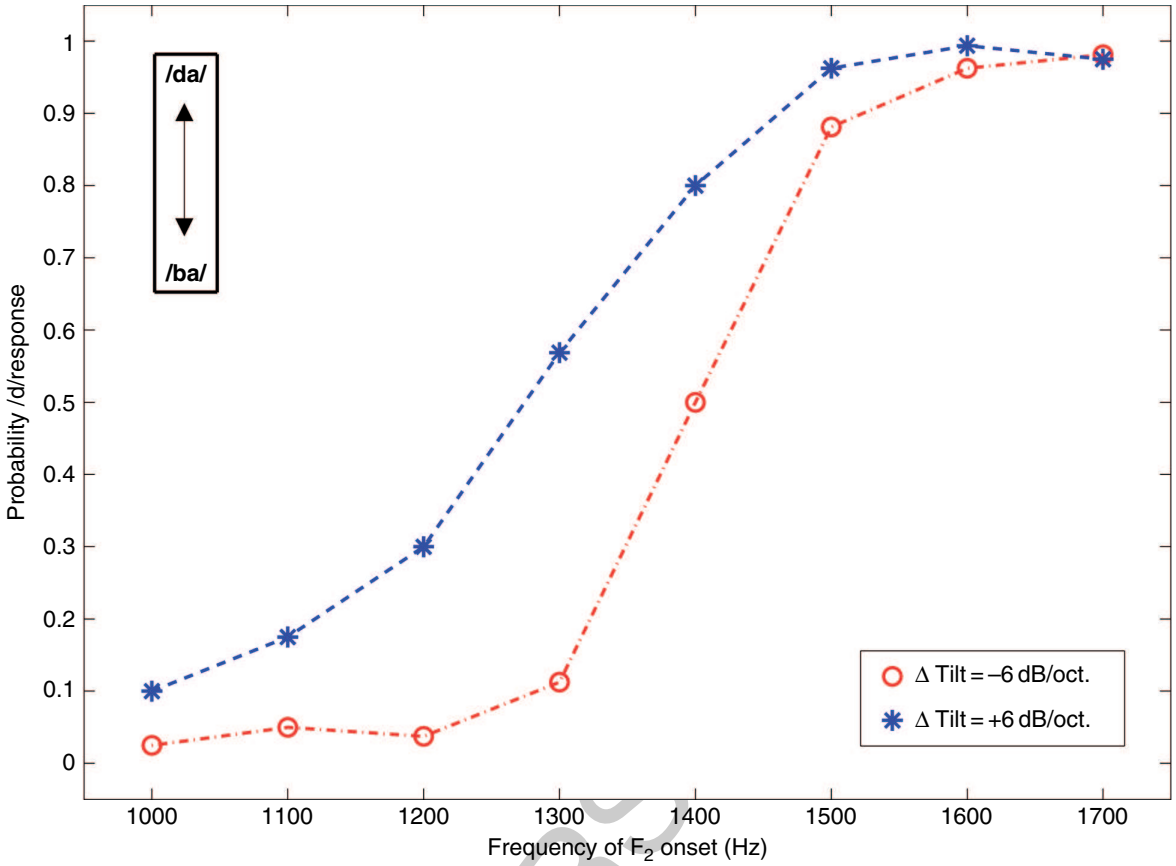

Figure 9 Listeners' use of multiple acoustic attributes is commonly demonstrated by experiments in which two attributes are independently manipulated. In this example from Alexander, J. M. and Kluender, K. R. (2005), when relative spectral tilt is positive (blue), listeners hear sounds with lower F_2 onset frequency as /d/. One attributes trades against another, as more of one compensates for less of another.

animals manage to navigate their worlds is ample testimony to their ability to maintain perceptual constancy. Nonhuman animals also have been shown to categorize both visual images and acoustic stimuli such as speech (Herrnstein, R. J., 1984; Kluender, K. R. *et al.*, 1987).

Confusion between perceptual constancy and perceptual categorization is common in descriptions of speech perception as somehow arriving at appropriate phonetic categories. However, there have been a fair number of instances for which researchers adopted perceptual constancy as the preferred description (see, e.g., Kuhl, P. K., 1978; 1979; 1980; 1987). Here it is proposed that categorical perception be thought of as perceptual constancy. To the extent that categorization is only a more abstract manifestation of constancy, choosing constancy may not be a particularly provocative choice. If one considers perception of speech to involve perceptual constancy, commonalities with similar perceptual achievements are revealed and surplus cognitive content typically ascribed to categorization is avoided.

1.59.5.2 Categorical Perception

Categorical perception is the most well-known pattern of perceptual performance with speech sounds. Three features define categorical perception: a sharp labeling (identification) function, discontinuous discrimination performance (near-perfect across identification boundary and near-chance to either side), and the ability to predict discrimination performance purely on the basis of labeling data (Wood, C. C., 1976). All three defining features of categorical perception arise naturally from the principle of discovering (and perceptually absorbing) predictability in the interest of maximizing sensitivity to change.

Returning to the fact that speech sounds are comprised of multiple acoustic attributes, many of which are redundant, one acoustic attribute serves to predict the occurrence of another. Through experience, perceptual processes come to absorb these correlations in a way that increases efficiency. There is no information in predictability. When perceptual systems encode correlations among attributes, there are

two consequences. First, there is a decrease in sensitivity to differences between two sounds that share the same pattern of correlation among the same set of attributes. Second, two sounds with different patterns of correlation become easier to distinguish. For speech, detection of differences between functionally different speech sounds is optimized to the extent that perceptual processes absorb redundancies across acoustic attributes that co-vary as properties of the same consonant or vowel (Figure 10).

1.59.5.2.1 Principal components analysis: An analogy

This perceptual processing can be compared with the statistical technique 'principal component analysis' (PCA; see, e.g., Dillon, W. R. and Goldstein, M., 1984). For PCA, one begins with a correlation matrix of multiple variables, created to assess the degree to which each variable is correlated with every other variable across many observations. From this correlation matrix, it is possible to determine weighted combinations of variables, vectors, which account for as much shared variance as possible. To the extent that multiple observations reveal covariance among variables, a limited number of vectors (few relative to the number of variables) can account for a high percentage of the total variance across observations.

PCA is being used here only as analogy because it is unlikely that real neurons adhere to formal restrictions on how vectors are chosen, and the ways PCA fails as analogy are themselves illuminating. First, PCA is a linear analysis, and it is well-known that sensory processes are nonlinear. Second, PCA assumes normally distributed values, and the real world complies with this assumption only to varying extents. A related analysis, independent component analysis (ICA; see, e.g., Hyvärinen, A. and Oja, E., 2000) does permit violations of assumption of normality, and may come a bit closer to modeling neural processing. Third, PCA, but not networks of neurons, requires that vectors be ordered from most to least amount of variance accounted for, and these vectors must be uncorrelated (eigenvectors).

The issue concerning orthogonality is interesting in two ways. First, while perfect efficiency is achieved if every vector shares no variance with any other vector, achieving this goal is unlikely in a neural system. A second point is more informative. Here, perception is being construed as a cascade of processes, each working to extract redundancy from the outputs of earlier processes. To the extent that outputs of prior processes even approach orthogonality, this would seem to imply

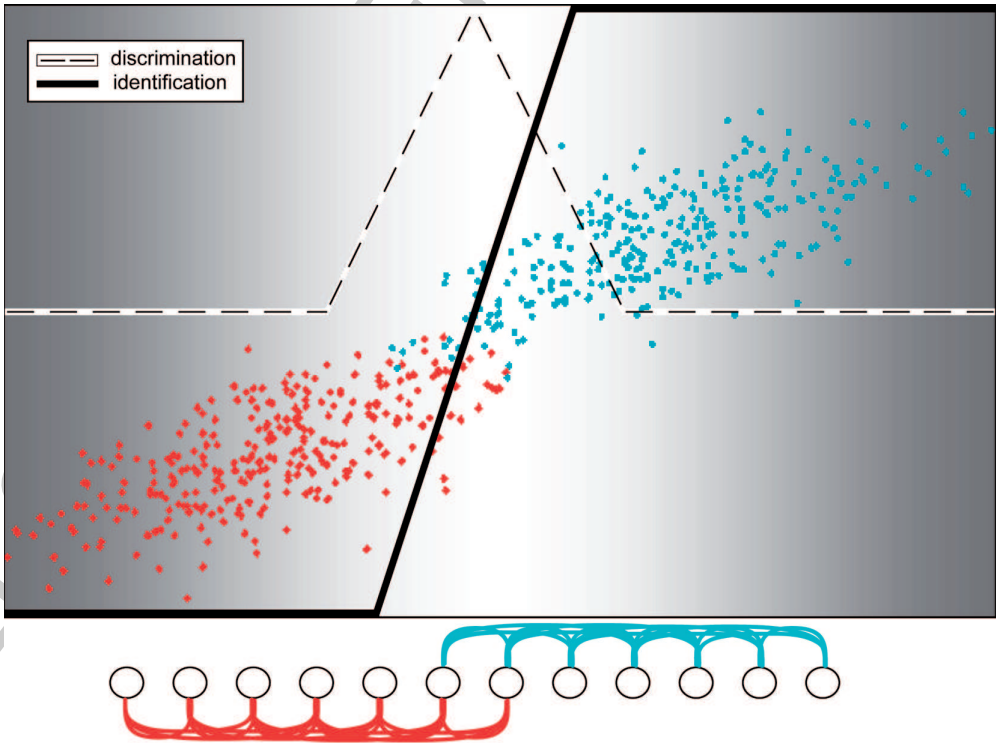

Figure 10 Categorical perception can be explained by auditory systems treating sounds with attributes that share the same correlation structure as equivalent. By this explanation, detection of differences between two complex sounds is greatest when each sound corresponds to a different pattern of experienced correlation.

that seizing upon correlation again would become increasingly implausible. The solution to this seeming dead end is that, with every successive reduction of redundancy, information over which processing operates expands in space, frequency, time, and any other dimension of interest. Thus, statistical relationships that hold relatively locally do not constrain correlations at the next coarser grain of processing. It is worthwhile to note the parallels between this statistical analogy and hierarchical organization of perceptual processing in the central nervous system.

Finally, there is a parallel between practical use of PCA and the prior argument that perceptual systems should efficiently optimize, not maximize, information transmission. Figure 4(c) depicts a sensori-neural system tuned to focus dynamic range in a way that optimizes information transmission relative to the distribution ($1/f$) of energy flux in the environment while neglecting the last bits of potential information. At an extreme, PCA also permits characterization of all of the variance across observations when the number of eigenvectors is equal to the number of observations. However, the virtue of PCA is that the majority of, but not all, variance can be accounted for by relatively few vectors. Efficiencies are gained by capturing correlations within a few vectors. More eigenvectors are discarded than are saved, despite the fact that some snippets of information are lost.

From the analogy to PCA, it is easy to envision efficient sensori-neural encoding that extracts reliable correlations between multiple attributes across observations. This process can be instantiated in perceptrons (Rosenblat, F., 1958), the simplest connectionist networks. Most important to the analogy is that experience is encoded efficiently in a way that decreases sensitivity to differences between stimulus inputs that share the same correlation vector (or attractor) and increases sensitivity to differences between inputs that correspond to different vectors (see, e.g., Olshausen, B. A. and Field, D. J., 1997; Simoncelli, E. P. and Olshausen, B. A., 2001).

1.59.5.2.2 Phonemes as Correlations?

Through experience with redundant attributes, simplified encoding of inputs as patterns of correlation serves as grist for consequent processing. It could be argued that vectors in a correlation matrix correspond to the putative linguistic units called phonemes (Kluender, K. R. and Lotto, A. J., 1999); however, the same basic principles apply continuously along the chain of processing, and such a demarcation would be artificial. The grain of synthesis steadily increases, from attributes that are spectrally and temporally local to those that extend across multiple attributes derived from preceding processing through to lexical organization. For example, Kingston, J. and Diehl, R. L. (1994) hypothesized intermediate perceptual properties (IPPs) may correspond to earlier analysis, and it is likely that statistical properties of acoustically simpler vowel and consonant sounds are extracted prior to those for more complex speech sounds. Further, as described below, additional reification of phoneme-like dimensions may await lexical organization.

Perhaps, no real harm may be done if one suggests that some correlations are phonemes per se, if only as a placeholder. However, it is important to distinguish two ways of thinking about this. The first way is common or even typical. One could suggest that the task for a listener is to identify consonants and vowels as individual psychological entities, and those entities are represented by correlations among attributes.

The second way, suggested here, is that consonants and vowels are revealed much more by what they are not than by what they are. Contrasts between sounds, not commonalities, are emphasized in speech perception. Through experience, perceptual processes become especially sensitive to acoustic differences that distinguish different consonants and vowels as a consequence ignoring differences among multiple acoustic properties that share a history of co-occurrence. What matters are distinctions between speech sounds, which are enhanced perceptually, not consonants and vowels themselves. Listeners hear the sounds of a language by virtue of learning how they are distinguished from all other consonants and vowels. This idea was most explicitly expressed by linguists Roman Jacobson and Morris Halle in their classic book *Fundamentals of Language* (1971) "All phonemes denote nothing but mere otherness" (p. 22).

1.59.5.2.3 Categorical Perception as Competing Correlations

Returning to categorical perception, one can understand how these patterns of performance emerge naturally from perceptual systems exploiting redundancies among attributes in the service of maximizing sensitivity to change. Following experience with correlations among acoustic attributes, listeners are relatively unlikely to detect differences among complex sounds that share the same correlation structure. The fact that modest changes are perceptually neglected is consistent with the fact that listeners can understand speech when some acoustic properties

(e.g., energy above 1500 Hz) are absent. When the correlation structure is not violated too severely, perception overcomes perturbations, and even absence, of some attributes that normally contribute to the correlation. All that is required are sufficient attributes to get to the right vector.

This lack of sensitivity to perturbations among inputs accounts for the finding that discrimination performance is near chance for different instances of the same consonant or vowel. If two different stimuli fit the same correlation structure relatively well, the same perceptual consequences obtain. Complementary to this lack of discrimination for sounds that share the same correlation structure, discrimination performance is exquisite when the speech sounds to be discriminated are associated with two competing correlation structures. For these cases, discrimination is especially good because, by virtue of perceptual processes extracting redundancies within separate correlation structures, detection of change (information transmission) is optimized.

If categorical perception is only another example of perceptual constancy operating within general principles of perceptual organization, why did so many researchers believe categorical perception to be unusual? One reason is because categorical perception was routinely contrasted with psychophysical data from experiments employing very simple stimuli (typically unidimensional) of limited ecological significance. Equally important, comparisons were made to stimuli with which subjects have little or no experience before coming to the experimental session. Classic psychoacoustic experiments using pure tones, noise bursts, and inharmonic complexes have great utility for interrogating operating characteristics of sensory transduction absent content or experience. Thresholds for energy detection or sensory change are valuable things to know, but these are not as informative with respect to perception as it guides real activities in a real world.

When investigators use stimuli that are complex and familiar, signature response patterns of categorical perception are found. Categorical perception has been reported for musical intervals (Burns, E. M. and Ward, 1974; 1978; Smith, J. D. *et al.*, 1994) and tempered triads (Locke, S. and Kellar, L., 1973). Visually, humans categorically perceive human faces (Beale, J. M. and Keil, F. C., 1995) and facial expressions (Etcoff, N. L. and Magee, J. J., 1992; Calder, A. J. *et al.,* 1996; de Gelder, B. *et al.*, 1997; Young, A. W. *et al.,* 1997), as well as faces of different species (Campbell, R. *et al.,* 1997). When human observers are trained with artificial categories, they demonstrate increased perceptual sensitivity for items that are categorized differently (Goldstone, R. L., 1994). When monkeys are trained to respond differentially to clear examples of cats versus dogs (initially novel categories for monkeys), behavioral responses to stimuli along a morphed cat/dog series exhibit sharp crossovers at the series midpoint (Freedman, D. J. *et al.*, 2001). Rather than being specific to speech, categorical perception is a general property of any perceptual system consequent to experience with rich regularities of natural objects and events.

Categorical perception appears to be an emergent property of any perceptual system that is shaped by experience. Damper, R. I. and Harnad, S. R. (2000) reviewed evidence from human and animal listeners as well as from neural network models. They concluded that any number of generalized learning mechanisms can account for categorical perception. Models ranging from simple associative networks (e.g., Anderson, J. A., *et al.*, 1977) to back-propagation networks with no hidden units (e.g., Damper, R. I. *et al.*, 2000) exhibit categorical perception. Because categorical performance arises from a variety of simple learning algorithms that seize upon reliable statistics in their inputs, Damper, R. I. and Harnad, S. R. (2000) conclude that specialized processing is not necessary, and that "any general learning system operating on broadly neural principles ought to exhibit the essentials of [categorical perception]" (p. 862).

1.59.5.2.4 Multimodal Interactions are Expected

Thus far, and for the remainder of this contribution, discussion typically will be restricted to auditory perception of speech. This should not be taken to imply that other modalities do not contribute to understanding speech. The approach outlined here is explicitly associationist and driven by experience. The brain is opportunistic. Sensori-neural systems seize upon redundancies to maximize information transmission whenever possible. Contemporary research (e.g., Bahrick, L.E. *et al.*, 2004) reveals how intersensory redundancy guides development of perception most generally. Whenever nonauditory information is redundant with speech acoustics, those correlations should contribute to efficient encoding of speech. For example, listeners have a wealth of experience simultaneously hearing speech and viewing talkers' faces, and the McGurk effect (McGurk, H. and MacDonald, J., 1976) is evidence of the profound effects visual information can have on the way speech sounds are perceived. Also, whenever people are

talking, they both hear the sounds they're producing and they experience the movements in their own vocal tracts. While ideally, this occurs less than half the time people hear speech, simultaneous activities of both hearing and talking provide exquisite conditions for extraction of correlations.

1.59.6 Experience and Sound Contrasts in the Native Language

Experience is essential for the development of every sensori-neural system. The profound role of experience is especially clear for speech perception. Infants gain significant experience with speech even before they're born. Late-term fetuses can discriminate vowel sounds (Lecanuet, J. P. *et al.*, 1986), and prenatal experience with speech sounds appears to have considerable influence on subsequent perception, as newborns prefer hearing their mother's voice (DeCasper, A., and Fifer, W., 1980). By the time French infants are 4 days old, they discriminate French from other languages (e.g., Russian) (Mehler, J. *et al.*, 1988). Perhaps most telling is the finding that newborns prefer hearing particular children's stories that were read aloud by their mothers during the third trimester (DeCasper, A. J. and Spence, M. J., 1986).

Experience plays a critical role in tuning speech perception to the distributions of sounds within one's language environment. Much, if not most, of this development as a native listener takes place during the first year of life. One of the most challenging obstacles for the infant is to use acoustic information that distinguishes one speech sound from another in the face of sometimes widely varying acoustic properties, many of which do not distinguish speech sounds in their language. Acoustic differences that convey a contrast in one language may be of little or no relevance to another language. Some of these differences simply may be unrelated (orthogonal) to distinctions used in a particular language and would not occur in a language context. In addition, clearly audible differences such as gender of talker, speaking rate, emotional state, and other factors have profound effects on the acoustic signal, yet the language learner must learn to understand speech across these variations.

1.59.6.1 Vowels

At least by the age of 6 months, infants have the ability to distinguish stimuli by vowel type even when different instances of the vowel differ considerably between presentations (Kuhl, P. K., 1983). In a reinforced head turn paradigm, Kuhl trained infants to turn their heads only when the vowel of the background stimulus changed during presentation of the closely related vowels [a] (as in tot) and [ɔ] (as in taught) spoken by a male talker. When tested on novel vowels produced by women and children (adding random variation in pitch contour in addition to shifting absolute frequencies of formants), infants provided the correct response on the first trial demonstrating that they recognized the novel instances as consistent with training vowels despite talker changes. Note that, by the shared covariance account offered above, the capacity to distinguish vowels across variation in irrelevant acoustic characteristics is a natural consequence of encoding stimuli on the basis of attributes that tend to co-occur. Attributes such as those accompanying changes in talker are irrelevant to particular consonants and vowels, so they do not play much role in phonetic distinctions.

While these studies attest to the ability of infants to respond to distinctions between vowels in the face of irrelevant variation, later studies have investigated how perception may be structured along acoustic/auditory dimensions that are directly relevant to distinguishing vowel sounds. What has become most apparent is that the degree to which infants treat acoustically different instances of the same vowel as equivalent is critically dependent upon their experience with a particular language. For example,6-month-old infants detect differences between vowel sounds differently depending upon whether they lived in an English-speaking (Seattle) or Swedish-speaking (Stockholm) home (Kuhl, P. K., *et al.*, 1992).

Further evidence for the role of experience can be found in experiments in which performance by European starlings (*Sturnus vulgaris*), having learned statistically controlled distributions of renditions of Swedish and English vowels, was highly correlated with performance of adult human listeners (Kluender, K. R. *et al.*,1998). A simple linear association network model, exposed to the same vowels heard by the birds, accounted for 95% of the variance in avian responses. Consistent with the principle that consonants and vowels are defined mostly be what sounds they are not, both human goodness judgments (Lively, S. E., 1993) and starling response rates illustrate an anisotropy such that peak responses are skewed away from competing vowel sounds more than they are defined by centroids of vowel distributions.

1.59.6.2 Consonants

Perception of differences between consonants is similarly tuned by experience. Werkerand her colleagues (Werker, J. F. *et al.*, 1981; Werker, J. F. and Logan, J. S., 1985; Werker, J. F. and Lalonde, C. E., 1988; Werker, J. F. and Tees, R. C., 1983, 1984a; 1984b) have demonstrated that, as a consequence of experience with consonants in their native language, infants' tendency to respond to differences between some consonants that are not in their language begins to attenuate. The series of studies by Werker. J. F. and Lalonde, C. E. (1988) permit a relatively complete description of the phenomena. They exploited the fact that speakers of English and Hindi use place of articulation somewhat differently for stop consonants. While for English, three places of articulation are used for voiced stop consonants: labial, alveolar, and velar (e.g. /b/, /d/, and /g/, respectively), in Hindi four places are used: labial, dental, retroflex, and velar (e.g. /b/, /d̲/, /d/, and /g/, respectively.) They created a synthetic series that varied perceptually from /b/ to /d/ (for native-English speaking adults) and from /b/ to /d̲/ to /d/ (for native-Hindi speaking adults) (Figure 11).

Using the same reinforced head turn procedure used by Kuhl, they found that 6- to 8-month-old infants from English-speaking families responded to changes in stimulus tokens that crossed perceptually from the English /b/ to /d/ and also responded to changes between Hindi stops [d̲] and [d]. A different group of infants from English speaking families aged 11- to 13-months of age responded reliably only to the English [b]-[d] contrast, and not to the Hindi [d̲]-[d] contrast. Essentially, 6- to 8-month-old infants responded in a manner typical of native-Hindi adults, while 11- to 13-month-olds responded like native-English adults treating both dental and retroflex stops as being the same. Werker and her colleagues have found analogous results in studies using different consonant contrasts from different languages.

For vowels and consonants, perception of speech is shaped during the first year of life in ways that respect the statistics of the linguistic environment. The challenge for the infant learning the speech sound distinctions in his or her native language is precisely this. Infants must learn how acoustic/auditory attributes tend to co-occur, and those correlations among attributes define perceptual organization that optimizes sensitivity to change.

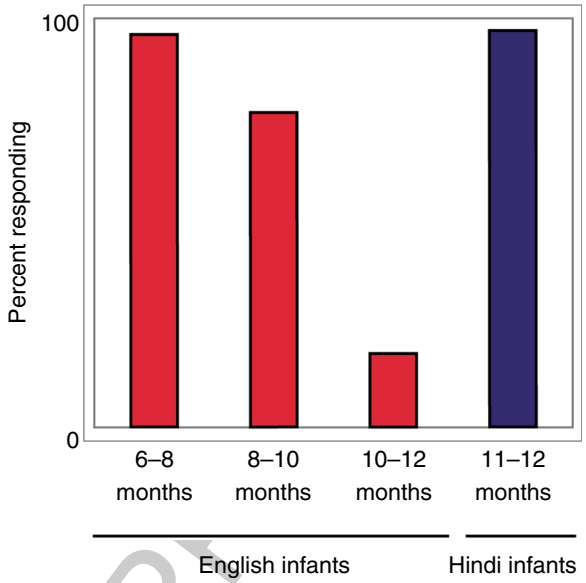

Figure 11 Perception of speech sounds is tuned by experience. As a consequence of experience with consonants in their native language, infants' tendency to respond to differences between some consonants that are not in their language begins to attenuate. Here, 6- to 8-month-old infants from English-speaking homes respond in a manner typical of native-Hindi adults and 1–12-month-old Hindi infants when hearing dental and retroflex Hindi stops. Before they are a year old, infants from English environments respond like native-English adults, treating both consonants the same. Adapted from Werker, J. F. and Lalonde, C. E., 1988 Cross-language speech perception: initial capabilities and developmental change. Dev. Psychol. 24, 672–683.

1.59.6.3 Second-Language Perception

The same principles that explain categorical perception and development of perceptual organization during the first year of life extend to predicting how difficult, or easy, it is to learn new speech contrasts in a second language. For the case of the sounds of a single language, correlated attributes distinguish each consonant or vowel from others in a fashion that optimizes sensitivity to differences. The same construct, habitual co-occurrence of acoustic attributes, constrains and predicts how listeners perceive familiar and unfamiliar sounds from a second language. There are three basic patterns of interaction between perceptual organization for the native language and the mapping of sounds from a second language.

First, acoustic attributes of two contrasting nonnative speech sounds can be equally well correlated with attributes corresponding to only a single native consonant or vowel. Consider the case for formant patterns contributing to categorization of stop consonants with varying place of articulation. For

example, both dental and retroflex stops such as those found in Hindi are acoustically realized in a manner quite similar to that for English alveolar stops. Given the range of ways English [d] is produced in the contexts of other speech sounds, there is ample overlap with acoustic attributes associated with both dental and retroflex stops. When [d] is produced in the environment of the retroflex continuant [r] as in drew, English [d] is shares multiple acoustic commonalities with the Hindi retroflex stop [d̪]. Similarly, when English [d] is produced in the environment of a dental fricative such as [θ] in words like width, it is acoustically quite similar to the Hindi dental stop [d̪] (Polka, L., 1991). Given the facts about the distributions of acoustic attributes for alveolar stops in fluent English, attributes consistent with dental or retroflex stops are well correlated with attributes that co-occur in alveolar stops. Dental-like or retroflex-like acoustic attributes are accommodated within correlation structures for English alveolar stops via an increase in overall variance reflective of the observed variability in English alveolar stop production. Werker, J. F. and Lalonde, C. E.'s (1988) adult identification data are entirely consistent with this scenario. Stimuli that are identified by native-Hindi listeners as dental or retroflex are all assimilated into the set of stimuli identified as alveolar by native-English listeners. Best, C. T. *et al.* (1988) referred to a similar process as single-category assimilation in her taxonomy of contrasts within an articulatory framework.

An analogous example of difficulty perceiving a distinction between two sounds is the well-known inability of native-Japanese listeners to detect the distinction between English [r] and [l] (e.g., Miyawaki, K. *et al.*, 1975). Japanese sounds include a consonant called a rhotic flap, and acoustic characteristics of these flaps overlap extensively with those of both English [r] and [l] (Sato, M. *et al.*, 2003). Consequently, acoustic attributes of [r] and [l] are equally well correlated with attributes corresponding to a single Japanese sound, and native-Japanese listeners are unable to hear, and consequently produce, the English distinction.

A related way non-native contrasts can be assimilated with native contrasts involves cases for which attributes of one sound from of a non-native contrast are very well correlated with attributes of a single native consonant or vowel, being effectively the same. Attributes of a second non-native sound fit less well with the correlation structure of the same native sound, but they do correspond better with that sound than with any other native sound. One example of this is the Farsi distinction between velar and uvular stops. Native-English listeners do not lose the ability to discriminate Farsi velars from uvulars. Instead, they perceive the Farsi voiced velar and uvular stops as being good and poor instances, respectively, of English /g/ (Polka, L., 1992). In this case, Farsi velar stops are perceived as good English velar stops because they share most or all of the acoustic/auditory attributes that comprise the correlated structure of English /g/. Farsi uvular stops share fewer attributes with those that co-occur for English [g]. Farsi uvulars somewhat, but not completely, fit the correlation structure of English [g]. A related process has been referred to as category-goodness assimilation by Best, C. T. *et al.* (1988).

The third way native and non-native contrasts can interact can be found in cases where the native language does not exactly share a contrast with a non-native language, but the native language does have a similar contrast that facilitates perception of the non-native contrast. For example, French does not include a voicing distinction for dental fricatives such as /ð/-/θ/ (as in than and thank), yet native-French listeners can discriminate voiced from voiceless English fricatives by perceiving them as versions of French dental stops /d̪/ and /t̪/, respectively (Jamieson, D. G. and Morosan, D. E., 1986). Best, C. T. *et al.* (1988) label this type of assimilation two-category because each sound of the non-native contrast maps more or less on to a different sound in the native language. Within the framework of correlated attributes, one would explain the fact that French listeners perceive the English fricatives as versions of French stops is because attributes of the dental fricatives are reasonably well correlated with attributes of the French dental stops as produced with typical allophonic variation in fluent speech.

This scenario leaves only those non-native distinctions that are roughly orthogonal to contrasts within the native language. Across the broad domain of possible mouth sounds, there always will remain some attributes that are not well correlated with any of attributes for any speech sounds within a given language. For example, attributes of the click sounds of Zulu are correlated with no sounds of English, and their perception should be, and is relatively unaltered by the process of learning English phonemes (Best, C. T. *et al.*, 1988). It should be noted that all of the patterns of performance above are consistent with Flege, J. E.'s (1995) speech learning model (SLM) and with patterns of experience-dependent learning of speech contrasts (e.g., Flege, J. E. *et al.*, 1997; Imai, S.

et al., 2002). However, Flege's explanations of underlying processes that give rise to these patterns of performance are distinct at least in level of analysis, and he may or may not agree with specific aspects of the present authors' explanations.

1.59.7 To the Lexicon and Beyond

1.59.7.1 Lexical Development and the Emergence of Phonemes (or Something like Them)

In the introduction to this chapter, discussion of phonetic segments and phonemes as independent entities was decidedly circumspect. Throughout the foregoing, consonants and vowels have been described either as sounds or as correlations among acoustic attributes in the service of maximizing information transmission. They have not been described as inherently linguistic or as a discrete stage in processing. To borrow Angell's (Angell, J. R., 1907) dichotomy between functionalism and structuralism, discussion has been more about how and why, and less about the structuralist what of linguistic theory. Particular emphasis has been about how, and the focus now turns to why. The why of speech perception is to recognize words, and the end goal must be getting from the acoustic signal to words that have meaning. Within the information-theoretic perspective adopted here, one can construe the process of speech perception as one of successively reducing uncertainty sufficiently to arrive at words.

Over the years, some researchers have made a case that speech perception really is word perception without intermediate levels of analysis. There have been simulations of lexicons constructed directly from acoustic/auditory input *sans* phonemes (e.g., Klatt, D. H., 1980; Johnson, 2000), and a number of investigators have argued for the primacy of holistic (word-size) organization in lexical development (e.g., Charles-Luce, J. and Luce, P. A., 1990; 1995; Jusczyk, P. W., 1986; Walley, A. C., 1993; Walley, A. C. *et al.*, 2003). For example, Charles-Luce and Luce (Charles-Luce, J. and Luce, P. A., 1990) argued that emergence of acoustic/auditory detail, such as consonants and vowels of words, within the lexicon is a consequence of – not antecedent to – learning more words. By such an account, as the number of words in the lexicon grows, increasing degrees of detail are required to sort each word from all the others.

Kluender, K. R. and Lotto, A. J. (1999), by contrast, suggested that neither words nor phonetic units may serve exclusively to structure the developing lexicon. Werker and Curtin (Werker, J. F. and Curtin, S., 2005), within their developmental framework for processing rich information from multidimensional interactive representation (PRIMIR), argue that infants have access to information for both phonetic units and the ways these units are grouped together into words. While PRIMIR is more of a framework than a detailed model, Werker and Curtin suggest that internal representations of phonemes become more firmly established and resistant to change as the lexicon grows.

The present emphasis of maximizing information transmission is consistent with concurrent development of statistical structures corresponding to consonants and vowels both preliminary to the lexicon and as an emergent property of lexical organization. In this chapter, speech perception has been described as a succession of processes operating upon the acoustic signal with varying levels of complexity. Common to all these processes is maximizing efficiency of information transmission by absorbing redundancies across inputs. For example, spectral contrast operates early, with trading relations and categorical perception operating later. Following these preliminary operations, one easily can imagine development of a nascent, but undetailed, lexicon of word forms. As a lexical space becomes increasingly populated, the covariance space becomes more complex. Predictable relationships among attributes, now with a phonetic segment-like grain, can be revealed resulting in a reduction in dimensionality of lexical items. The classic definition of phonemes is that they provide the minimal distinction between meaningful elements (morphemes) of language (Trubetzkoy, N. S., 1939/1969). Consequently, phonemes provide one of the most efficient ways to describe differences within the lexicon. Because phonemes provide efficient descriptors of lexical space, they emerge as dimensions of the developing lexical space by quite the same process that explains categorical perception, now operating over a larger time window.

One might suggest that positing phoneme-like dimensions as emergent properties of a lexical space obviates the need for anything resembling consonants and vowels preliminary to the lexicon. However, doing so violates principles of sensori-neural organization, these being that redundancies are extracted continuously with ascending levels of processing. Werker and Curtin's (Werker, J. F. and Curtin, S., 2005) proposal that phonemes become more established, and presumably increasingly tuned to

phonotactic regularities in a language, is more consistent with persistent successive absorption of redundancy in the service of maximizing sensitivity to change.

1.59.7.2 Finding Word Boundaries

One final example of auditory perception using predictability to enhance sensitivity to change is found in studies demonstrating how infants find boundaries between words. In connected speech, acoustic realization of the beginning and end of one word also overlaps with sounds of preceding and following words. Unlike white spaces between words on a page, there are no silent intervals that mark beginnings and ends of words. Interestingly, perception is at odds with this acoustic reality. When listening to someone talking, most individual words stand out quite clearly as discrete entities. But listening to someone speak in a different language is often a very different experience. Every phrase or sentence may sound like a single very long word. This is the situation faced by infants.

Saffran and colleagues (Saffran, J. R. *et al.*,1996) demonstrated that infants can use transitional probabilities between successive sounds within a speech stream as evidence for breaks between words. In their studies, they used streams of connected pseudowords, for which the probability of some sequences of consonant-vowels (CVs) was very high (1.0) while probability of other sequences was relatively low (0.33). Infants were sensitive to whether two sounds share a history of co-occurrence. When they heard successive CVs that rarely co-occurred with one another in their experience, they recognized this as a sign that there was a break between words. This discontinuity corresponds to a spike in information because one CV did not predict the occurrence of the next. Statistics of English support this emphasis upon word boundaries, as the ends of most words cannot be identified prior to the onset of the next (Luce, R. D., 1986). Infant sensitivity to boundaries is yet another example of using predictability to enhance sensitivity to change, and hence enhance transmission of information (Figure 12).

Because this is a principle of perceptual systems most broadly, one expects this use of predictability to apply most generally. Indeed, these patterns of performance extend to infants experiencing tonal sequences (Saffran, J. R.et *al.*, 1999), visual shapes (Kirkham N. Z.et *al.*, 2002), and visual-feature combinations (Fiser, J. and Aslin, R. N., 2002). In fact, even nonhuman primates (Hauser, M. D. *et al.*, 2001) exhibit this sensitivity to transitional probabilities.

1.59.8 Speech in the Brain

Given the plethora of relatively recent studies concerning speech processing using electroencephalography (EEG), magnetoencecphalography (MEG), positron emission tomography (PET), and functional magnetic resonance imaging (fMRI), extensive depiction of speech processing in cerebral cortex alone requires one or more chapters. Here, a very brief review of current understanding will be placed within the framework of information processing provided above.

Hearing sounds of any kind activates primary auditory cortex (AI). Processing of complex sounds relies upon additional areas of cortex adjacent to AI, which is functionally divided into ventral, anterior,

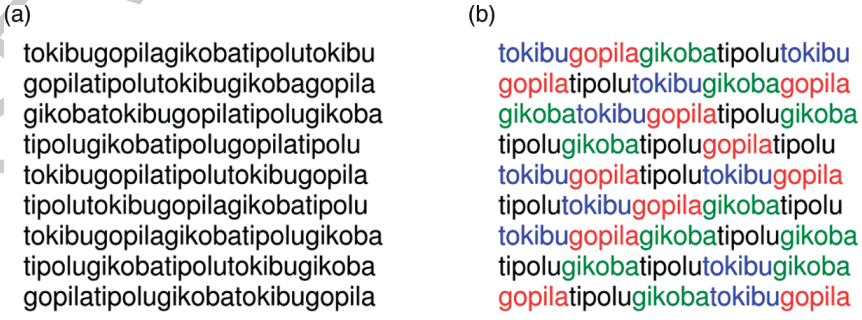

Figure 12 There are no acoustic markers between most words in a stream of fluent speech, analogous to the strings of letters on the right (a). Following minutes of experience with streams of connected speech in which probability of some sequences of consonant-vowels (CVs) is very high, while probability of other sequences was relatively low, infants are sensitive to whether two sounds share a history of co-occurrence. From Saffran, J. R. Aslin, R. N., and Newport, E. L. 1996. Statistical learning by 8-month-old infants. Science 274, 1926–1928.

and posterior sections. Neurons from AI project to the surrounding belt of cortex, and neurons from this belt synapse with neurons in the adjacent parabelt area. Just about any sound will cause activation in some part of AI. However, in the belt and parabelt areas, referred to as secondary or associational auditory areas, simple sounds such as sine waves and white noise elicit less activity, particularly if they have limited temporal structure. Thus, as in the visual system, processing proceeds from simpler to more complex stimuli farther along the auditory pathway, and there is also greater evidence of cross-modal processing (e.g., combining acoustic and optic information), particularly in parabelt areas. Of course, this general property of hierarchical organization is consistent with continuous and successive extraction of redundancy across increasing spans of space and time. As one might expect, areas beyond AI are activated when listeners hear speech and music. Further, at these early levels of cortical processing, activity in response to music, speech, and other complex sounds is relatively balanced across the two hemispheres.

When listening to speech, additional areas of both left and right superior temporal lobes along the superior temporal sulcus (STS) activate more strongly in response to speech than to nonspeech sounds such as tones and noise (e.g., Binder, J.R., 2000). While language processing is typically lateralized to one hemisphere, this activity in response to speech signals is relatively balanced across both sides of the brain when researchers have been very careful to avoid higher-level effects of language (Zatorre, R. J. and Binder, J., 2000).

At some point, however, processing of speech should become more lateralized as perceiving speech becomes part of understanding language. However, one challenge for researchers has been to create control stimuli that have all the complex properties of speech without being heard as speech. Because listeners are very good at understanding even severely distorted speech, it is very difficult to construct stimuli that are complex like speech without being heard as speech.

Liebenthal and colleagues (Liebenthal, E. *et al.*, 2005) adopted a creative way to control for acoustic complexity while varying whether sounds would be heard as speech. They synthesized speech syllables varying incrementally from [ba] to [da]. Nonspeech control stimuli were the same series of syllables, except characteristics of F_1 transitions were flipped upside down, decreasing in center frequency following syllable onset. It is impossible for a human vocal tract to create such sounds, and listeners could not identify them as consonants.

During scanning, listeners participated in a categorical perception task, discriminating pairs of stimuli from one or the other series of stimuli. For the [ba-da] speech series, performance was typical for categorical perception experiments, with stimuli that would be labeled differently (e.g., /ba/ versus /da/) being almost perfect, while discrimination of other stimulus pairs was rather poor. For the nonspeech sounds, discrimination was above chance and pretty much the same for all pairs of stimuli. Listening to both series of stimuli resulted in increased activation in STS in both temporal lobes. When sounds were speech [ba-da] however, there was increased activation in STS superior temporal cortex, a bit anterior and ventral from activation for control stimuli and mostly in the left hemisphere. This research (Liebenthal, E. *et al.*, 2005) may have revealed a place in the left temporal lobe where experience with English [b] and [d] shapes neural activity.

This experience-dependent cortical organization for English consonants may have been predicted by previous work with visual processing of well-learned objects. In the visual system, perception of faces results in activation of the middle fusiform gyrus (or fusiform face area) of the cortex. This area is disproportionately activated when subjects view faces versus viewing other objects or scenes. Similar patterns of activation, in the same general brain region, can be found in response to other types of stimuli with which subjects are very familiar (Gauthier, I *et al.*, 1999).

A study by Scott and colleagues (Scott, S. K. *et al.*, 2000) suggest a cortical locus for the next step of processing of speech into language. They controlled for acoustic complexity while changing whether sentences were intelligible by using spectrally rotated sentences that could not be understood by listeners. Rotated speech signals were spectrographically upside down (i.e., the frequency scale was inverted.). They could not be understood by listeners, but they were as complex acoustically as right-side-up sentences. Rotated stimuli activated STS comparably to intact nonrotated sentences, suggesting that auditory processing in these areas is related more to complexity than to being speech per se. The essential difference between cortical responses to these two types of stimuli was that, on the left temporal lobe, activation in response to intact sentences continued

further anterior and ventral (including superior temporal gyrus, STG) to the region activated by rotated sentences. Activation was a little anterior and ventral to the area in which Liebenthal and colleagues (Liebenthal, E. *et al.,* 2005) found activation for [b] and [d]. Relative left lateralization consequent to recognizability parallels findings for other environmental sounds (e.g., Lewis, J. W. *et al.*, 2004), albeit in different brain regions (Figure 13).

The extension of lateralized activation, increasingly anterior and ventral, discovered by Scott and her colleagues (Scott, S. K. *et al.,* 2000) inspires a tantalizing hypothesis concerning the next level of processing beyond speech perception to word recognition. Based upon neuropsychological data from patients, cortical areas essential to semantic memory may reside within far anterior temporal lobe (see, e.g., Rogers, T. T. *et al.*, 2004). Consider a scenario through which increasingly sophisticated redundancies are wrenched out of the speech signal, through differentiation of sounds such as [b] and [d], through detection of word boundaries, to regularities within word boundaries (i.e., words) and concomitant associations with semantic properties of words. Such a notion is clearly speculative given the present state of knowledge; however, this would be an elegant view of successive processing along increasingly anterior and ventral areas of temporal lobe.

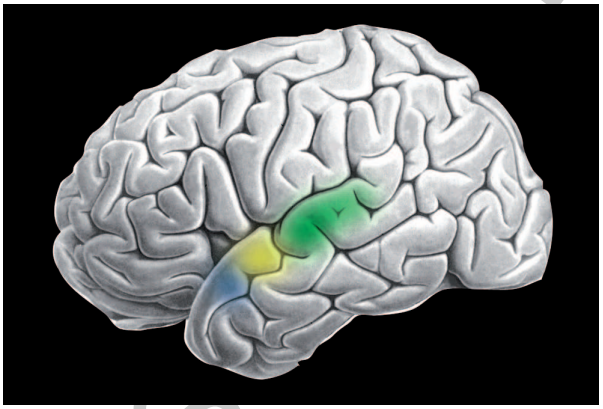

Figure 13 When listening to speech, as well as other complex sounds, cortical activity along the superior temporal sulcus (STS, green) is relatively balanced across the hemispheres. When acoustical complexity is carefully controlled, there is evidence of increased activation in STS (yellow), more anterior and mostly in the left hemisphere (Liebenthal, E. *et al.*, 2005). Cortical activation in response to recognizable sentences appears further anterior and ventral (blue), including superior temporal gyrus (STG). From Scott, S. K. Scott, S. K., Blank, C. C., Rosen, S., and Wise, R. J. S. 2000. Identification of a pathway for intelligible speech in the left temporal lobe. Brain 123, 2400–2406.

1.59.9 Conclusion

Speech perception is grounded in general principles that apply to other acoustic events and to other modalities. Classic principles that guide perception, none of which is wholly original to the authors, explain processes underlying multiple phenomena of speech perception. This information-theoretic model, operating from sensory transduction to word learning, is biologically realistic. It is intended that this framework will serve, not only to reveal processes underlying normative aspects of speech perception, but also to extend understanding of clinical conditions of speech and language processing. Finally, beyond being amenable to study like any other form of perception, speech perception holds promise as a fertile domain for research that can reveal and extend fundamental understanding of perception most generally.

Acknowledgments

Comments provided by Randy Diehl, Ray Kent, Michael Kiefte, Timothy Rogers, and Christian Stilp on part or of all of this chapter are greatly appreciated, but the authors have not accepted all their advice and are alone responsible for any shortcomings of editing or content. Laurel Steinmeyer's and Chrsitian Stilp's assistance with figures is also much appreciated. Chapter written in January 2006. Work supported by NIDCD.

References

Alexander, J. M. and Kluender, K. R. 2005. Contributions of gross spectral properties to perception of stop consonants. J. Acoust. Soc. Am. 118(3) pt 2.

Anderson, J. A., Silverstein, J. W., Ritz, S. A., and Jones, R. S. 1977. Distinctive features, categorical perception, and probability learning: some applications of a neural model. Psychol. Rev. 84, 413–451.

Anderson, B. L. and Winawer, J. 2005. Image segmentation and lightness perception. Nature 434, 79–83.

Angell, J. R. 1907. The province of functional psychology. Psychol. Rev. 14, 61–91.

Attneave, F. 1954. Some informational aspects of visual perception. Psychol. Rev. 61, 183–193.

Attneave, F. 1959. Applications of Information Theory to Psychology: A Summary of Basic Concepts, Methods, and Results. Henry Holt and Company, Inc.

Bahrick, L. E., Lickliter, R., and Flom, R. 2004. Intersensory redundancy guides development of selective attation, perception, and cognition in infancy. Curr. Dir. in Psychol. Sci. 13, 99–102.

Barbour, D. L. and Wang, X. 2003. Contrast tuning in auditory cortex. Science 299, 1073–1075.
Barlow, H. B. 1961. Possible Principles Underlying the Transformations of Sensory Messages. In: Sensory Communication (*ed*. W. A. Rosenblith), pp. 53–85. MIT Press.
Barlow, H. B. 1997. The knowledge used in vision and where it comes from. Philos. Trans. R. Soc. Lond. B 352, 1141–1147.
Barlow, H. B. 2001. The exploitation of regularities in the environment by the brain. Behav. Brain Sci. 24, 602–607.
Beale, J. M. and Keil, F. C. 1995. Categorical effects in the perception of faces. Cognition 57, 217–239.
Best, C. T., McRoberts, G. W., and Sithole, N. M. 1988. Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. J. Exp. Psychol. 14, 345–360.
Binder, J. R. 2000. The neuroanatomy of speech perception (Editorial). Brain 123, 2371–2372.
Bladon, R. A. W. and Lindblom, B. 1981. Modeling the judgment of vowel quality differences 1981. J. Acuost. Soc. Am. 69, 1414–1422.
Burns, E. M. and Ward, J. D. 1974. Categorical perception of musical intervals. J. Acoust. Soc. Am. 55, 456.
Burns, E. M. and Ward, W. D. 1978. Categorical perception – phonemenon or epiphenomenon: evidence from experiments in the perception of melodic musical intervals. J. Acoust. Soc. Am. 63, 456–468.
Calder, A. J., Young, A. W., Perrett, D. I., Etcoff, N. L., and Rowland, D. 1996. Categorical perception of morphed facial expressions. Vis. Cogn. 3, 81–117.
Campbell, R., Pascalis, O., Coleman, M., Wallace, S. B., and Benson, P. J. 1997. Are faces of different species perceived categorically by human observers? Proc. R. Soc. Lond. B. Biol. Sci. 264, 1429–1434.
Cardozo, B. L. 1967. Ohm's law and masking. *IPO Annual Progress Report, Institute for Perception Research 2*, 59–64.
Cathcart, E. P. and Dawson, S. 1928/29. Persistence 2. Br. J. Psychol. 19, 343–356.
Charles-Luce, J. and Luce, P. A. 1990. Similarity neighbourhoods of words in young children's lexicons. J. Child Lang. 17, 205–215.
Charles-Luce, J. and Luce, P. A. 1995. An examination of similarity neighbourhoods in young children's receptive vocabularies. J. Child Lang. 22, 727–735.
Christman, R. J. 1954. Shifts in pitch as a function of prolonged stimulation with pure tones. Am. J. Psychol. 67, 484–491.
Cleveland, J. and Snowdon, C. T. 1982. The complex vocal repertoire of the adult cotton-top tamarin (*Saguinus oedipus oedipus*). Zeitschrift fur Tierpsychologie 58, 231–270.
Coady, J. A., Kluender, K. R., and Rhode, W. S. 2003. Effects of contrast between onsets of speech and other complex spectra. J. Acoust. Soc. Am. 114(4), 2225–2235.
Damper, R. I., Gunn, S. R., and Gore, M. O. 2000. Extracting phonetic knowledge from learning systems: perceptrons, support vector machines and linear discriminants. Appl. Intell. 12, 43–62.
Damper, R. I. and Harnad, S. R. 2000. Neural network models of categorical perception. Percept. Psychophys. 62, 843–867.
DeCasper, A. and Fifer, W. 1980. Of human bonding: newborns prefer their mothers' voices. Science 208, 1174–1176.
DeCasper, A. J. and Spence, M. J. 1986. Prenatal maternal speech influences newborns' perception of speech sounds. Infant Behav. Dev. 9, 133–150.
De Gelder, B., Teunisse, J. P., and Benson, P. J. 1997. Categorical perception of facial expressions: categories and their internal structure. Cogn. Emot. 11, 1–23.
Delgutte, B. 1980. Representation of speech-like sounds in the discharge patterns of auditory nerve fibers. J. Acoust. Soc. Am. 68, 843–857.
Delgutte, B. 1986. Analysis of French Stop Consonants with a Model of the Peripheral Auditory System. In: Invariance and Variability of Speech Processes (*eds*. J. S. Perkell and D. H. Klatt), pp. 131–177. Lawrence Earlbaum Associates.
Delgutte, B. 1996. Auditory Neural Processing of Speech. In: The Handbook of Phonetic Sciences (*eds*. W. J. Hardcastle and J. Laver), pp. 507–538. Blackwell.
Delgutte, B., Hammond, B. M., Kalluri, S., Litvak, L. M., and Cariani, P. A. 1996. Neural Encoding of Temporal Envelope and Temporal Interactions in Speech. In: Auditory Basis of Speech Perception (*eds*. W. Ainsworth and S. Greenberg), pp. 1–9. European Speech Communication Association.
Delgutte, B. and Kiang, N. Y. S. 1984. Speech coding in the auditory nerve IV: sounds with consonant-like dynamic characteristics. J. Acoust. Soc. Am. 75, 897–907.
Diehl, R. L. and Kluender, K. R. 1989. On the objects of speech perception. Ecol. Psychol. 1(2), 121–144.
Dillon, W. R. and Goldstein, M. 1984. Multivariate Analysis, Methods and Applications, John E Wiley and Sons.
Escabi, M. A., Miller, L. M., Read, H. L., and Schreiner, C. E. 2003. Naturalistic auditory contrast improves spectrotemporal coding in the cat inferior colliculus. J. Neurosci. 23, 11489–11504.
Etcoff, N. L. and Magee, J. J. 1992. Categorical perception of facial expressions. Cognition 44, 227–240.
Festen, J. M. and Plomp, R. 1981. Relations between auditory functions in normal hearing. J. Acoust. Soc. Am. 70, 356–369.
Fiser, J. and Aslin, R. N. 2002. Statistical learning of new visual feature combinations by infants. Proc. Natl. Acad. Sci. 99, 15822–15826.
Flege, J. E. 1995. Second language speech learning: Theory, Findings and Problems. In: Speech Perception and Linguistic Experience: Issues in Cross-language Research (*eds*. W. Strange York and M. D. Timonium), pp. 233–272.
Flege, J. E., Bohn, O. S., and Jang, S. 1997. Effects of experience on non-native speakers' production and perception of English vowels. J. Phon. 25(4), 437–470.
Fletcher, H. 1953/1995. Speech and Hearing in Communication. Krieger.
Fowler, C. A. 1986. An event approach to the study of speech perception from a direct-realist perspective. J. Phon. 14, 3–28.
Fowler, C. A., Best, C. T., and McRoberts, G. W. 1990. Young infants' perception of liquid coarticulatory influences on following stop consonants. Percept. Psychophys. 48, 559–570.
Freedman, D. J., Riesenhuber, M., Poggio, T., and Miller, E. K. 2001. Categorical perception of visual stimuli in the primate prefrontal cortex. Science 291, 312–316.
Gardner, W. R. and Hake, H. W. 1951. The amount of information in absolute judgments. Psycho. Rev. 58, 446–459.
Gauthier, I., Tarr, M. J., Anderson, A. W., Skudlarski, P., and Gore, J. C. 1999. Activation of the middle fusiform 'face area' increases with expertise in recognizing novel objects. Nat. Neurosci. 2(6), 568–573.
Goldstone, R. L. 1994c. Influences of categorization on perceptual discrimination. J. Exp. Psychol. General 123, 178–200.
Gordon, C., Webb, D. L., and Wolpert, S. 1992. One cannot hear the shape of a drum. Bull. Am. Math. Soc. 27, 134–138.
Green, D. M., McKey, M. J., and Licklider, J. C. R. 1959. Detection of a pulsed sinusoid in noise as a function of frequency. J. Acoust. Soc. Am., 31, 1146–1152.
Hauser, M. D., Newport, E. L., and Aslin, R. N. 2001. Segmentation of the speech stream in a non-human primate: statistical learning in cotton-top tamarins. Cognition 78, B53–B64.

Herrnstein, R. J. 1984. Objects, categories and discriminative stimuli. In: Animal Cognition, (*eds*. H. L. Roitblat, T. G. Bever, and H. S. Terrace), pp. 233–261. Erlbaum.
Hoagland, H. 1933. Quantitative aspects of cutaneous sensory adaptation I. J. Gen. Physiol. 16, 911–923.
Holt, L. L. 2005. Temporally non-adjacent non-linguistic sounds affect speech categorization. Psychol. Sci. 16, 305–312.
Holt, L. L., Lotto, A. J., and Kluender, K. R. 2000. Neighboring spectral content influences vowel identification. J. Acoust. Soc. Am. 108(2), 710–722.
Hood, J. D. 1950. Studies in auditory fatigue and adaptation. Acta Oto-Laryngol. Suppl. 92, 1–57.
Houtgast, T. 1972. Psychophysical evidence for lateral inhibition in hearing. J. Acoust. Soc. Am. 51, 1885–1894.
Houtgast, T. 1974. Auditory analysis of vowel-like sounds. Acustica 31, 320–324.
Hume, D. 1748/1963. An Enquiry Concerning Human Understanding. The Harvard Classics. P. F. Collier and Son Corporation. (Original published in 1748 entitled My own life.)
Hyvärinen, A. and Oja, E. 2000. Independent component analysis: algorithms and applications. Neural Netw. 13, 411–430.
Imai, S., Flege, J., and Wayland, R. 2002. Perception of cross-language vowel differences: a longitudinal study of native Spanish learners of English. Acoustical Society of America Journal 111(5), 2364–2364.
Jakobson, R. and Halle, M. 1971. The Fundamentals of Language. Mouton.
Jamieson, D. G. and Morosan, D. E. 1986. Training non-native speech contrasts in adults: acquisition of English /q/-/d/ contrast by francophones. Percept. Psychophys. 40, 205–215.
Jusczyk, P. W. 1981. Infant Speech Perception: A Critical Appraisal. In: Perspectives on the Study of Speech (*eds*. P. D. Eimas and J. LMiller), pp. 113–64. Erlbaum.
Jusczyk, P. W. 1986. Towards a Model for the Development of Speech Perception. In: Invariance and Variability in Speech Processes (*eds*. J. Perkell and D. H. Klatt), pp. 1–19. Erlbaum.
Kiefte, M. and Kluender, K. R. 2005. The relative importance of spectral tilt in monophthongs and diphthongs. J. Acoust. Soc. Am. 117(3), 1395–1404.
Kiefte, M. and Kluender, K. R. (in press). Cancellation of reliable spectral characteristics in auditory perception. J. Acoust. Soc. Am.
Kingston, J. and Diehl, R. L. 1994. Phonetic knowledge. Language 70, 419–454.
Kirk, E. C. and Smith, D. W. 2003. Protection from acoustic trauma is not a primary function of the medial olivocochlear efferent system. J. Assoc. Res. Otolaryngol. 4, 445–465.
Kirkham, N. Z., Slemmer, J. A., and Johnson, S. P. 2002. Visual statistical learning in infancy: evidence for a domain general learning mechanism. Cognition 83, B35–B42.
Klatt, D. H. 1980. Software for a cascade/parallel formant synthesizer. J. Acoust. Soc. Am. 67(3), 971–995.
Klatt, D. H. 1982. Prediction of perceived phonetic distance from critical band spectra: a first step. Proc. ICASSP 1278–81.
Kluender, K. R. 1988. Auditory constraints on phonetic categorization: Trading relations in humans and nonhumans. Unpublished Ph.D. dissertation, University of Texas at Austin.
Kluender, K. R. 1994. Speech Perception as a Tractable Problem in Cognitive Science. In: Handbook of Psycholinguistics (*ed*. M. A. Gernsbacher), pp. 173–217. Academic.
Kluender, K. R., Diehl, R. L., and Killeen, P. R. 1987. Japanese quail can learn phonetic categories. Science 237, 1195–1197.
Kluender, K. R. and Lotto, A. J. 1999. Virtues and perils of empiricist approaches to speech perception. J. Acoust. Soc. Am. 105, 503–511.
Kluender, K. R., Lotto, A. J., and Holt, L. L. 2005. Contributions of Nonhuman Animal Models to Understanding Human Speech Perception. In: Listening to Speech: An Auditory Perspective. (*eds*. S. Greenberg and W. Ainsworth), Oxford University Press.
Kluender, K. R., Lotto, A. J., Holt, L. L., and Bloedel, S. L. 1998. Role of experience for language-specific functional mappings of vowel sounds. J. Acoust. Soc. Am. 104, 3568–3582.
Kuhl, P. K. 1978. Perceptual constancy for speech-sound categories. N.I.C.H.D. *Conference on Child Phonology: Perception, Production, and Deviation*. Bethesda, Maryland.
Kuhl, P. K. 1979. Speech perception in early infancy: perceptual constancy for spectrally dissimilar vowel categories. J. Acoust. Soc. Am. 66, 1668–1679.
Kuhl, P. K. 1980. Perceptual Constancy for Speech-Sound Categories in Early Infancy. In: Child Phonology Vol. 2: Perception (*eds*. G. H. Yeni-Komshian, J. F. Kavanagh, and C. A. Ferguson), Academic Press.
Kuhl, P. K. 1983. Perception of auditory equivalence classes for speech in early infancy. Infant Behav. Dev. 6, 263–285.
Kuhl, P. K. 1987. Perception of Speech and Sound In Early Infancy. In: Handbook of Infant Perception Vol. 2 (*eds*. P. Salapatek and L. Cohen), pp. 257–381. Academic Press.
Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., and Lindblom, B. 1992. Linguistic experience alters phonetic perception in infants six-months of age. Science 255, 606–608.
Lecanuet, J. P., Granier-Deferre, C., Cohen, H., LeHouezec, R., and Busnel, M. C. 1986. Fetal responses to acoustic stimulation depend on heart rate variability pattern, stimulus intensity and repetition. Early Hum. Dev. 13, 269–283.
Lewis, J. W., Wightman, F., Brefczynski, J. A., Phinney, R. E., Binder, J. R., and DeYoe, E. A. 2004. Human brain regions involved in recognizing environmental sounds. Cereb. Cortex 14, 1008–21.
Lieberman, P. 1984. The Biology and Evolution of Language, Harvard University Press.
Liberman, A. M. and Mattingly, I. G. 1985. The motor theory of speech perception revisited. Cognition 21, 1–36.
Liebenthal, E., Binder, J. R., Spitzer, S. M., Possing, E. T., and Medler, D. A. 2005. Neural substrates of phonemic perception. Cereb. Cortex 15(10), 162–163.
Liljencrantz, J. and Lindblom, B. 1972. Numerical stimulation of vowel quality systems: the role of perceptual contrast. Language 48, 839–862.
Lindblom, B. E. F. 1963. Spectrographic study of vowel reduction. J. Acoust. Soc. Am. 35, 1773–1781.
Lindblom, B. and Maddieson, I 1988. Phonetic Universals in Consonant Systems. In: Language, Speech and Mind: Studies in Honour of Victoria A. Fromkin (*eds*. L. M. Hyman and C. N. Li), pp. 62–78. Routledge.
Lindblom, B. E. F. and Studdert-Kennedy, M. 1967. On the role of formant transitions in vowel recognition. J. Acoust. Soc. Am. 42, 830–843.
Lippman, R. P. 1996. Speech Perception by Humans and Machines. In: Workshop on the Auditory Basis of Speech Perception (*eds*. W. Ainsworth and S. Greenberg) pp. 309–316. Keele University.
Lisker L. 1978. Rapid versus rabid: a catalogue of acoustical features that may cue the distinction *Haskins Laboratories Status Report on Speech Research SR*-54, 127–132.
Lively, S. E. 1993. An examination of the perceptual magnet effect. J. Acoust. Soc. Am. 93, 2423.
Locke, S. and Kellar, L. 1973. Categorical perception in a non-linguistic mode. Cortex 9, 355–369.

Lotto, A. J. 2000. Reply to ''An analytical error invalidates the 'depolarization' of the perceptual magnet effect.''. J. Acoust. Soc. Am. 107, 3578–3580.
Lotto, A. J. and Kluender, K. R. 1998. General contrast effects in speech perception: effect of preceding liquid on stop consonant identification. Percept. Psychophys. 60, 602–619.
Lotto, A. J., Kluender, K. R., and Holt, L. L. 1997. Perceptual compensation for coarticulation by Japanese quail (*Coturnix coturnix japonica*). J. Accost. Soc. Am. 102, 1134–1140.
Luce, R. D. 1986. Response Times, Oxford University Press.
Maddieson, I. 1984. Patterns of Sound. Cambridge University Press.
Mann, V. A. 1980. Influence of preceding liquid in stop-consonant perception. Percept. Psychophys. 28, 407–412.
Mann, V. A. 1986. Distinguishing universal and language-dependent levels of speech perception: evidence from Japanese listeners' perception of English ''l'' and ''r''. Cognition 24, 169–196.
Mann, V. A. and Repp, B. H. 1981. Influence of preceding fricative on stop consonant perception. J. Accost. Soc. Am. 69, 548–558.
McGurk, H. and MacDonald, J. 1976. Hearing lips and seeing voices. Nature 264(5588), 746–748.
Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., and Amiel-Tison, C. 1988. A precursor of language acquisition in young infants. Cognition 29, 143–178.
Miller, G. A. 1956. The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychol. Rev. 63, 81–97.
Miyawaki, K., Strange, W., Verbrugge, R. R., Liberman, A. M., Jenkins, J. J., and Fujimura, O. 1975. An effect of linguistic experience: the discrimination of (r) and (l) by native speakers of Japanese and English. Percept. Psychophys. 18, 331–340.
Moore, B. C. J. and Glasberg, B. R. 1983. Suggested formulas for calculating auditory-filter bandwidths and excitation patterns. J. Accost. Soc. Am. 74, 750–753.
Nearey, T. M. 1997. Speech perception as pattern recognition. J. Accost. Soc. Am. 101, 3241–3254.
Olshausen, B. A. and Field, D. J. 1997. Sparse coding with an overcomplete basis set: a strategy employed by V1? Vision Res. 37, 3311–3325.
Olsho, L. W., Koch, E. G., Carter, E. A., Halpin, C. F., and Spetner, N. B. 1988. Pure-tone sensitivity of human infants. J. Acoust. Soc. Am. 84, 1316–1324.
Olsho, L. W., Koch, E. G., and Halpin, C. F. 1987. Level and age effects in infantfrequency discrimination. J. Accost. Soc. Am. 82, 454–464.
Pérez-González, D., Malmierca, M. S., and Covey, E. 2005. Novelty detector neurons in the mammalian auditory midbrain. Eur. J. Neurosci. 22, 2879–2885.
Polka, L. 1991. Cross-language speech perception in adults: phonemic, phonetic, and acoustic contributions. J. Accost. Soc. Am. 89(6), 2961–2977.
Polka, L. 1992. Characterizing the influence of native language experience on adult speech perception. Percept. Psychophys. 52(1), 37–52.
Port, R. F. (in press). The Graphical Basis of Phones and Phonemes. In: Second-Language Speech Learning: The Role Of Language Experience In: Speech Perception And Production (eds. M. Munro and O-S Bohn).
Remez, R. E., Rubin, P. E., Pisoni, D. B., and Carrell, T. D. 1981. Speech perception without traditional speech cues. Science 212, 947–950.
Repp, B. H. 1982. Phonetic trading relations and context effects: new evidence for a speech mode of perception. Psychol. Bull. 92, 81–110.
Riggs, L. A., Ratliff, F., Cornsweet, J. C., and Cornsweet, T. N. 1953. The disappearance of steadily fixated visual test objects. J. Opt. Soc. Am. 43(6), 495–501.
Rogers, T. T., Lambon Ralph, M. A., Garrard, P., Bozeat, S., McClelland, J. L., Hodges, J. R., and Patterson, K. 2004. The structure and deterioration of semantic memory: a neuropsychological and computational investigation. Psychol. Rev. 111, 205–235.
Rosenblat, F. 1958. The perceptron: a probabilistic model for information storage and organization in the brain. Psychol. Rev. 65, 386–408.
Saffran, J. R., Aslin, R. N., and Newport, E. L. 1996. Statistical learning by 8-month-old infants. Science 274, 1926–1928.
Saffran, J. R., Johnson, E. K., Aslin, R. N., and Newport, E. L. 1999. Statistical learning of tone sequences by human infants and adults. Cognition 70, 27–52.
Sato, M, Lotto, A. J., and Diehl, R. L. 2003. Patterns of acoustic variance in native and non-native phonemes. The J. Accost. Soc. Am. 114(4), 2392.
Schouten, J. F. 1940. The residue, a new component in subjective analysis. Proc. Kon. Akad. Wetensch 43, 356–365.
Schwartz, O. and Simoncelli, E. P. 2001. Natural signal statistics and sensory gain control. Nature: Neuroscience 4, 819–825.
Scott, S. K., Blank, C. C., Rosen, S., and Wise, R. J. S. 2000. Identification of a pathway for intelligible speech in the left temporal lobe. Brain 123, 2400–2406.
Shannon, C. E. 1948. A mathematical theory of communication. BSTJ 27, 379–423.
Shannon, R. V., Zeng, F. G., Wygonski, J., Kamath, V., and Ekelid, M. 1995. Speech recognition with primarily temporal cues. Science 270, 303–304.
Simoncelli, E. P. and Olshausen, B. A. 2001. Natural image statistics and neural representation. Ann. Rev. Neurosci. 24, 1193–1215.
Smith, R. L. 1979. Adaptation, saturations, and physiological masking in single auditory-nerve fibers. J. Accost. Soc. Am. 65, 166–178.
Smith, R. L., Brachman, M. L., and Frisina, R. D. 1985. Sensitivity of auditory-nerve fibers to changes in intensity: a dichotomy between decrements and increments. J. Accost. Soc. Am. 78, 1310–1316.
Smith, J. D., Kemler Nelson, D. G., Grohskopf, L. A., and Appleton, T. 1994. What child is this? What interval was that? Familiar tunes and music perception in novice listeners. Cognition 52, 23–54.
Smith, R. L. and Zwislocki, J. J. 1971. Responses of some neurons of the cochlear nucleus to tone-intensity increments. J. Accost. Soc. Am. 50, 1520–1525.
Stevens, K. N. and Blumstein, S. E. 1981. The Search for Invariant Acoustic Correlates of Phonetic Features. In: Perspectives in the Study of Speech (*eds*. P. D. Eimas and J. L. Miller), pp. 1–38. Erlbaum.
Stilp, C. and Kluender, K. R. 2006. Perceptual absorption of listening context when perceiving musical instruments. J. Accost. Soc. Am. 119, 3241.
Summerfield, Q., Haggard, M. P., Foster, J., and Gray, S. 1984. Perceiving vowels from uniform spectra: phonetic exploration of an auditory aftereffect. Percept. and Psychophys. 35, 203–213.
Trubetzkoy, N. S. 1969. Principles of Phonology (C. Baltaxe, Trans.). University of California Press. (Original work published in 1939).
Tuller, B., Case, P., Ding, M., and Kelso, J. A. S. 1994. The nonlinear dynamics of speech categorization. J. Exp. Psychol. Hum. Percept. Perform. 20(1), 3–16.
Ulanovsky, N., Las, L., and Nelken, I. 2003. Processing of low-probability sounds by cortical neurons. Nat. Neurosci. 6, 391–398.

Urbantschitsch 1876, *cf.* Abrahams, Krakauer, and Dallenbach (1937). Gustatory adaptation to salt. Am. J. Psychol. 49, 462–469.
Viemeister, N. F. 1980. Adaptation of Masking. In: Psychophysical, Physiological, and Behavioral Studies in Hearing (*eds*. G. van den Brink and F. A. Bilsen), pp. 190–197. Delft: University Press.
Viemeister, N. F. and Bacon, S. P. 1982. Forward masking by enhanced components in harmonic complexes. J. Accost. Soc. Am. 71, 1502–1507.
Voss, R. F. and Clarke, J. 1975. '1/*f* noise' in music and speech. Nature 258, 317–318.
Voss, R. F. and Clarke, J. 1978. ''1/*f* noise'' in music: Music from 1/*f* noise. J. Accost. Soc. Am. 63, 258–263.
Walley, A. C. 1993. The role of vocabulary development in children's spoken word recognition and segmentation ability. Dev. Rev. 13, 286–350.
Walley, A. C., Metsala, J. L., and Garlock, V. M. 2003. Spoken vocabulary growth: its role in the development of phoneme awareness and early reading ability. Read. and Writ. 16, 5–20.
Weiner, N. 1948. Cybernetics. Wiley.
Werker, J. F. and Curtin, S. 2005. PRIMIR: a developmental framework of infant speech processing. Lang. Learn. Dev. 1(2), 197–234.
Werker, J. F., Gilbert, J. H. V., Humphrey, K., and Tees, R. C. 1981. Developmental aspects of cross-language speech perception. Child Dev. 52, 349–355.
Werker, J. F. and Lalonde, C. E. 1988. Cross-language speech perception: initial capabilities and developmental change. Dev. Psychol. 24, 672–683.
Werker, J. F. and Logan, J. S. 1985. Cross-language evidence for three factors in speech perception. Percept. and Psychophys. 37, 35–44.
Werker, J. F. and Tees, R. C. 1983. Developmental changes across childhood in the perception of non-native speech sounds. Can. J. Psychol. 37, 278–286.
Werker, J. F. and Tees, R. C. 1984a. Cross-language speech perception: evidence for perceptual reorganization during the first year of life. Infant Behav. Dev. 7, 49–63.
Werker, J. F. and Tees, R. C. 1984b. Phonemic and phonetic factors in adult cross-language speech perception. J. Accost. Soc. Am. 75, 1866–1878.
Werner, L. and Gillenwater, J. 1990. Pure-tone sensitivity of 2-to 5-week-old infants. Infant Behav. Dev. 13, 355–375.
Wightman, F. L., McGee, T., and Kramer, M. 1977. Factors Influencing Frequency Selectivity in Normal and Hearing-Impaired Listeners. In: Psychophysics and Physiology of Hearing (*eds*. E. F. Evans and J. P. Wilson), pp. 295–310. Academic Press.
Wood, C. C. 1976. Discriminability, response bias, and phoneme categories in discrimination of voice onset time. J. Accost. Soc. Am. 60, 1381–1389.
Young, A. W., Rowland, D., Calder, A. J., Etcoff, N. L, Seth, A., and Perrett, D. I. 1997. Facial expression megamix: tests of dimensional and category accounts of emotional recognition. Cognition 63, 271–313.
Zatorre, R. J. and Binder, J. 2000. Functional and Structural Imaging of the Human Auditory System, In: Brain Mapping the Systems. (*eds*. A. Toga and J. Mazziotta), pp. 365–402. Academic Press.
Zwaardemaker, H. 1895. Die physiologie des geruchs, Engelmann.