Topics in speech perception

Diane Kewley-Port
Department of Speech and Hearing Sciences, Indiana University, USA
https://doi.org/10.36505/ExLing-2006/01/0005/000005

Abstract

The study of speech perception over the past 60 years has tried to determine the human processes that underlie the rapid understanding of fluent speech. A first step was to determine the units of speech that could be experimentally manipulated. Years of examining the acoustic properties associated with phonemes led to theories such as the Motor Theory which postulate larger units that integrate vowel and con-sonant information. Current approaches find better support for the syllable as the most robust and coherent unit of speech. A complete theory of speech perception should systematically map how speech acoustic information is processed bottom-up through the peripheral and central auditory system, as well as how linguistic knowl-edge interacts top-down with the acoustic-phonetic information to extract meaning.

Introduction

The goal of the study of speech perception is to understand how fluent speech in typical environments is processed by humans to extract the talker's intended message. Towards this goal, the present overview will address three topics: (1) What are the units of speech perception? (2) How is speech processed from the peripheral to central auditory system? (3) What are the effects of the enormous variability observed in speech on speech perception?

Units of speech

Consider a fluent spoken sentence, such as "But that explanation is only partly true" (from TIMIT, Garfolo et al., 1993) recorded in the quiet (Fig. 1). The observed rapidly changing spectrotemporal properties in the spectrogram are typical of normal speech and permit a high transmission rate of information between human beings. What is even more remarkable is that communication does not usually take place in quiet, but rather in listening environments that are noisy or reverberant or have competing talkers, or in all three degrading circumstances, and yet speech understanding remains high. What do we know about how humans perceive speech?

A primary theoretical issue in speech perception is to determine the units of speech that are essential to describe human communication. Given a particular unit, various experiments can be conducted to manipulate speech and

examine the resulting perceptual consequences. Writing systems have relatively clear units such as alphabets (phonemes, Roman alphabet), syllables (Japanese hiragana syllabary) and words (Mayan) to represent graphically some of the information found in spoken language. Linguists have additionally postulated feature systems (Jacobson, Fant and Halle, 1952; Chomsky and Halle, 1968) as the basic units of speech. Thus although speech generally consists of multiword sequences (phrases and sentences), the largest unit typically used to represent speech is the word. For example, a variety of computer-based speech recognition systems have been designed to identify whole words, and those that identify words in isolation are considerably more successful than continuous word recognition in sentences such as those in Figure 1 (Lippmann, 1997).

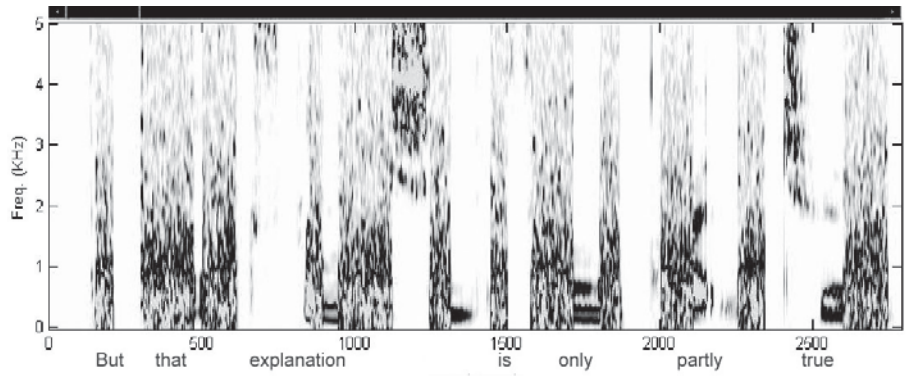

Figure 1. Spectrogram of a sentence with text roughly aligned in time.

The point of view taken in this overview of speech perception is that good experimental support must be demonstrated for postulated units of speech. Consider linguistic features. Stevens and his colleagues (1998) have long studied the acoustic properties of features such as place of articulation (Stevens and Blumstein, 1978) or sonorant/nonsonorant (/n/ versus /d/) to demonstrate that some of these properties are invariant across considerable talker variability. Recently Stevens (2002) has proposed a model that specifically states that lexical access from speech is based on the processing of feature bundles that are structured into phoneme segments. The primary evidence against this approach using discrete units is found in the substantial acoustic effects of coarticulation between segments in fluent speech (Liberman et al., 1967; Diehl et al., 2004), as well as the influence of the temporal properties of speech on linguistic categories (Port and Leary, 2005). Moreover, the details of speech acoustics required for a feature-based model such as Stevens (2002) are generally only available in quiet conditions. As noted above, human speech perception is robust under substantial

amounts of noise. This is because the speech signal is highly redundant and therefore speech perception only requires partial information to successfully extract the intended message. In the past ten years a great deal of research on speech processed through simulated cochlear implants has demonstrated that only a small number of frequency channels, four to seven, are needed to recognize sentences (Shannon et al., 1996; Dorman et al., 1997). In fact, in the extreme Remez and his colleagues (Remez et al., 1981, 1994) have demonstrated that sentences can be recognized from only three frequency modulated tones that approximate the first three formants of speech (sinewave speech), even though each individual tone sounds nothing like speech.

Given the strong evidence against discrete features or segments as being the units of speech perception, what is the alternative? There is a long history of proposing that the primary unit of speech is the gesture, starting with Liberman and his colleagues who postulated the Motor Theory of Speech Perception (Liberman et al., 1967) as based on CV units. This was followed at Haskins by Fowler (1984) whose Direct Realist Theory also referenced the speech gesture as the basic unit, but one described as having the vowel and consonant information coproduced and perceived. Additional support also from Haskins has been given by the speech production research of Browman and Goldstein (1992) whose theory of Articulatory Phonology provides details about the organization of consonants and vowels into coordinated gestures. How do these models of articulation and speech production relate to speech perception? As Browman and Goldstein (1988) initially argued, and as Studdert-Kennedy (1998) clearly states, the central unit of speech is the syllable. The syllable is the smallest unit in which the acoustic properties and temporal structure of speech are integrated into a coherent structural unit with the underlying articulatory gesture. This syllabic unit has properties related directly to other units larger (words) and smaller (features, phonemes) than the syllable. However, the syllable is the central unit that has structural descriptions that correspond across speech production, speech perception, the motor control of articulation, stored cognitive categories of speech and even language acquisition by infants (Studdert-Kennedy, 1998). Studdert-Kennedy (1998; Studdert-Kennedy and Goldstein, 2003) has argued that the relation between the syllable and other units in speech can be described in terms of the *particulate principle* in which the combination of smaller structures creates a functionally different set of objects at the next higher level. Of special importance to the view of speech perception described here is the fact that the strong coherence of acoustic information across frequency and time in syllables means that syllables can still be perceived when only fragmentary information is available, for example due to strange processing schemes (cochlear implant simulators) or noisy conditions.

Peripheral and central mechanisms in speech processing

Whatever linguistic units are the basis of speech perception, processing of the acoustic signal starts at the auditory periphery. Kewley-Port and her colleagues have attempted to describe processing of vowels using psychophysical methods to understand how the acoustic signal is represented at the most peripheral levels of the auditory system (Kewley-Port, 1991; Kewley-Port and Watson, 1994), and then describe how more central levels of linguistic processing interact with that information (Kewley-Port, 2001; Liu and Kewley-Port, 2004a). This research program began by establishing the smallest detectable difference (threshold) in a vowel formant between a standard vowel and a test vowel that can be discriminated under optimal listening conditions (after extensive training while listening to only one formant per block in the quiet). Results demonstrated that fine detail in the vowels is represented in the peripheral auditory system. Threshold differences across vowels and talkers (Kewley-Port and Zheng, 1999), and in quiet and noise (Liu and Kewley-Port, 2004b) can be modelled using loudness patterns derived from computational models for simple non-speech stimuli developed by Moore and Glasberg (1997). The conclusion of this research is that the first stages of processing of vowels yield considerable more detail about these complex, harmonic spectra than is needed to categorize vowels.

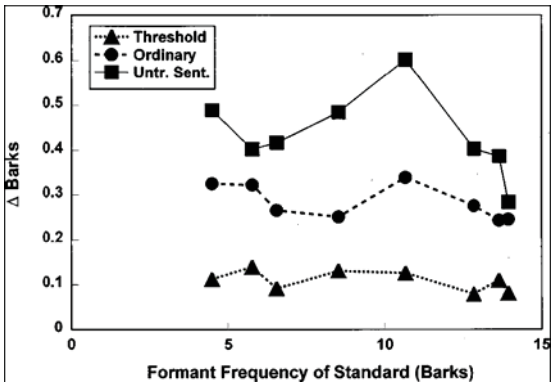

Figure 2. Thresholds for formant discrimination in Δ barks are displayed as a function of the center frequency of F1 and F2 for eight vowels. The function labelled "Threshold" is for discrimination of isolated vowels under optimum listening conditions. The function labelled "Ordinary" is for vowels embedded in phrases and sentences. The function labelled "Untr. Sent." is for the same listeners as for Ordinary, but for the Δ bark values obtained in the first half hour of testing, before listeners were trained (after Kewley-Port and Zheng, 1999).

As more variability is included in the vowel stimuli or the task complexity increases, more central levels of processing are required to perform the vowel discrimination task. In Fig. 2 the baseline vowel thresholds (in barks) under optimal conditions are shown to be relatively constant at about 0.11 barks. In a study exploring vowels in more ordinary listening conditions (Kewley-Port and Zheng, 1999) where many vowel formants were tested in phrases and sentences, higher levels of linguistic context elevated thresholds (labelled Ordinary) by a factor of 3 to 0.33 barks. However, formant discrimination (labelled Untr. Sent.) that was measured in sentences before the subjects were trained was elevated by a factor of 5.

Our results from vowel discrimination and identification tasks demonstrate that the information available about vowels in the periphery becomes degraded as more central processes are needed to process sentences or to learn new tasks. However, when the well learned task of identifying the words in sentences was added to the discrimination task in sentences (Liu and Kewley-Port, 2004a), vowel thresholds for discrimination remained similar, just above the 0.33 bark threshold measured under more ordinary listening conditions. The implication is that when adults listen to their native language, auditory processing of vowels has a threshold norm of one-third of a bark that represents the human ability to extract critical vowel spectral information in fluent sentences. This norm limits the bottom-up processing capabilities for vowel spectra. However, predictive information from top-down processing may enhance listeners' abilities to categorize vowels, as well as visual information from the face. A complete picture of speech perception needs to establish a systematic relation between peripheral and central mechanisms for processing consonants and vowels both in syllables and in sentences.

Variability in speech

A hallmark of spoken language is the large amount of variability observed in speech. For example, if different talkers in different environments all spoke the same sentence, normal native listeners would all write down the same sentence in spite of the high variability in the acoustic signal. This "nonlinguistic" variability includes considerable information about the speaker, referred to as the indexical properties of speech (gender, emotion, Nygaard and Pisoni, 1998), as well as speaking style variation (rate, "clear speech", Ferguson, 2004), and the cross-linguistic interference of accented speech (Flege, 1988). In the scenario above, little of this variability is preserved in the written transcription of the sentence, i.e. this nonlinguistic information is stripped off. But is it correct to treat this variability as random noise? The

clear answer is no for at least two reasons. First, listeners clearly use the indexical properties of speech in the give-and-take of every day conversation. Moreover, evidence has accumulated that this information can be stored as part of the representation of words in memory (episodic memory, Goldinger, 1998). Perhaps more important in normal discourse is that different speaking styles and rates affect the successful transmission of the intended message to the listener. Thus what has been considered nonlinguistic variability in speech can be manipulated for the purposes of improving speech intelligibility (Ferguson, 2004), and therefore represents structured information, and not random noise, in speech.

And finally, after this brief overview of many factors found to be important to understanding speech perception over the past 60 years, let's consider whether or not a comprehensive theory of speech perception is possible, at least in the near future. Stevens (2002) clearly believes that his theory is close to describing perceptual processes that span cognitive mechanisms representing the fine detail of speech in features through the retrieval of the associated words in the lexicon. However, the arguments proposed here suggest that this type of discrete unit model is not an adequate approach to understanding the mechanisms of speech perception. Rather, the approach taken by Studdert-Kennedy (1998) that uses the particulate principle for describing the structures of human behaviour is more likely to succeed. That is, we should agree that fine detail in speech may be captured by acoustic features as shown by Stevens (2002), but also acknowledge that this detail is restructured into higher level objects that have inherently different properties than feature bundles have by themselves. The particulate principle approach suggests that the syllable is the central unit that provides the most coherent relations between the structures of other units, both smaller and larger than the syllable. Whether or not this is true, our knowledge is incomplete for describing the relation between these units in the quiet, and research on the robustness of speech in noise (the typical environmental condition) is in its infancy. In fact, mechanisms for processing speech under the variety of adverse circumstances that humans encounter may differ substantially from one another (e.g. is listening in noise the same as trying to understand accented speech?). Building more comprehensive models of speech perception will require much more research.

Acknowledgements

Preparation of this manuscript supported by NIHDCD-02229.

References

Browman, C.P. and Goldstein, L. 1988. Some Notes on Syllable Structure in Articulatory Phonology. Phonetica 45, 140-155.

Browman, C. and Goldstein, L. 1992. Articulatory phonology: an overview. Phonetica. 49, 155-80

Chomsky, N. and Halle, M. 1968. The sound pattern of English. New York:Harper and Row.

Diehl, R., Lotto, A. and Holt, L. 2004. Speech Perception. Annu. Rev. Psychol. 55, 149-179.

Dorman, M.F., Loizou, P.C. and Rainey, D. 1997. Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs. J. Acous. Soc. Am. 102, 2403-2410.

Fowler, C.A. 1984. Segmentation of coarticulated speech in perception. Percept. & Psychophy. 36, 359-368.

Garofolo, J., Lamel, L., Fisher, W, Fiscus, J., Pallett, D., and Dahlgren, N. 1993. DARPA TIMIT: Acoustic-Phonetic Continuous Speech Corpus.

Goldinger, S. 1998. Echoes of echoes? An episodic theory of lexical access. Psych. Rev., 105, 251-279.

Goldstein, L. and Fowler, C.A. 2003. Articulatory phonology: A phonology for public language use. In Schiller, N.O. and Meyer, A.S. (eds.), Phonetics and Phonology in Language Comprehension and Production, 159-207. Mouton de Gruyter.

Jakobson, R., Fant, G., and Halle, M. 1952. Preliminaries to speech analysis: The distinctive features. Cambridge, MA: MIT Press.

Kewley-Port, D. 1991. Detection thresholds for isolated vowels. J. Acoust. Soc. Am. 89, 820-829.

Kewley-Port, D. 2001. Vowel formant discrimination II: Effects of stimulus uncertainty, consonantal context and training. J. Acoust. Soc. Am. 110, 2141-2155.

Kewley-Port, D. and Watson, C.S. 1994. Formant-frequency discrimination for isolated English vowels, J. Acoust. Soc. Am. 95, 485-496.

Kewley-Port, D. and Zheng, Y. 1999. Vowel formant discrimination: Towards more ordinary listening conditions. J. Acoust. Soc. Am. 106, 2945-2958.

Liberman, A., Cooper, F., Shankweiler, D. and Studdert-Kennedy, M. 1967. Perception of the speech code. Psychol. Rev. 74, 431-461.

Lippmann, R. 1997. Speech recognition by machines and humans. Speech Com. 22, 1–15.

Liu, C. and Kewley-Port, D. 2004a. Vowel formant discrimination in high-fidelity speech. J. Acoust. Soc. Am 116, 1224-1233.

Liu, C. and Kewley-Port, D. 2004b. Formant discrimination in noise for isolated vowels. J. Acoust. Soc. Am. 116, 3119-3129.

Moore, B. C. J., and Glasberg, B. R. 1997. A model of loudness perception applied to cochlear hearing loss. Auditory Neurosci. 3, 289-311.

Nygaard, L. and Pisoni. D. 1998. Talker-specific perceptual learning in speech perception. Percept. & Psychophy. 60, 355-376.

Port, R. and Leary, A. 2005. Against formal phonology. Language 72, 927–964.
Remez, R.E., Rubin, P.E., Pisoni, D.B., and Carrell, T.D. 1981. Speech perception without traditional speech cues. Science 212, 947-950.
Remez, R.E., Rubin, P.E., Berns, S.M., Pardo, J.S., and Lang, J.M. 1994. On the Perceptual Organization of Speech. Psych. Rev. 101, 129-156.
Shannon, R., Zeng, F-G, and Wygonski, J. 1996. Altered temporal and spectral patterns produced by cochlear implants: Implications for psychophysics and speech recognition. J. Acoust. Soc. Am. 96, 2470-2500.
Stevens, K.N. 1998. Acoustic Phonetics. Cambridge, MA, MIT Press.
Stevens, K.N. 2002. Toward a model for lexical access based on acoustic landmarks and distinctive features. J. Acoust. Soc. Am 111, 1872-1891.
Stevens, K.N. and Blumstein, S. 1978. Invariant cues for place of articulation in stop consonants. J. Acoust. Soc. Am. 64, 1358-1368.
Studdert-Kennedy, M. 1998. The particulate origins of language generativity: from syllable to gesture. In: Hurford, J., Studdert-Kennedy, M., and Knight, C. (eds.), Approaches to the evolution of language, Cambrdge University Press, Cambridge, U.K.
Studdert-Kennedy, M. and Goldstein, L. 2003. Launching language: The gestural origin of discrete infinity. In Morten Christiansen and Simon Kirby (eds.), Language Evolution, 235-254 Oxford University Press, Oxford, U.K.