

Chapter Published In:

Auditory Perception of Sound Sources

Series: [Springer Handbook of Auditory Research](#), Vol. 29

Yost, William A.; Popper, Arthur N.; Fay, Richard R. (Eds.)

2008, XIV, 332 p. 59 illus., Hardcover

ISBN: 978-0-387-71304-5

Chapter 10: Speech as a Sound Source

Andrew J. Lotto and Sarah C. Sullivan

1. What is the sound source for speech?

Speech is one of the most salient and important sound sources for the human listener. As with many other natural sound sources, a listener can localize the direction from which a signal originated and can even determine some of the physical characteristics of the sound-producing object and event. But the real value of the speech signal lies not just in where the sound came from or by whom the sound was created, but in the linguistic message that it carries. The intended message of the speaker is the real sound source of speech and the ability of listeners to apprehend this message in spite of varying talker and communication characteristics is the focus of this chapter.

This is not to say that the “where” and “by whom” questions related to the speech sound source are inconsequential. Localizing a speaker can be important for the segregation of their speech stream from competing speakers or noise (see Darwin, Chapter 7). Given the continuously varying nature of the speech signal, the segregation of speech from a particular talker is non-trivial and there is a long history of research into this problem (see Hafter and Sarampalis, Chapter 4; Carlyon and Gockel, Chapter 9). In

addition to perceiving the location of a speaker, listeners can learn quite a bit about the speaker from their productions. The information in the signal that specifies characteristics of the speaker, such as their gender, size or affect, is referred to as *indexical* information. The indexical information is similar to the shape, size, and material composition information for other sound producing objects/events (see Lutfi, Chapter 2; Patterson, Ives, and Walters, Chapter 3). It is clear that listeners can identify particular talkers from their speech (e.g., Bachorowski and Owren 1999) and this knowledge can color the interpretation of the incoming message. In the end, however, when one refers to *speech perception*, the task that comes to mind is the determination of the linguistic message intended by the speaker¹.

Even if one accepts that the true source perception problem for speech is identification of the message carried by the signal, it is still unclear what the unit of identification is. Words may seem to be a reasonable candidate, as they are the smallest units carrying semantic information. Some theorists have suggested that coherent theories of speech perception can be developed from the assumption of the word as the fundamental unit (e.g., Lindblom, MacNeilage, and Studdert-Kennedy 1984; Stevens 1986, 2002; Kluender and Lotto 1999). However, the vast majority of research in speech perception is focused on the identification of phonemes or phonetic categories. In fact, the study of speech perception and spoken word recognition do not overlap as much as one might expect and the fields generally cleave at the level of the phoneme. Most theories and models of speech perception explicitly state or implicitly assume that the goal or outcome of speech perception is a mapping from acoustics onto a phoneme or phonetic category representation. It is this mapping on which we will focus this review.

Whether the fundamental unit of speech perception turns out to be the phoneme or the word or the syllable (or the di-phone or tri-phone), it is likely that the concepts and results summarized here will apply generally.

In this chapter, we will present speech perception as a specific case of sound source identification. As with other source identification tasks, speech sound identification is based on the integration of multiple acoustic cues into a decision. However, the actual mapping from acoustic dimensions to phonetic categories is complicated by variability arising from speaker-specific characteristics, phonetic context and the vicissitudes of listening conditions. After reviewing studies that explore the mechanisms by which listeners accommodate this variability, we will attempt to synthesize the results by describing auditory perception as “relative”. That is, the perception of a particular sound is influenced by preceding (and following) sounds over multiple temporal windows. These effects of context (both temporally local and global) are likely to be important for any real-world perception of complex sounds (i.e., sound source perception).

2. Phonetic Categorization

Much of the tradition of speech perception research can be summarized as the study of *phonetic categorization*. That is, it has been focused on the ability of humans (and in some cases non-human animals) to map a set of sounds onto a discrete response typically corresponding to a phonetic segment (or minimal pair of syllables or words). For example, listeners are presented a synthesized series of syllables varying on a single acoustic dimension and are asked to press buttons labeled “da” and “ga” to identify the

sound. While it has not been established that this mapping is a necessary step in normal speech perception (Lotto and Holt 2000; Scott and Wise 2003), robust phonetic categorization in the face of many sources of acoustic variance remains one of the most remarkable achievements of human auditory perception.

The task in phonetic categorization studies is quite similar to the non-speech sound source identification tasks discussed by Lutfi (Chapter 2). For example, Lutfi and Oh (1997) presented participants with synthesized approximates of struck clamped bars that differed in material. The participants pressed a button to indicate which of two intervals contained the sound produced by a target material (e.g., iron versus glass). The sounds varied along the acoustic attributes that distinguished the two materials. In traditional *phonetic categorization* tasks, listeners are asked to identify a phonetic category (typically from a closed-set) based on sounds varying on just those dimensions that distinguish the categories. In fact, both of these tasks would be correctly referred to as *categorization* tasks. That is, a set of exemplars that vary on one or more physical dimensions are mapped onto a single response or label. The listener must be able to discriminate between members of each category but to generalize their response across members of the same category.

Another similarity between phonetic categorization and other sound source identification tasks is that the category distinction is defined by a number of acoustic attributes or cues. Lutfi (Chapter 2) enumerates a number of acoustic cues that are related to the material or length of clamped struck bars, including the amplitude, frequency and decay of different partials. Likewise, phonetic categories typically differ on a number of acoustic dimensions. For example, Lisker (1986) catalogued 16 different

acoustic cues to the English voicing distinction, e.g., /b/ versus /p/, in syllable-initial position². These include measures of relative amplitude, frequency and duration of various components.

Even vowel categories are distinguished by a large number of acoustic attributes. Most people are familiar with the defining nature of formant frequencies for vowels. Formants are peaks in the spectral envelope corresponding to resonances of the vocal tract. The center frequencies of the first two formants (F1 and F2, labeled in order of increasing frequency) do a fairly good job of segregating categories for steady-state vowels³. However, in natural speech, the steady state-vowel is a bit of a mythical creature. Vowel categories can also be distinguished by overall duration (Peterson and Lehiste 1960; Strange 1989; Hillenbrand, Clark, and Houde 2000), by the extent and direction of changes in formant values within the vowel or *vowel-inherent spectral change* (Nearey and Assmann 1986) and by the steepness of the spectral envelope slope or tilt (Kiefte and Kluender 2005).

Of course, the fact that sounds differ on a number of acoustic dimensions as a function of their category membership does not mean that all dimensions are equally informative for the category distinction. For example, the English distinction between /l/ and /r/ in the syllable-initial position is realized, in part, by the starting frequencies of F2 and F3 (which then transition to the formant values of the following vowel). If one plots the initial F2 and F3 frequencies for exemplars of /l/ and /r/ produced by a number of speakers with a variety of following vowels, the resulting distributions show very little overlap in the F3 dimension and quite a bit of overlap in F2 (Dalston 1975; Lotto, Sato, and Diehl 2004). That is, initial F3 is a far more reliable cue for distinguishing /l/ and /r/

than is initial F2. And, in fact, native English speakers rely much more on initial F3 than on F2 when categorizing these sounds (Yamada and Tohkura 1990; Iverson et al. 2003).

A particularly interesting aspect of speech perception is the salient effects of differential experience on phonetic categorization and discrimination. It is well documented that native Japanese speakers have difficulty perceiving and producing the English /r/ - /l/ distinction. One reason for this difficulty is that Japanese listeners appear to apply ineffective weighting functions to the cues for this distinction. That is, Japanese listeners tend to rely on initial F2 (as opposed to F3) when categorizing /l/ and /r/ exemplars (Yamada and Tohkura 1990; Iverson et al. 2003). Japanese productions of this contrast also result in distributions that are differentiated more by initial F2 than by F3 (Lotto, Sato, and Diehl 2004). This weighting strategy appears to be a result of learning Japanese, which contains a distinction between /w/ and a flap consonant that is similar to /l/ and /r/ but is distinguished by F2 (Lotto, Sato, and Diehl 2004). Thus, experience with a particular phonetic system can result in the application of sub-optimal weighting strategies for non-native contrasts (see also, Francis and Nusbaum 2002; Kim and Lotto 2002).

In summary, phonetic categorization is a process by which a listener determines a sound's category by integrating and weighting multiple cues and these weighting functions are not always optimal. This description should strike the reader as equally applicable to categorizing sounds on the basis of whether it was the result of a struck iron bar or a dropped wooden dowel, that is, it is a general description of sound source identification. One of the concerns in sound source identification is determining whether listeners are using optimal decision and weighting rules for a given task. Lutfi (2001),

for example, derives optimal weighting functions for hollowness detection analytically from equations describing the acoustic outputs of vibrating hollow and solid bars. Such an approach is unlikely to be feasible for determining optimal weighting strategies for phonetic categorization. While there are good models for predicting the acoustic output for different vocal tract configurations, it is doubtful that one will be able to develop analytical solutions that capture all of the variability inherent in different productions of the same phonetic segment. In fact, it is this variability in the mapping between acoustics and phonetic categories (or intended gestures) that is the bugaboo for the understanding and modeling of human speech perception.

The sources of the variability range from perturbations common to all sound sources such as room acoustics, channel transmission characteristics, and competing sources, to changes that are characteristic of speech like coarticulation and differences between talkers. Several of the other chapters in this volume that review the particular challenges of competing sources include discussion of speech signals (e.g., Hafter and Sarampalis, Chapter 4; Kidd, Mason, Richards, Gallun, and Durlach, Chapter 6; Darwin, Chapter 7). Here, we will concentrate on how the auditory system accommodates acoustic variation due to surrounding phonetic environment and talker-specific characteristics in phonetic categorization tasks. After reviewing some of the relevant empirical results, we will suggest that it is useful to conceptualize this accommodation as being the result of adaptive encoding by the auditory system working on multiple time scales.

3. Phonetic Context Effects

The acoustic pattern that is associated with a particular phonetic segment is notoriously context-dependent. One reason for this context dependence is that articulation is constrained by the physics of mass and inertia. At reasonable rates of speech production, it is difficult to move the articulators quickly enough to fully reach the targets that would characterize an articulation produced in isolation. For example, the vowel /ʊ/ (as in *but*) is produced in isolation with the tongue body relatively retracted. However, when producing *dud* the tongue moves anterior to produce the initial and final /d/ and may not completely retract for the vowel, leading to a “fronted” articulation of /ʊ/. However, a more retracted version of the vowel will occur in a /g_g/ context, where the /g/ articulation requires the tongue to make a more posterior occlusion. That is, the articulation of the vowel is assimilated to the articulations of the surrounding context consonants; it is *coarticulated*.

Coarticulation is not just the result of physical constraints on articulators. The articulation of a phoneme can be influenced by following phonemes (*anticipatory coarticulation*) and coarticulation occurs even when there is relatively little interdependence of the articulators involved in the target and context phonemes. It appears that coarticulation is in part a result of the motor plan for speech (Whalen 1990). In fact, some cases of coarticulation or context-dependent production may be specified at the level of linguistic rules (e.g., regressive place assimilation, Gaskell and Marslen-Wilson 1996).

Whatever the underlying causes are, the result of coarticulation is context-dependent acoustics for phonetic categories. This acoustic variability is not evident simply as noise on non-essential dimensions, but is present in those very dimensions that

serve as substantial cues to phoneme identification. This provides a difficulty for simple template- or feature-matching models of phonetic categorization because there are few acoustic invariants that one can point to as defining a particular category. In the vowel coarticulation example provided above, the result of coarticulation is that the formant frequency values during the “vowel portion” vary as a function of the surrounding consonants (See Fig. 10.1). At quick speaking rates, the formant values for /ʊ/ in “dud” resemble the values for the vowel /ʊ/ (the vowel in *bet*) spoken in isolation (Lindblom 1963; Nearey 1989). Thus, the approach of defining vowels simply by their formant frequencies is thwarted.

Another example, from Mann (1980), demonstrates coarticulation effects that cross over a syllable boundary between two consonants. As mentioned above, /d/ is articulated by creating an occlusion of the air stream relatively anterior in the mouth (at the alveolar ridge when produced in isolation). The exact placement of the tongue tip in creating this occlusion can be influenced by the context phonemes being produced. Producing /d/ after the matching anterior production of /l/ leads to a more anterior place of articulation. Producing /d/ after /r/ (produced with a more retracted tongue) results in a more posterior place of articulation. You can demonstrate this by producing /al da/ and /ar da/ with a natural or quick speaking rate. (You may try this at home, even without direct supervision.) The same coarticulation effects occur for /g/, which is produced with a relatively posterior occlusion. These articulation changes result in acoustic changes along the very dimensions that best distinguish /d/ from /g/. Figure 10.2 is a schematic of the formants for the four context-target conditions /al da/, /ar da/, /al ga/, /ar ga/ (based loosely on measures from Mann 1980). One can see in the context-consistent conditions

(consistent in anterior or posterior articulation), /al da/ and /ar ga/, that the main distinction between /d/ and /g/ is the onset frequency of F3. However, in the context-inconsistent conditions, /al ga/ and /ar da/, the F3 of /ga/ is drawn higher and the F3 of /da/ is drawn lower. The resulting syllables are nearly indistinguishable. How do listeners deal with this ambiguity?

Mann (1980) demonstrated that listeners accommodate context-dependent acoustics through context-sensitive perception. She presented listeners with a synthesized series of consonant-vowel (CV) stimuli that varied acoustically in initial F3 frequency and, consequently, varied perceptually from /da/ to /ga/. These stimuli were preceded by recordings of /al/ and /ar/ (with a 50-ms silent gap between syllables). Listeners' identifications of the target CVs indicated that the perceived identity of a phoneme was dependent on the preceding context. Following /al/ stimuli were identified as /ga/ more often than when the same stimuli followed /ar/. That is, identical acoustics lead to different perceptions depending on the preceding context. Note that the direction of this context-dependent shift is opposite the direction of coarticulation. In production, a preceding /al/ makes the CV more anterior or /da/-like. In perception, a preceding /al/ is identified as more /ga/-like. It appears that perception is compensating for the acoustic effects of coarticulation. In this way, the intended phoneme can be perceived despite variability in the acoustic form. This perceptual constancy is a hallmark of adaptive object or event perception (Brunswik 1956). It should stand as a particularly informative case of successful sound source identification.

A similar compensation for coarticulation is demonstrable for the case of a vowel coarticulated with, preceding, and following consonants, such as presented in Figure

10.1. Lindblom and Studdert-Kennedy (1967) first demonstrated this context-sensitive perception for Swedish vowels with liquid ('w' and 'y') contexts. To protect the average English reader from hurting themselves while attempting to produce Swedish vowels in /w_w/ frames, we describe here similar results obtained by Nearey (1989) and Holt et al. (2000). Listeners were presented vowels varying in F2 midpoint frequency, from a good /ɪ/ to a good /ʊ/, in either isolation or /d_d/ context. More /ɪ/ responses were made to the vowels in /d_d/ context than in isolation. This again reverses the effects of coarticulation, which would result in vowel acoustics more appropriate for /ɪ/ in this context. Several other examples of apparent compensation for coarticulation have been examined (see Repp 1982 for a review).

These demonstrations leave one wondering what aspect of the context is used by the auditory system to derive context-specific identifications of the target. Does the perceptual system recognize the phonemic content of the context and then shift identification based on this context identity? While this is an explanation preferred by some word recognition models (e.g., TRACE, Elman and McClelland 1988), it is inconsistent with some of the data. Infants as young as 4 months old demonstrate shifts in responses to /da-/ga/ stimuli as a function of /a/ or /ar/ context despite not having a developed phonological system (Fowler, Best, and McRoberts 1990) and native Japanese speakers show similar effects to English speakers despite not being able to discriminate /a/ and /ar/ (Mann 1986). Perhaps even more damaging to an account that relies on phonemic content of the context, birds trained to peck keys in response to presentations of /da/ or /ga/, show the same context-sensitive shift in responses that humans show for /a/ and /ar/ contexts (Lotto, Kluender, and Holt 1997). Thus, it does not appear that the

listener needs to access the phonemic identity of the context in order to compensate for coarticulation.

Lotto and Kluender (1998) proposed that it may be the spectral makeup of the context that determines the context effects in perception as opposed to its status as a phonemic entity. They presented listeners with a /ga/-/da/ series preceded by a frequency glide that tracked the transition of F3 for /r/ or /l/ (with a 50-ms silent interval). This context, which was not identifiable as a speech sound, was sufficient to result in a target categorization shift in the same direction as if the CVs were preceded by /al/ and /ar/. In a similar demonstration, Holt et al. (2000) replaced the /d_d/ context for a /ɪ/ to /ɪ/ series with tonal glides that matched the trajectory of F2. Once again, these non-phonetic contexts resulted in similar shifts as obtained for speech contexts. Non-speech context effects can be obtained from band-pass noise, sine-wave tones, or single formants filtered from speech (Holt 1999; Lotto 2004).

Regardless of whether the context is speech or non-speech, the effects on target identification can be described as contrastive. For example, /al/ with its *high*-frequency F3 offset (see Fig. 10.2) results in more /ga/ responses, which is characterized by a *low*-frequency F3 onset. Alternatively, an /ar/ context with a low-frequency F3 offset results in more high-frequency F3 onset, or /da/, responses. Tone or noise-band contexts centered on the F3 offsets for /al/ and /ar/ also result in contrastive shifts in target identification. Similarly, the high frequency F2 onset and offset of /d/ contexts (or FM glide analogs) result in a lower perceived F2 for vowels in /d_d/ context compared to isolation (/ɪ/ being the vowel typically containing a lower F2). Thus, the effects of

context can be predicted by the relative distribution of acoustic energy across frequencies regardless of the source of the context sound.

This constellation of findings implicates a rather general auditory process, which is insensitive to whether the sounds involved are speech. In addition to the demonstrations of non-speech contexts affecting speech target perception, one can obtain contrastive effects of speech contexts on the perception of target non-speech sounds (Stephens and Holt 2003) and non-speech context effects on non-speech targets (Aravamudhan 2005). If, in fact, a general auditory process is partly responsible for compensation for coarticulation, then it is not surprising that the effects are present in infants, or non-native language listeners (e.g., Japanese listeners and English stimuli), or even birds. One may also conclude that this process would play a role in sound source identification for sources that are not speech; that the identification of any complex sound may be affected by its acoustic context.

The original descriptions of these speech-non-speech context effects referred to the results as demonstrations of *frequency contrast* (Lotto and Kluender 1998). However, this is a misnomer, because the “frequencies” present in the speech contexts don’t change but the relative energy present at each frequency does change. The /al/ and /ar/ contexts contain harmonics at the same frequencies when produced with the same fundamental frequency. The difference between them is the distribution of energy amplitude across those harmonics with the peaks in energy defining the formants. Likewise, the targets /da/ and /ga/ differ in the relative amplitude of the harmonics in the F3 region. It is the amplitude differences between the spectral patterns that are being enhanced. *Spectral contrast* is a more appropriate description of these effects. Thus, one should be able to

predict the effect of a context by the frequency regions of its spectral prominences. Conversely, one should be able to predict a complementary effect for contexts that have spectral troughs. Coady et al. (2003) preceded a CV series varying in F2 onset (/ba-/da/) with a harmonic spectrum (rolling off at -6 dB/octave approximating the spectral tilt of speech) that contained either a low-frequency or high-frequency trough (0 energy at several consecutive harmonics) in the F2 region. The results demonstrated a contrastive effect of context. A context with a low-frequency *trough* leads to more target identifications consistent with a low-frequency *prominence* (i.e., /ba/).

The Coady et al. (2003) experiment is reminiscent of experiments conducted by Summerfield and colleagues on vowel “negative” aftereffects (Summerfield et al. 1984; Summerfield and Assmann 1987). They presented a uniform harmonic spectrum composed of equal-amplitude harmonics preceded by a spectral complement for a particular vowel (with troughs replacing formant prominences). Listeners reported hearing the vowel during presentation of the uniform spectrum. This result is in line with predictions of spectral contrast. Regions that are relatively prominent in the context are attenuated in the target and troughs in the contexts are enhanced in the target, in this case, leading to a pattern that resembles a vowel. Summerfield et al. (1984) note that the results are also consistent with the psychoacoustic phenomenon of auditory enhancement (Green, McKay, and Licklider 1959; Viemeister 1980; Viemeister and Bacon 1982). Auditory enhancement can be demonstrated by presenting an equal amplitude harmonic complex with one of the harmonics omitted followed by the same complex with the harmonic included. The replaced harmonic will stand out perceptually and its auditory representation appears to be enhanced, as it can lead to increased forward masking of a

tone relative to the complex being presented without the context (Viemeister and Bacon 1982). It is quite possible that the mechanisms underlying auditory enhancement produce some of the spectral contrast witnessed for speech sounds.

However, it doesn't appear that auditory enhancement can provide the complete story. Enhancement seems to be a largely monaural effect. Summerfield and Assmann (1989) failed to find effects of a precursor stimulus in their vowel experiments when the precursor was presented to the contralateral ear to the target. On the other hand, spectral contrast effects for speech are maintained even when the context (/al/ or /ar/) is presented to the opposite ear from the target (/da/-/ga/, Holt and Lotto 2002). Non-speech effects are also present for dichotic presentation of context and target (Lotto, Sullivan, and Holt 2003). For both speech and non-speech contexts, the effect is smaller for dichotic presentation versus diotic. These results suggest that peripheral mechanisms such as VIIIth nerve adaptation or adaptation of suppression may play a partial role, but that interactions are occurring more centrally as well. More evidence for non-peripheral mechanisms comes from examining the time course of speech effects. Holt and Lotto (2002) varied the duration of the silent gap between /al/-/ar/ contexts and CV targets from 25 to 400 ms (50 ms being the value used in all previously described experiments). There was a monotonic decrease in effect size with increasing interval duration, but the effect was still significant at 275 ms. Lotto et al. (2003) demonstrated an effect of non-speech context with a gap of 175 ms. These results again are consistent with both peripheral and more central mechanisms, because the effect is strongest for short intervals of 25 ms, within the temporal window of peripheral interactions, but lasts several hundred milliseconds, which is an unlikely window for purely peripheral

mechanisms. Viemeister and Bacon (1982) reported no appreciable auditory enhancement for their masking study beyond about 100 ms of silent gap.

Perhaps the best evidence that speech effects cannot be accounted for solely by peripheral interactions is that context can affect preceding targets. Wade and Holt (2005) had subjects identify words as “got” or “dot” with an embedded tone following the vowel. The tone was either high or low frequency. When the tone followed the consonant by 40 ms, it resulted in contrastive shifts in consonant identity (more “got” responses for embedded high frequency tone). Whether the mechanisms responsible for “forward” and “backward” contrast effects are the same remains an open question. But it is clear that the identification of a complex sound can be heavily influenced by its surrounding context.

Another question that is unanswered is how sound source segregation influences context effects. The fact that sounds obviously originating from different sources (e.g., speech and tones) can affect each other in perception suggests that context effects may precede or be independent from source segregation. However, strict tests of the priority of segregation and context effects have not been conducted. Whereas non-speech can affect speech when presented to opposite ears (Lotto, Sullivan, and Holt 2003), no one has tested whether a context that is localized to a specific region of exterior space will affect a target perceived as coming from a different location. Nor have there been attempts to manipulate the segregation of context and target by providing alternative perceptual organizations such as in an auditory streaming paradigm (Bregman 1990). It wouldn't be surprising if segregation influenced context effects. Empirical results from the visual modality demonstrate that context effects are malleable in relation to

perceptual organization. For example, Gilchrist (1977) has reported that brightness contrast occurs only for luminances that are perceived as coplanar (see also Gogel 1978). Source segregation may explain a finding from Lotto and Kluender (1998). They preceded a /da/-/ga/ series modeled on a male voice with /al/-/ar/ contexts produced by the same male or a female. The female contexts did result in a significant shift in target identification but the effect was significantly smaller than that obtained for the male contexts. Whether this difference was due to the listener perceiving the change in sources or because the spectral patterns for the female were not optimal for shifting the targets was not investigated.

Further investigation will also be required to resolve how spectral contrast interacts with linguistic information such as lexical status and phonological rules to determine a speech sounds identity. It is interesting to note that perceptual accommodation of linguistically-determined assimilation does not appear to require that one has experience with the particular language being presented (Gow and Im 2004). General perceptual mechanisms and principles may be involved even in these cases, which previously were accounted for by appealing to linguistic-specific knowledge (e.g., Gaskell and Marslen-Wilson 1996).

4. Talker Normalization

As discussed in section 2, an examination of the distributions of phonetic categories in acoustic space allows one to determine an optimal weighting and decision strategy for distinguishing contrasts in a particular language. Several theorists have proposed that language learners derive phonetic categories from these distributions averaged over many

encountered talkers (Kuhl 1993; Jusczyk 1997; Lotto 2000). However, whereas average distributions will provide a best guess as to phonetic identity across all talkers, they will be sub-optimal for any particular talker. While the acoustic variability associated with different talkers is useful when one's task is indexical identification (e.g., distinguishing the gender of the speaker), it can be a challenge for the robust identification of the intended phoneme. In order to effectively identify phonemes in all communication settings, listeners must be able to "tune" their auditory representations to the particular talker. This accommodation of talker-specific characteristics is referred to as *talker normalization* and has been a focus of speech perception research since the inception of the field (Potter and Steinberg 1950).

Peterson and Barney (1952) presented an early description of talker differences in vowel acoustics that has structured much of the work on talker normalization over the past 50 years. They measured formant frequency values for adult males and females and children for the vowels of English spoken in /h_d/ context. Despite the lack of context-induced variability, the distributions for the vowel categories show a great deal of dispersion and overlap (see Hillenbrand et al. 1995 for an updated data set). However, listeners can still identify the phoneme intended by the speaker. One way to account for this ability is to propose that listeners are using less-variable ratios of formants rather than treating each formant as an independent informative dimension (Fujisaki and Kawashima 1968; Traunmüller 1981; Syrdal and Gopal 1986; Miller 1989). This approach is exemplified by the suggestion of Potter and Steinberg (1950) that a vowel may be equivalent to a pattern of stimulation on the basilar membrane regardless of its location along the membrane. Sussman (1986) proposes that columns of combination-

sensitive neurons could encode the same formant ratios across changes in absolute frequencies. Talker normalization solutions, such as these, that rely on information contained solely within the vowel (or, more generally, the phonetic segment) have been referred to as *intrinsic* by Ainsworth (1975).

Whereas formant ratios can decrease some of the talker variability, there is clear evidence that listeners also apply normalization strategies that utilize information *extrinsic* to the target vowel. In 1957, Ladefoged and Broadbent conducted a classic experiment in which they presented listeners with a synthesized context sentence (*Please say what this word is __.*) followed by a synthesized vowel embedded in the frame /b_t/. Listeners identified the final target word as *bit*, *bet*, *bat* or *but*. Using a synthesizer, the researchers varied the spectral characteristics of the context sentence. For example, in one condition they lowered the range of F1 frequencies and in another raised it. Manipulations of the context sentence had large effects on the perceived vowel category. A vowel that was categorized as /ɪ/ (*bit*) by 88% of listeners following the unaltered context was categorized as /ɛ/ (*bet*) by 90% of listeners following a context with lowered F1 range. That is, the categorization of the vowel sound was strongly dependent on the characteristics of the context sentence.

Ladefoged and Broadbent (1957) also report that the manipulated context sentences maintained intelligibility but sounded as if they were produced by different speakers. From this view, the results can be interpreted as indicative of talker normalization. Listeners' responses appeared to be the result of tuning phonetic categories given information about typical values of formant frequencies for that speaker. In general, productions of the vowels /ɪ/ and /ɛ/ differ in F1 frequency with /ɪ/ having a

lower-frequency F1 (Peterson and Barney 1952). However, the actual value that corresponds to a “lower-frequency F1” is relative to speech produced by a talker. When the range of F1 in the context sentence is lowered, a moderate F1 value appears “high” and encourages an /ɪ/ categorization. Similar demonstrations of context sentences on target identifications have been made by a number of researchers (Broadbent and Ladefoged 1960; Ainsworth 1975; Assmann, Nearey, and Hogan 1982; Remez et al. 1987; Darwin, McKeown, and Kirby 1989; Ladefoged 1989; Nearey 1989; Johnson 1990; Watkins and Makin 1994).

Ladefoged and Broadbent (1957) proposed that their findings are consistent with a proposal by Joos (1948) that listeners make vowel identifications by referencing a talker-specific formant space created from the context material. That is, the identity of a vowel is dependent on its formant values relative to other vowels produced by the same talker. A vowel token positioned with respect to a low range of F1 values will be perceived differently from the same token positioned relative to high F1 values. This explanation is consistent with a number of theories of normalization in which it is proposed that the listener recalibrates their perception of a segment by making reference to talker-specific information (e.g., Ainsworth 1975; Nordstrom and Lindblom 1975; Nearey 1989). Common to these approaches is the requirement that the listener retains some distributional information about the speech of the particular talker, whether that information is just the average F3 values for back vowels (Nordstrom and Lindblom 1975), the ranges of the formant frequencies (Joos 1948) or an entire mapping of the vowel space. Results from word recognition and memory studies make it clear that some talker-specific information is retained in the speech representation (Goldinger 1998).

Much of the investigation of talker normalization has been concentrated on perceptual compensation for anatomical differences between talkers, such as gender differences. One approach to these differences is to rescale speech based on an estimate of vocal tract length (e.g., Nordstrom and Lindblom 1975). Alternatively, Patterson et al. (Chapter 3) present a transform that normalizes vowels by extracting variance related to vocal tract length. The problem with approaches that explicitly relate to vocal-tract length is that many speaker specific differences are not due strictly to length. Even the difference between males and females is partly due to differences in the proportions of different regions of the vocal tract, in addition to overall length. And even when one accounts for these structural differences, there is not complete overlap between male and female vowel spaces, suggesting a difference in articulation style (Fant 1966; Nordstrom 1977). In fact, there are many individual differences in production (Johnson, Ladefoged, and Lindau 1993) that are unrelated to anatomy, including dialect and accent. It is likely that a problem as complicated as talker normalization for speech will be accomplished by a number of different processes working sequentially or in parallel.

The fact that general auditory processes appear to play some role in compensation for coarticulation may lead one to question whether there are general processes that aid in talker normalization. The size normalization process proposed by Patterson et al. (Chapter 3) may be an example of an auditory process not specialized for human speech that is involved in talker normalization. Recently, Holt (2005) described a new auditory phenomenon that may also play an important role in normalization. The stimulus paradigm appears to be a mix of the normalization experiment of Ladefoged and Broadbent (1957) and the non-speech / speech context effect experiments of Lotto and

Kluender (1998). The target that listeners had to identify was a member of a /da/-/ga/ series (modified from natural speech tokens). The target was preceded by a 70-ms tone situated at a frequency that was shown to be a neutral context (set at a frequency that was in the middle of the F3 range for the CV). This *standard tone* was, in turn, preceded by a series of 21 70-ms tones varying in frequency. The 21 tones were randomly sampled from a rectangular distribution of tone frequencies that either had a low or high mean. The low mean corresponded to the F3 offset-frequency of /ar/ and the high mean corresponded to the F3 for /al/ from the experiments by Lotto and Kluender (1998). The context tones are referred to as the *acoustic history*. Representations of the stimuli are presented in Figure 10.3. As in the context effects experiments, listeners were asked to identify the final syllable as /da/ or /ga/. However, unlike the previous experiments, the context was not adjacent to the syllable (the neutral standard tone always directly preceded the target) and the difference in the context conditions cannot be described in terms of a specific spectral pattern (the order of tones in the acoustic histories changed on each trial). Nevertheless, the results resemble those obtained in the context effects experiments. Listeners identified the target as /ga/ more often following the high mean history and as /da/ more often following the low mean history.

This is a contrastive response pattern, except that the contrast is not with a particular spectral pattern but with the spectral energy averaged over a relatively long (over 2 s) temporal window. In support of the conclusion that this is another contrast effect, Holt (in press) demonstrated that complementary results can be obtained when the acoustic history is a series of noise bursts with troughs at sampled frequencies instead of tones.

The acoustic histories of Holt (2005) resemble in some respect the carrier sentences of Ladefoged and Broadbent (1957). Both are extended contexts that differ in the range of frequencies that contain amplitude peaks (tones or first formants). Given this correspondence, one may propose that a similar process plays a role in both demonstrations. The results of Ladefoged and Broadbent (1957) can also be re-described in contrastive terms. If one lowers the average frequency of F1 in the carrier sentence then the F1 for the vowel in the target word is perceived as higher (i.e., more /ɪ/). That is, it may be that talker normalization is, in part, another example of spectral contrast influencing speech perception.

There are several other studies that demonstrate that perception of a target syllable is influenced by the spectral makeup of the carrier phrase and in each case the effect can be described as contrast with the average spectral pattern of the precursor. Watkins (1988; 1991) applied a filter to a carrier phrase and demonstrated that a target vowel was perceived as if it was filtered with an inverse of the phrase filter (see also Watkins and Makin 1994; 1996). Similarly, Kiefte and Kluender (2001) presented carrier phrases that varied in the slope of their spectral tilt (the slope of the amplitude fall-off for higher-frequency harmonics). Steeper spectral tilts led to target vowel identifications that were more consistent with a shallow spectral tilt. One can conceive of these demonstrations as examples of talker normalization or normalizing for the effects of filtering by a transmission channel. Whatever the cause of these deviations, the effects appears to be that target identification is made relative to the preceding (and following, Watkins and Makin 1996) spectral patterns.

The demonstrations of context-based perception discussed thus far are related to spectral differences in the context, but what of temporal differences? One salient difference between talkers is speaking rate. Given that temporal cues (such as voice onset time) are important for phonetic categorization, it would appear necessary that listeners compensate for inherent temporal variations among talkers. As an example, the distinction between /ba/ and /wa/ in English is, in part, defined by the duration of the formant transitions from onset to the vowel; short duration transitions are associated with /b/. (Think of the production in each case as movement away from approximated lips. This movement is faster for /ba/). However, these transition durations also vary with speaking rate (Miller and Baer 1983). Listeners appear to accommodate speaking rate variation by perceiving the transition duration relative to the following vowel duration, which could be considered a correlate of speaking rate. A synthesized CV that is perceived as /wa/ when the vowel is short will be perceived as /ba/ when the vowel is lengthened (Miller and Liberman 1979). This is again a contrastive response pattern in phonetic categorization. The effective perceived transition duration is shortened when the vowel is lengthened. The same pattern can be witnessed in non-speech categorization. Pisoni et al. (1983) reported analogous shifts for sine-wave analogs of /ba/ and /wa/ that were categorized as beginning with an “abrupt” or “gradual” transition (see also Diehl and Walsh 1989). The implication that a general contrast process may underlie this context effect is consistent with the findings of vowel length effects for infants (Jusczyk et al. 1983) and non-human animals (macaques: Stevens, Kuhl, and Padden 1988; budgerigars: Dent et al. 1997).

As with spectral effects, one can demonstrate that changing the average durations for segments (speaking rate) in a carrier phrase will affect target identification (Diehl, Souther, and Convis 1980; Summerfield 1981; Kidd 1989; Wayland, Miller, and Volaitis 1994). Wade and Holt (2005) utilized the acoustic histories paradigm described above to examine whether carrier phrase effects could be induced with non-speech precursors. They preceded members of a /ba/-/wa/ series (varying in formant transition duration) with a series of tones sampled from a single rectangular distribution with a range from F1 to F2. The context conditions differed in terms of the duration of these tones, with short (30 ms) and long (110 ms) conditions. The precursors had a reliable contrastive effect on the categorization of the target CV (more /ba/ responses for long condition). Thus, it appears that rate normalization shares much in common with the other versions of talker normalization reviewed above.

Whereas the correspondence of speech and non-speech effects presented here implicates general auditory processes in talker normalization, it should be noted again that normalization is a complex problem that likely requires a multitude of mechanisms. Many of the differences between talkers cannot be summarized as overall changes in rate or average spectra. For example, the perturbations of production resulting from a foreign accent or dialect difference are often specific to individual phonetic categories. Yet listeners appear to be able to adjust their categorization on the basis of a talker's dialect (Evans and Iverson 2004). Listeners also appear to make use of information from vision when normalizing for a talker (Johnson, Strand, and D'Imperio 1999; Glidden and Assmann 2004). In order to account for the entire constellation of findings, it is likely that the proposal of many perceptual processes will need to be entertained. However, a

subset of these processes appear to be of a general auditory nature and are likely to play a role in any real-world sound source identification task.

5. A Synthesis: Relative Perception

In this review, we have proposed that phonetic categorization is an example of a sound source identification task. As such, the results of investigations into perceptual weighting strategies, source segregation, auditory attention and memory, etc. discussed in the other chapters of this volume may be applied to the complex problem of speech perception. Another implication of this proposal is that phenomena in speech perception may provide insights into the auditory processes that are active for categorization of any complex sound. The demonstrations of phonetic context effects (or compensation for coarticulation) and talker normalization reviewed here indicate that the identification of a target sound can be influenced by the acoustic makeup of surrounding context sounds. To the extent that sound sources are not perceived in isolation, contextual sounds may be an important determiner of behavior in many non-speech identification tasks.

The effects of context on identification can be described as contrastive. For example, energy in a particular frequency region is perceived as less intense in contrast to a preceding (or following) peak of energy in that region. What general mechanisms in the auditory system lead to this type of perceptual contrast? There are a number of candidate neural mechanisms that emphasize the difference between sounds. Delgutte and his colleagues (1996; Delgutte 1997) have established a case for a broad role for neural adaptation in perception of speech, noting that the adaptation may enhance spectral contrast between sequential segments. This contrast is predicted to arise because

neurons adapted by stimulus components close to their preferred (characteristic) frequency are relatively less responsive to subsequent energy at that frequency, whereas components not present (or weakly present) in a prior stimulus are encoded by more responsive unadulterated neurons. Adaptation of suppression is another possible contrast-inducing mechanism that has been implicated in auditory enhancement (Palmer, Summerfield, and Fantini 1995). Clearly, neural adaptation is a mechanism that would be active in both speech and non-speech source identification tasks.

Recent studies have provided strong evidence that the auditory system, like the visual system (e.g., Movshon and Lennie 1979; Saul and Cynader 1989), exhibits another form of adaptation – known as stimulus-specific adaptation (SSA) – that has intriguing parallels to the spectral contrast effects reviewed above. Ulanovsky et al. (2003; 2004) have demonstrated SSA in primary auditory cortex using a version of the “oddball” paradigm common to mismatch negativity studies (Näätänen, Gaillard, and Mäntysalo 1978). In particular, they presented a repeating tone as a standard that was sporadically replaced by a deviant tone with a different frequency. The response to the deviant tone was enhanced relative to when the tones were presented equally often in a sequence. That is, the cortical neurons provide an enhanced response to acoustic novelty. This is a contrastive response pattern. The effects of context in speech can also be viewed as an enhancement to change from the prevailing acoustic environment. The acoustic histories of Holt (2005) establish a context with energy centered in high or low frequency regions and the introduction of components outside of those regions leads to a perceptual emphasis of those components.

Abrupt changes in sound waves or light are indicative of novel forces working on an object or of the presence of multiple sources. Emphasis of change, whether it is spectral contrast or brightness contrast, can help the perceiver in directing attention to new information or to segregate different sources. Thus, contrast appears to be not just a single process or the result of a single mechanism, but is instead an operating characteristic of adaptive perceptual systems.

In order to detect change, perceptual systems need to retain information about context stimuli. This retention appears to operate over multiple time scales. In phonetic context effects, the time scale is on the order of 10s to 100s of milliseconds. In the carrier phrase and acoustic history experiments, the time scale appears to be seconds. One could consider this retention to be an example of *auditory memory* (see Demany and Semal, Chapter 5). However, memory is a term that is usually associated with cognition as opposed to perception. We prefer to think of the tracking of statistical regularities in the input and the encoding of targets relative to those regularities as fundamental to perception.

Given the purported importance of tracking statistics to source perception, it is incumbent on us to determine what “statistics” are computed and over what temporal windows they are computed. Data from carrier phrase and acoustic history experiments suggest that the average spectra of contexts are likely computed. In the carrier phrase experiments of Kiefte and Kluender (2001), listeners appear to extract the average spectral tilt of the precursor and perceive the target relative to that average spectrum. In Holt’s (2005) acoustic history experiments the mean of the tone distributions seem to be extracted for comparison with the target. In a follow-ups study, Holt (in press)

demonstrated that repeated presentation of a tone with the mean frequency had the same effect on identification as presentation of the entire distribution and that, in general, the variance of the distribution plays little or no role in the effect.

The extraction of the average spectrum by the auditory system provides a possible means of normalizing for talker differences. Work on speech production models by Story and colleagues (2002; Story 2005) has provided evidence that individual talker differences are apparent in the vocal tract shape used in the production of a neutral or average vowel. The productions of other vowels and consonants can be considered as perturbations of this neutral vowel shape. These perturbations are remarkably consistent across talkers, so that much of the talker variability is captured by the differences in the neutral vowel shape. If the auditory system is extracting an average spectra and then enhancing deviations from that average (contrast), then one can think of the perceiver as extracting the acoustics of the neutral (average) vocal tract shape and enhancing the perturbations from this average, which result from the phonetic articulations of the speaker. This tuning of perception to the average spectra (and by extension neutral vocal tract) of a speaker would drastically lower the variability associated with talker differences. Again, this beneficial result for robust speech communication is only a specific case of the general processes involved in auditory source perception.

It is likely that the auditory system can track regularities beyond mean spectra. The results of temporal contrast studies, such as those involving /ba/ versus /wa/, indicate that average duration of segments or something like it is tracked. The work on SSA in audition and vision (e.g., Fairhall et al. 2001; Ulanovsky et al. 2004) suggests that there are a variety of statistical regularities in sensory signals that can be tracked. Similar

conclusions come from the literature on mismatch negativity studies, which have demonstrated that the auditory system reacts to deviants from a standard stimulus repetition based on amplitude, intensity, spatial location and even phonetic category (see Naatanen and Winkler 1999 for a review). Certainly, there must be some restrictions on the types of regularities that are extracted but, to date, there has not been systematic study of these constraints. The context studies reviewed here provide a possible paradigm for testing the limits of the auditory system's abilities in this regard.

The concept of perceiving a target sound with respect to previous statistical or distributional information can be extended to the entire process of categorization as discussed in section 2 of this review. We presented the idea of optimal cue weighting strategies as determinable from the category distributions described in acoustic space. If listeners do develop weighting strategies based on the distributions of experienced exemplars, then they must retain some description of these distributions that is created over time. It is unclear what exactly is retained. It could be something as detailed as a full representation of each exemplar (e.g., Goldinger 1997; Johnson 1997) or a "tally" of the values of experienced exemplars on a constrained set of acoustic attributes. Whatever the answer turns out to be, it is becoming clear that listeners retain a fairly good representation of the distributions of experienced sounds. Sullivan et al. (2005) presented bands of noise varying in center frequency from two overlapping distributions that were arbitrarily labeled as categories "A" and "B" to listeners who learned to categorize the sounds with feedback. Within 6 minutes of training (one repetition of the 50 stimuli in each distribution), the participants were able to categorize the sounds with near optimal performance. In order to do this, they had to calculate the cross-over point of the two

distributions and use it as a decision criterion. Listeners appeared to do this with notable precision. Obviously, phonetic categories and categories for other sound sources are developed over a longer time interval than a single experimental session, but the parallels between phonetic context effects and the formation of phonetic categories are intriguing. In each case, the perception of a target is made relative to a larger context, whether it is a carrier phrase or all experienced tokens of different phonemes. There is even evidence for contrastive effects at the category level. The exemplars of vowel categories that are judged as “Best” members of the category or result in the strongest responses are not those exemplars that are most typical but those that are most different from competing categories (Johnson, Flemming, and Wright 1993; Kluender et al. 1998). Also, a vowel that is ambiguous between two categories preceded by a good exemplar from one of the categories will be perceived as a member of the contrasting vowel category (Repp, Healy, and Crowder 1979; Healy and Repp 1982; Lotto, Kluender, and Holt 1998). Thus, there appear to be similarities in response patterns and importance of context that extends from peripheral neural adaptation to categorization, across time scales differing in many orders of magnitude.

Whether these similarities are superficial or whether they reveal something fundamental about auditory perception remains to be seen. But as hearing scientists move towards an understanding of sound source perception in the environment, it is clear that it will not be sufficient to examine the ability of listeners to detect an acoustic feature or register a value along an acoustic dimension in isolation. Perception in the real world is about perception in context.

Acknowledgements: Preparation of this chapter was supported in part by grants from NIH-NIDCD and NSF.

Endnotes

¹ It will become clear in the remainder of this chapter that talker specific characteristics play a role in speech perception. Here we are describing indexical identification as an outcome of sound source perception.

² It should be noted that the acoustic cues (and their perceptual weighting) differ for sounds that we label with the same phoneme when they appear in different positions in a syllable. For example, the acoustic cues that best distinguish English /l/ and /r/ described later in the text are only relevant when these sounds appear in a syllable-initial position. When the sounds occur in the syllable-final position, the relative importance of the cues changes (Sato, Lotto, and Diehl 2003). Whereas we label these sounds with the same phoneme and orthographic symbols regardless of position, they may be most appropriately considered different phonetic categories that are provided the same labels when we learn to read.

³ Patterson et al. (Chapter 3) provide a description of the acoustic characteristics of vowels as developed from the source – filter theory. In this chapter, we have opted to omit an overview of speech acoustics in favor of providing specific acoustic descriptions for phonetic distinctions as they are discussed.

References

Ainsworth WA (1975) Intrinsic and extrinsic factors in vowel judgments. In: Fant G, Tatham M (eds) *Auditory Analysis and Perception of Speech*. London: Academic Press, pp. 103-113.

Aravamudhan R (2005) *Perceptual overshoot with speech and nonspeech sounds*. Ph.D. Dissertation, Kent State University, Kent, OH.

Assmann PF, Nearey TM, Hogan JT (1982) Vowel identification: Orthographic, perceptual and acoustic aspects. *J Acoust Soc Am* 71:975-989.

Bachorowski JA, Owren MJ (1999) Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in running speech. *J Acoust Soc Am* 106:1054-1063.

Bregman AS (1990) *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA: Bradford Books, MIT Press.

Broadbent DE, Ladefoged P (1960) Vowel judgments and adaptation level. *Proc Biol Sci* 151:384-399.

Brunswik E (1956) *Perception and the Representative Design of Psychological*

Experiments. Berkeley, CA: University of California Press.

Coady JA, Kluender KR, Rhode WS (2003) Effects of contrast between onsets of speech and other complex spectra. *J Acoust Soc* 114:2225-2235.

Dalston RM (1975) Acoustic characteristics of English /w, r, l/ spoken correctly by young children and adults. *J Acoust Soc* 57:462-469.

Darwin CJ, McKeown JD, Kirby D (1989) Perceptual compensation for transmission channel and speaker effects on vowel quality. *Spe Com* 8:221-234.

Delgutte B (1997) Auditory neural processing of speech. In: Hardcastle WJ, Laver J (eds) *The Handbook of Phonetic Sciences*. Oxford: Blackwell, pp. 507-538.

Delgutte B, Hammond BM, Kalluri S, Litvak LM, Cariani P (1996) Neural encoding of temporal envelope and temporal interactions in speech. In: Ainsworth W, Greenberg S (eds) *Proceedings of the ESCA Research Workshop on the Auditory Basis of Speech Perception*. pp. 1-11.

Dent ML, Brittan-Powell EF, Dooling RJ, Pierce A (1997) Perception of synthetic /ba/ /wa/ speech continuum by budgerigars (*Melopsittacus undulatus*). *J Acoust Soc* 102:1891-1897.

Diehl RL, Walsh MA (1989) An auditory basis for the stimulus-length effect in the perception of stops and glides. *J Acoust Soc Am* 85:2154-2164.

Diehl RL, Souther AF, Convis CL (1980) Conditions on rate normalization in speech perception. *Percept Psychophys* 27:435-443.

Elman JL, McClelland JL (1988) Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *J Mem Lang* 27:143-165.

Evans BG, Iverson P (2004) Vowel normalization for accent: An investigation of best exemplar locations in northern and southern British English sentences. *J Acoust Soc Am* 115:352-361.

Fairhall AL, Lewen GD, Bialek W, de Ruyter van Steveninck RR (2001) Efficiency and ambiguity in an adaptive neural code. *Nature* 412:787-792.

Fant G (1966) A note on vocal tract size factors and non-uniform F-pattern scalings. *Speech Trans Lab Quart Prog Stat Rep* 7:22-30.

Fowler CA, Best CT, McRoberts GW (1990) Young infants' perception of liquid coarticulatory influences on following stop consonants. *Percept Psychophys* 48:559-570.

Francis AL, Nusbaum HC (2002) Selective attention and the acquisition of new phonetic categories. *J Exp Psychol [Hum Percept]* 28:349-366.

Fujisaki H, Kawashima T (1968) The roles of pitch and higher formants in the perception of vowels. *IEEE Trans Audio Elect AU-16*:73-77.

Gaskell G, Marslen-Wilson WD (1996) Phonological variation and inference in lexical access. *J Exp Psychol [Hum Percept]* 22:144-158.

Gilchrist A (1977) Perceived lightness depends on perceived spatial arrangement. *Science* 195:185-187.

Glidden CM, Assmann PF (2004) Effects of visual gender and frequency shifts on vowel category judgments. *Acoust Res Let Online* 5:132-138.

Gogel WC (1978) The adjacency principle in visual perception. *Sci Am* 238:126-139.

Goldinger SD (1997) Words and voices: Perception and production in an episodic lexicon. In: Johnson K, Mullennix JW (eds) *Talker Variability in Speech Processing*. San Diego, CA: Academic Press, pp. 33-66.

Goldinger SD (1998) Echoes of echoes? An episodic theory of lexical access. *Psychol Rev* 105:251-279.

Gow DW, Im AM (2004) A cross-linguistic examination of assimilation context effects. *J Mem Lang* 51:279-296.

Green DM, McKay MJ, Licklider JCR (1959) Detection of a pulsed sinusoid in noise as a function of frequency. *J Acoust Soc Am* 31:1446-1452.

Healy AF, Repp BH (1982) Context independence and phonetic mediation in categorical perception. *J Exp Psychol [Hum Percept]* 8:68-80.

Hillenbrand JM, Clark MJ, Houde RA (2000) Some effects of duration on vowel recognition. *J Acoust Soc Am* 108:3013-3022.

Hillenbrand JM, Getty L, Clark MJ, Wheeler K (1995) Acoustic characteristics of American English vowels. *J Acoust Soc Am* 97:3099-3111.

Holt LL (1999) Auditory constraints on speech perception: An examination of spectral contrast. *Diss Abstr Int (Sci)* 61:556.

Holt LL (2005) Temporally non-adjacent non-linguistic sounds affect speech categorization. *Psychol Sci* 16:305-312.

Holt LL (in press) The mean matters: Effects of statistically-defined non-speech spectral

distributions on speech categorization. *J Acoust Soc Am*.

Holt LL, Lotto AJ (2002) Behavioral examinations of the level of auditory processing of speech context effects. *Hear Res* 167:156-169.

Holt LL, Lotto AJ, Kluender KR (2000) Neighboring spectral content influences vowel identification. *J Acoust Soc Am* 108:710-722.

Iverson P, Kuhl PK, Akahane-Yamada R, Diesch E, Tohkura Y, Kettermann A, Siebert C (2003) A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition* 87:B47-B57.

Johnson K (1990) The role of perceived speaker identity in F0 normalization of vowels. *J Acoust Soc Am* 88:642-654.

Johnson K (1997) Speech perception without speaker normalization: An exemplar model. In: Johnson K, Mullennix JW (eds) *Talker Variability in Speech Processing*. San Diego, CA: Academic Press, pp. 145-166.

Johnson K, Flemming E, Wright R (1993) The hyperspace effect: Phonetic targets are hyperarticulated. *Language* 69:505-528.

Johnson K, Ladefoged P, Lindau M (1993) Individual differences in vowel production. *J*

Acoust Soc Am 94:701-714.

Johnson K, Strand EA, D'Imperio M (1999) Auditory-visual integration of talker gender in vowel perception. *J Phonet* 27:359-384.

Joos M (1948) *Acoustic Phonetics*. *Language* 24:1-136.

Jusczyk PW (1997) *The Discovery of Spoken Language*. Cambridge, MA: MIT Press.

Jusczyk PW, Pisoni DB, Reed M, Fernald A, Myers M (1983) Infants' discrimination of the duration of a rapid spectrum change in nonspeech signals. *Science* 222:175-177.

Kidd GR (1989) Articulatory-rate context effects in phoneme identification. *J Exp Psychol [Hum Percept]* 15:736-748.

Kiefte M, Kluender KR (2001) Spectral tilt versus formant frequency in static and dynamic vowels. *J Acoust Soc Am* 109:2294-2295.

Kiefte M, Kluender KR (2005) The relative importance of spectral tilt in monophthongs and diphthongs. *J Acoust Soc Am* 117:1395-1404.

Kim M-RC, Lotto AJ (2002) An investigation of acoustic characteristics of Korean stops produced by non-heritage learners. In: Ree JJ (ed) *The Korean Language in America*.

American Association of Teachers of Korean, pp. 177-188.

Kluender KR, Lotto AJ (1999) Virtues and perils of an empiricist approach to speech perception. *J Acoust Soc Am* 105:503-511.

Kluender KR, Lotto AJ, Holt LL, Bloedel SL (1998) Role of experience for language specific functional mappings of vowel sounds. *J Acoust Soc Am* 104:3568-3582.

Kuhl PK (1993) Early linguistic experience and phonetic perception: Implications for theories of developmental speech perception. *J Phonet* 21:125-139.

Ladefoged P (1989) A note on 'Information conveyed by vowels'. *J Acoust Soc Am* 85:2223-2224.

Ladefoged P, Broadbent DE (1957) Information conveyed by vowels. *J Acoust Soc Am* 29:98-104.

Lindblom B (1963) Spectrographic study of vowel reduction. *J Acoust Soc Am* 35:1773-1781.

Lindblom B, Studdert-Kennedy M (1967) On the role of formant transitions in vowel recognition. *J Acoust Soc Am* 42:830-843.

Lindblom B, MacNeilage P, Studdert-Kennedy M (1984) Self-organizing processes and the explanation of language universals. In: Butterworth B, Comrie B, Dahl Ö (eds) Explanations for Language Universals. Berlin: Walter de Gruyter and Co, pp. 181-203.

Lisker L (1986) "Voicing" in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees. *Lang Speech* 29:3-11.

Lotto AJ (2000) Language acquisition as complex category formation. *Phonetica* 57:189-196.

Lotto AJ (2004) Perceptual compensation for coarticulation as a general auditory process. In: Agwuele A, Warren W, Park S-H (eds) Proceedings of the 2003 Texas Linguistic Society Conference. Sommerville, MA: Cascadilla Proceedings Project, pp. 42-53.

Lotto AJ, Holt LL (2000) The illusion of the phoneme. In: Billings SJ, Boyle JP, Griffith AM (eds) Chicago Linguistic Society, Volume 35: The Panels. Chicago: Chicago Linguistic Society, pp. 191-204.

Lotto AJ, Kluender KR (1998) General contrast effects of speech perception: Effect of preceding liquid on stop consonant identification. *Percept Psychophys* 60:602-619.

Lotto AJ, Kluender KR, Holt LL (1997) Perceptual compensation for coarticulation by Japanese quail (*Coturnix coturnix japonica*). *J Acoust Soc Am* 102:1134-1140.

Lotto AJ, Kluender KR, Holt LL (1998) The perceptual magnet effect depolarized. *J Acoust Soc Am* 103:3648-3655.

Lotto AJ, Sato M, Diehl RL (2004) Mapping the task for the second language learner: The case of Japanese acquisition of /r/ and /l/. In: Slifka J, Manuel S, Matthies M (eds) *From Sound to Sense: 50+ Years of Discoveries in Speech Communication*. Electronic Conference Proceedings, pp. C181-C186.

Lotto AJ, Sullivan SC, Holt LL (2003) Central locus for nonspeech context effects on phonetic identification. *J Acoust Soc Am* 113:53-56.

Lutfi RA (2001) Auditory detection of hollowness. *J Acoust Soc Am* 110:1010-1019.

Lutfi RA, Oh EL (1997) Auditory discrimination of material changes in a struck-clamped bar. *J Acoust Soc Am* 102:3647-3656.

Mann VA (1980) Influence of preceding liquid on stop-consonant perception. *Percept Psychophys* 28:407-412.

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

Miller JD (1989) Auditory-perceptual interpretation of the vowel. *J Acoust Soc Am* 85: 2114-2134.

Miller JL, Baer T (1983) Some effects of speaking rate on the production of /b/ and /w/. *J Acoust Soc Am* 73:1751-1755.

Miller JL, Liberman AM (1979) Some effects of later-occurring information on the perception of stop consonant and semivowel. *Percept Psychophys* 25:457-465.

Movshon JA, Lennie P (1979) Pattern-selective adaptation in visual cortical neurons. *Nature* 278:850-852.

Naatanen R, Winkler I (1999) The concept of auditory stimulus representation in cognitive science. *Psychol Bull* 125:826-859.

Naatanen R, Gaillard AW, Mantysalo S (1978) Early selective attention effect on evoked potential reinterpreted. *Acta Psychol* 42:313-329.

Nearey TM (1989) Static, dynamic, and relational properties in vowel perception. *J Acoust Soc Am* 85:2088-2113.

Nearey TM, Assmann PF (1986) Modeling the role of inherent spectral change in vowel

identification. *J Acoust Soc Am* 80:1297-1308.

Nordstrom PE (1977) Female and infant vocal tracts simulated from male area functions. *J Phonet* 5:81-92.

Nordstrom PE, Lindblom B (1975) A normalization procedure for vowel formant data. In: *Proceedings of the 8th International Congress of Phonetic Sciences*. Leeds, England, pp. 212.

Palmer AR, Summerfield Q, Fantini DA (1995) Responses of auditory-nerve fibers to stimuli producing psychophysical enhancement. *J Acoust Soc Am* 97:1786-1799.

Peterson GE, Barney HL (1952) Control methods used in a study of the vowels. *J Acoust Soc Am* 24:175-184.

Peterson GE, Lehiste I (1960) Duration of syllable nuclei in English. *J Acoust Soc Am* 32:693-703.

Pisoni DB, Carrell TD, Gans SJ (1983) Perception of the duration of rapid spectrum changes in speech and nonspeech signals. *Percept Psychophys* 34:314-322.

Potter RK, Steinberg JC (1950) Toward the specification of speech. *J Acoust Soc Am* 22:807-820.

Remez RE, Rubin PE, Nygaard LC, Howell WA (1987) Perceptual normalization of vowels produced by sinusoidal voices. *J Exp Psychol [Hum Percept]* 13:40-61.

Repp BH (1982) Phonetic trading relations and context effects: New evidence for a speech mode of perception. *Psychol Bull* 92:81-110.

Repp BH, Healy AF, Crowder RG (1979) Categories and context in the perception of isolated steady-state vowels. *J Exp Psychol [Hum Percept]* 5:129-145.

Sato M, Lotto AJ, Diehl RL (2003) Patterns of acoustic variance in native and non-native phonemes: The case of Japanese production of /r/ and /l/. *J Acoust Soc Am* 114:2392.

Saul AB, Cynader MS (1989) Adaptation in single units in visual cortex: the tuning of aftereffects in the spatial domain. *Vis Neurosci* 2:593-607.

Scott SK, Wise RJS (2003) Functional imaging and language: A critical guide to methodology and analysis. *Spe Com* 41:7-21.

Stephens JDW, Holt LL (2003) Preceding phonetic context affects perception of non-speech sounds. *J Acoust Soc Am* 114:3036-3039.

Stevens EB, Kuhl PK, Padden DM (1988) Macaques show context effects in speech

perception. *J Acoust Soc Am* 84(Suppl. 1):577.

Stevens KN (1986) Models of phonetic recognition II: A feature-based model of speech recognition. In: Mermelstein P (ed) *Proceedings of the Montreal Satellite Symposium on Speech Recognition*, pp. 67-68.

Stevens KN (2002) Toward a model for lexical access based on acoustic landmarks and distinctive features. *J Acoust Soc Am* 111:1872-1891.

Story BH (2005) A parametric model of the vocal tract area function for vowel and consonant simulation. *J Acoust Soc Am* 117:3231-3254.

Story BH, Titze IR (2002) A preliminary study of vowel quality transformation based on modifications to the neutral vocal tract area function. *J Phonet* 30:485-509.

Strange W (1989) Evolving theories of vowel perception. *J Acoust Soc Am* 85:2081-2087.

Sullivan SC, Lotto AJ, Diehl RL (2005) Optimal auditory categorization on a single dimension. In: Forbus K, Gentner D, Regier T (eds) *Proceedings of the Twenty-Sixth Annual Conference of the Cognitive Science Society*. Mahwah, NY: Lawrence Erlbaum Associates Inc, pp. 1639.

Summerfield Q (1981) Articulatory rate and perceptual constancy in phonetic perception. J Exp Psychol [Hum Percept] 7:1074-1095.

Summerfield Q, Assmann PF (1987) Auditory enhancement in speech perception. In: Schouten MEH (ed) NATO Advanced Research Workshop on The Psychophysics of Speech Perception. Dordrecht, Netherlands: Martinus Nijhoff Publishers, pp 140-150.

Summerfield Q, Assmann PF (1989) Auditory enhancement and the perception of concurrent vowels. Percept Psychophys 45:529-536.

Summerfield Q, Haggard M, Foster J, Gray S (1984) Perceiving vowels from uniform spectra: Phonetic exploration of an auditory aftereffect. Percept Psychophys 35:203-213.

Sussman HM (1986) A neuronal model of vowel normalization and representation. Brain Lang 28:12-23.

Syrdal AK, Gopal HS (1986) A perceptual model of vowel recognition based on the auditory representation of American English vowels. J Acoust Soc Am 79:1086-1100.

Trautmüller H (1981) Perceptual dimension of openness in vowels. J Acoust Soc Am 69:1465-1475.

Ulanovsky N, Las L, Nelken I (2003) Processing of low-probability sounds by cortical

neurons. *Nat Neurosci* 6:391-398.

Ulanovsky N, Las L, Farkas D, Nelken I (2004) Multiple time scales of adaptation in auditory cortex neurons. *J Neurosci* 24:10440-10453.

Viemeister NF (1980) Adaptation of masking. In: van den Brink G, Bilsen FA (eds) *Psychophysical, Physiological, and Behavioural Studies in Hearing*. Delft, Netherlands: Delft University Press, pp. 190-199.

Viemeister NF, Bacon SP (1982) Forward masking by enhanced components in harmonic complexes. *J Acoust Soc Am* 71:1502-1507.

Wade T, Holt LL (2005) Effects of later-occurring nonlinguistic sounds on speech categorization. *J Acoust Soc Am* 118:1701-1710.

Wade T, Holt LL (2005) Perceptual effects of preceding non-speech rate on temporal properties of speech categories. *Percept Psychophys* 67:939-950.

Watkins AJ (1988) Spectral transitions and perceptual compensation for effects on transmission channels. In: Ainsworth W, Holmes J (eds) *Proceedings of the 7th Symposium of the Federation of Acoustical Societies of Europe: Speech '88*, pp. 711-718.

Watkins AJ (1991) Central, auditory mechanisms of perceptual compensation for

spectral-envelope distortion. *J Acoust Soc Am* 90:2942-2955.

Watkins AJ, Makin SJ (1994) Perceptual compensation for speaker differences and for spectral-envelope distortion. *J Acoust Soc Am* 96:1263-1282.

Watkins AJ, Makin SJ (1996) Some effects of filtered contexts on the perception of vowels and fricatives. *J Acoust Soc Am* 99:588-594.

Wayland SC, Miller JL, Volaitis LE (1994) The influence of sentential speaking rate on the internal structure of phonetic categories. *J Acoust Soc Am* 95:2694-2701.

Whalen DH (1990) Coarticulation is largely planned. *J Phonet* 18:3-35.

Yamada RA, Tohkura Y (1990) Perception and production of syllable-initial English /r/ and /l/ by native speakers of Japanese. In: *Proceedings of the International Conference on Spoken Language Processing*, pp. 757-760.

Figure 10.1.

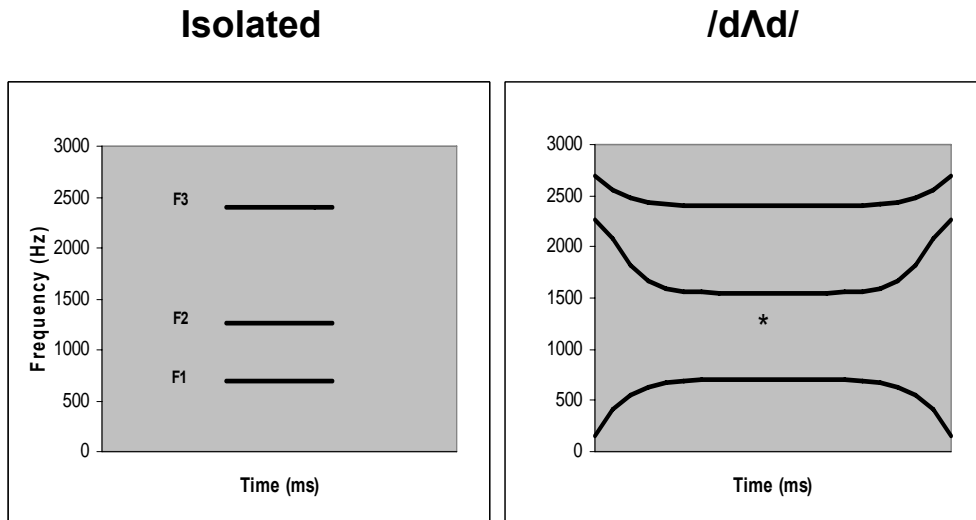


Figure 10.2.

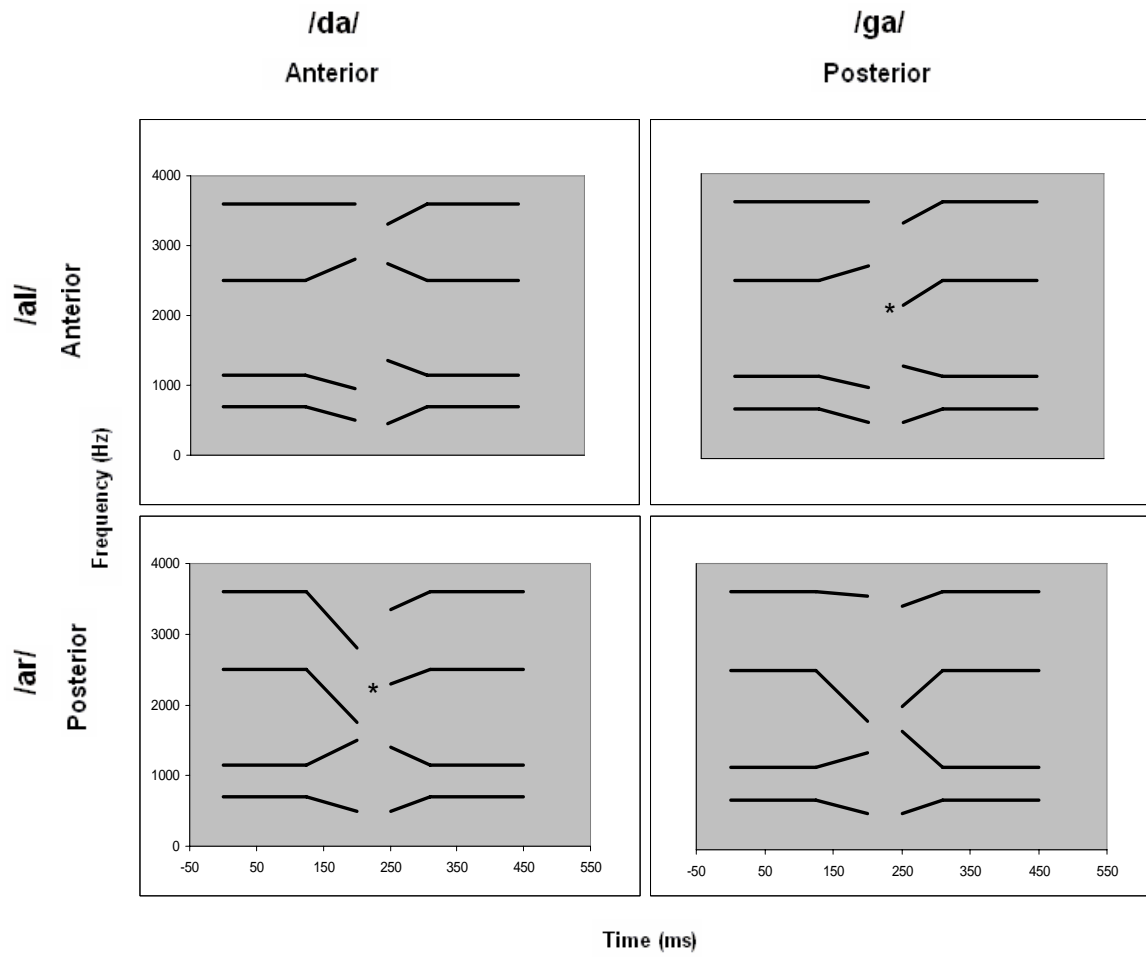


Figure 10.3.

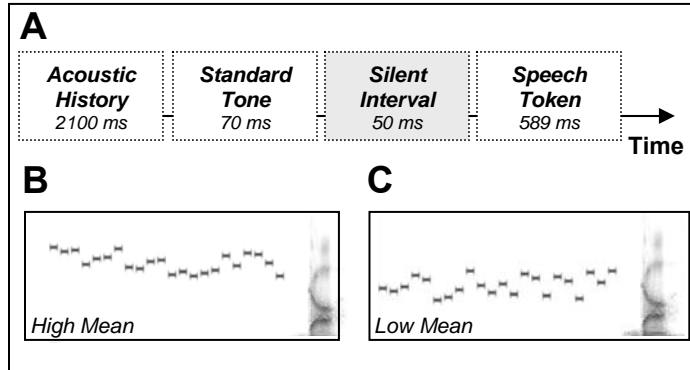


Figure Captions

Figure 10.1. Formant tracks for a vowel spoken in isolation and in a /d_d/ context. The star indicates the frequency for the second formant when produced in isolation. The shift in frequency for this formant in context demonstrates coarticulation.

Figure 10.2. Schematic spectrograms representing the frequencies of the first four formants for productions of /ga/ and /da/ in the context of /al/ and /ar/. The stars indicate third formant onset values that are nearly equivalent resulting in syllables that are acoustically similar but represent different phonetic categories.

Figure 10.3. Description of the stimulus paradigm of Holt (2005). A) time interval for each stimulus event; B) and C) spectrograms of two stimuli with tones sampled from either a high-mean (B) or a low-mean (C) distribution.