A Critical Assessment of Speech as a Sound Source

By

Dr. Henry W. BRAIN
Faculty of Arts
Golden Gate University, California
The United States of America

ABSTRACT

Spoken language communication is arguably the most important activity that distinguishes humans from non-human species. This paper provides an overview of speech as a sound source.While many animal species communicate and exchange information using sound, humans are unique in the complexity of the information that can be conveyed using speech, and in the range of ideas, thoughts and emotions that can be expressed. 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 study.

KEYWORDS: Language, communication, Speech, Sound Source

Introduction

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.

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).

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.

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.

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 /□/ (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 /□/. 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 inter-dependence 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 /□/ in "dud" resemble the values for the vowel /□/ (the vowel in bet) spoken in isolation

(Lindblom1963; Nearey 1989). Thus, the approach of defining vowels simply by their formant frequencies is thwarted.

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 /□/ to a good /□/, in either isolation or /d_d/ context. More /□/ 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 /□/ in this context. Several other examples of apparent compensation for coarticulation have been examined (see Repp 1982 for a review).

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).

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.

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).

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 /□/). 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 Volaitis1994). 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.

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 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 (Naatanen, Gaillard, and Mantysalo1978). 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.

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.

Conclusion

The ability to not just create a sound, but to make a speech is a unique attribute of mankind. From the study, it was concluded that 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. When one refers to speech perception, the task that comes to mind is the determination of the linguistic message intended by the speaker. The best evidence that speech effects cannot be accounted for solely by peripheral interactions is that context can affect preceding targets. 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. One salient difference between talkers is speaking rate.

Recommendations

1. 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.

2. Since message identification poses great problem to listeners, speech must be communicated in its simplest form to ease message recognition and interpretation.

3. The educational system should ensure that the range of ideas, thoughts and emotions that are expressed by students are broadened and unrestricted, as this will boost their communication performances and their ability to convey meaning in speech.

4. The government must provide all necessary assistance/aid to schools, to ensure that communication is effective in order to ensure that useful information is being conveyed through speech.

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

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

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

Coady, J. A, Kluender, K. R, &Rhode W. S. (2003).*Effects of contrast between onsets of speech and other complex spectra*. J AcoustSoc 114:2225-2235.

Delgutte, B., Hammond, B. M, Kalluri, S., Litvak, L. M, &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, M. L, Brittan-Powell E. F., Dooling, R. J., &Pierce A (1997) *Perception of synthetic /ba/ /wa/ speech continuum by budgerigars (Melopsittacusundulatus)*. J AcoustSoc 102:1891-1897.

Diehl, R. L, &Walsh, M. A. (1989) *An auditory basis for the stimulus-length effect in the perception of stops and glides*. J AcoustSoc Am 85:2154-2164.

Diehl, R. L., Souther, A. F, &Convis C. L. (1980) *Conditions on rate normalization in speech perception*. Percept Psychophys 27:435-443.

Gaskell, G.&Marslen-Wilson, W. D. (1996) *Phonological variation and inference in lexical access*. J ExpPsychol [Hum Percept] 22:144-158.

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

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

Goldinger, S. D. (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.

Green, D. M., McKay, M. J., &Licklider, J. C. R. (1959) *Detection of a pulsed sinusoid in noise as a function of frequency*. J AcoustSoc Am 31:1446-1452.

Healy, A. F.,&Repp, B. H. (1982) *Context independence and phonetic mediation in categorical perception*. J ExpPsychol [Hum Percept] 8:68-80.

Holt, L. L. (2005) *Temporally non-adjacent non-linguistic sounds affect speech categorization*. PsycholSci 16:305-312.

Holt, L. L. (in press) *The mean matters*: Effects of statistically-defined non-speech spectral distributions on speech categorization. J AcoustSoc Am.

Holt, L. L., Lotto, A. J., &Kluender, K. R. (2000) *Neighboring spectral content influences vowel identification*. J AcoustSoc Am 108:710-722.

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.

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

Jusczyk, P. W., Pisoni, D. B., 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, G. R. (1989) *Articulatory-rate context effects in phoneme identification*. J ExpPsychol [Hum Percept] 15:736-748.

Kiefte, M., &Kluender, K. R. (2001) *Spectral tilt versus formant frequency in static and dynamic vowels*. J AcoustSoc Am 109:2294-2295.

Kluender, K. R., Lotto, A. J., Holt, L. L.&Bloedel, S.L. (1998) *Role of experience for language specific functional mappings of vowel sounds*. J AcoustSoc Am 104:3568-3582.

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

Ladefoged, P. & Broadbent, D. E. (1957) *Information conveyed by vowels*. J AcoustSoc Am 29:98-104.

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

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

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

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

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

Lotto, A. J., Kluender, K. R. & Holt, L. L. (1998) *Perceptual compensation for coarticulation by Japanese quail (Coturnixcoturnix japonica).* J AcoustSoc Am 102:1134-1140.

Lotto, A. J., Sullivan, S. C. &Holt, L. L. (2003) *Central locus for nonspeech context effects on phonetic identification*. J AcoustSoc Am 113:53-56.

Lutfi, R. A. (2001) *Auditory detection of hollowness*. J AcoustSoc Am 110:1010-1019.

Lutfi, R. A. & Oh, E. L. (1997) *Auditory discrimination of material changes in a struck-clamped bar*. J AcoustSoc Am 102:3647-3656.

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

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

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

Naatanen, R., Gaillard, A. W. &Mantysalo, S. (1978) *Early selective attention effect on evoked potential reinterpreted*. ActaPsychol 42:313-329.

Nearey, T. M. (1989) *Static, dynamic, and relational properties in vowel perception*. J AcoustSoc Am 85:2088-2113.

Palmer, A. R., Summerfield, Q. &Fantini, D. A. (1995) *Responses of auditory-nerve fibers to stimuli producing psychophysical enhancement*. J AcoustSoc Am 97:1786-1799.

Pisoni, D. B. Carrell, T. D.&Gans, S. J. (1983) *Perception of the duration of rapid spectrum changes in speech and nonspeech signals*. Percept Psychophys 34:314-322.

Potter, R. K. & Steinberg, J. C. (1950) *Toward the specification of speech*. J AcoustSoc Am 22:807-820.

Repp, B. H. Healy, A. F.&Crowder, R. G (1979) *Categories and context in the perception of isolated steady-state vowels*. J ExpPsychol [Hum Percept] 5:129-145.

Sato, M., Lotto, A. J. & Diehl, R. L. (2003) *Patterns of acoustic variance in native and non-native phonemes*: The case of Japanese production of /r/ and /l/. J AcoustSoc Am 114:2392.

Saul, A. B. &Cynader, M. S. (1989) *Adaptation in single units in visual cortex: the tuning of aftereffects in the spatial domain*. Vis Neurosci 2:593-607.

Scott, S. K.&Wise, R. J. S. (2003) *Functional imaging and language*: A critical guide to methodology and analysis. Spe Com 41:7-21.

Stephens, J. D. W. & Holt, L. L. (2003) *Preceding phonetic context affects perception of non-speech sounds*. J AcoustSoc Am 114:3036-3039.

Stevens, E. B. Kuhl, P.K. &Padden, D. M. (1988) *Macaques show context effects in speech perception*. J AcoustSoc Am 84(Suppl. 1):577.

Story, B. H. (2005) *A parametric model of the vocal tract area function for vowel and consonant simulation*. J AcoustSoc Am 117:3231-3254.

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

Sullivan, S. C., Lotto, A. J. & Diehl, R. L. (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 ExpPsychol [Hum Percept] 7:1074-1095.

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

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.

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, N. F. (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, N. F. & Bacon, S. P. (1982) *Forward masking by enhanced components in harmonic complexes*. J AcoustSoc Am 71:1502-1507.

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

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

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

Watkins, A. J. (1991) *Central, auditory mechanisms of perceptual compensation for spectral-envelope distortion*. J AcoustSoc Am 90:2942-2955.

Watkins, A. J. & Makin, S. J. (1994) *Perceptual compensation for speaker differences and for spectral-envelope distortion*. J AcoustSoc Am 96:1263-1282.

Watkins, A. J. & Makin, S. J. (1996) *Some effects of filtered contexts on the perception of vowels and fricatives*. J AcoustSoc Am 99:588-594.

Wayland, S. C., Miller, J. L. &Volaitis, L. E. (1994) *The influence of sentential speaking rate on the internal structure of phonetic categories*. J AcoustSoc Am 95:2694-2701.

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