ISSUES IN SPEECH PERCEPTION

Dominic W. Massaro, Department of Psychology
University of Wisconsin, Madison, Wisconsin, 53706, USA

My goal in the present paper is to address what I believe to be some important issues in the study of perception of speech and nonspeech sounds. The issues are discussed in the framework of binary contrasts. The binary framework was deemed appropriate because of both linguistic precedent and limited psychological capacity. Some hierarchical organization of the issues is probably optimal but I have been reluctant to provide one; the reader can sort, add to, delete, and order the issues as she or he chooses.

Templates versus features

Speech sounds may be gestalt units that cannot be further analyzed or reduced in terms of other attributes. If speech consisted of a sequence of indivisible sounds, then speech analysis would be limited to some variation of a template matching scheme. For successful analysis, an additional template would be needed for every unique speech sound. Although this possibility may be linguistically and psychologically correct, it leaves the student very little to do beyond a general recording and tabulation.

Not only does the template matching scheme leave time on the student's hands, it is not very appealing to those of us who wish to impose simplicity and order upon Mother Nature (or Mother Tongue). Luckily, Jakobson and his colleagues of the Prague school successfully argued that phoneme units could in fact be further analyzed in terms of distinctive features that represent similarities and differences with respect to other phonemes. Given this theoretical perspective, it follows naturally that all of the phonemes of a language can be characterized in terms of a set of distinctive features. Feature analysis is appealing because it allows the units to be subjected to a more abstract classification.

Feature analysis is also preferred over template matching in the study of perception. Template matching schemes would not illuminate any perceived similarities or differences among speech sounds. The applicability of feature analysis proves useful in understanding the findings that two sounds are perceived as similar to one another or are in fact confused with one another to the extent they share the same features. Independent evidence for

features comes from well-known neurophysiological findings that individual cells in the cortex respond selectively to a class of stimuli that share a particular property, such as the direction of the frequency change in a sound. Feature analysis is a worthwhile enterprise as long as we sometimes remind ourselves that it must stop somewhere. When a set of descriptors is no longer analyzable we are left with miniature templates.

Binary versus continuous features

Although it is not unreasonable to describe a speech sound in terms of the degree to which a feature is present in the sound, Jakobson made the important assumption that distinctive features[1] were binary in that each feature is either present or absent in all-or-none fashion. Jakobson argued that "the dichotomous scale is superimposed by language upon the sound matter." The idea of binary features is appealing in terms of parsimony but most importantly in terms of ease of classification. The integration of binary information from two or more feature dimensions requires only logical conjunction of pluses and minuses. The elegance of binary classification is probably responsible for what might sometimes be viewed as an excessive observance of the principle.

In terms of speech perception, it seems more reasonable to assume that the listener has information about the degree to which each feature is present in the speech sound. This assumption of continuous rather than all-or-none featural information contrasts with the traditional view of binary features in linguistic theory. More recently, Chomsky and Halle and Ladefoged have allowed a multi-valued representation of featural information at the perceptual level. In our model, each feature is evaluated in terms of a fuzzy predicate that specifies the degree to which it is true that the sound has a particular feature. Given the fuzzy information passed on by feature evaluation, it is apparent that the integration of this information across several features is more complex than in traditional all-or-none classificatory schemes. Much of our work has supported the idea that features are combined in

---

(1) The reader should be reminded that the issue of binary versus continuous features is independent of other issues such as phonetic versus acoustic features. Accordingly, even though some examples are drawn from linguistic analyses, the use of features is intended to be general and not limited to one level of analysis.

terms of a multiplicative rule. This combinatorial process is extremely simple but has the nice consequence that the less ambiguous features carry more weight.

Phonetic versus acoustic features

It is readily transparent that the concept of phonetic features has advanced the study of the linguistic classification of speech sounds. Students of speech perception must further inquire, however, whether speech perception is mediated by phonetic and/or acoustic features. The seminal work at Haskins Laboratories using synthetic speech evolved around the assumption that phonetic features were perceptually real. Many experiments were carried out to determine which acoustic properties of speech sounds were responsible for the perceived presence or absence of phonetic features. Given our analysis in the discussion of templates versus features it follows that the acoustic properties of speech sounds could be evaluated in terms of templates or features. If you agree that feature analysis is more desirable, then the speech perception theorist must be concerned with the analysis of speech sounds in terms of acoustic, not just phonetic, features.

Single factor versus multifactor experiments

In most experiments, speech sounds are varied along a single relevant dimension and observers are asked to perceive a given contrast between two sounds. For example, in the study of the acoustic features for a voicing contrast, all acoustic properties relevant to the contrast are made relatively natural except one, such as voice onset time, and this property dimension is varied through a continuum of values. Very few experiments independently vary more than one property within a particular experiment. The few exceptions in the early literature essentially reduced the data analysis to single-property experiments. In our work we utilize factorial designs and functional measurement techniques to study how acoustic features are evaluated and integrated together. With this procedure, two or more acoustic dimensions are independently varied so that all combinations of the values of one property are paired with all combinations of the values of another property. This design allows a direct assessment of how the acoustic features are evaluated and integrated together in speech perception.

Independent versus dependent features

This issue centers around whether the value for a given feature is modified by the value of another feature. Some support for featural independence was provided by studies demonstrating that separate sets of acoustic properties were relevant for perception of different contrasts. However, this result does not necessarily rule out the possibility that the perception of one contrast is dependent on the perception of another. Nonindependence has been proposed to account for the observed shifts in a voicing-contrast boundary as a function of a contrast in terms of place of articulation. However, these boundary shifts may occur even if each of the features makes independent contributions to the analyses. The observed interaction may result from the manner in which the independent featural information is integrated together. A quantitative model based on this idea has been successful in providing a quantitative account of boundary shifts and, therefore, the shifts do not imply nonindependence of feature evaluation.

Phoneme versus syllable units

Speech sounds of phoneme size have proven to be valuable in linguistic analysis. For the student of speech perception, however, it is important to ask what sound units are perceptually real. Although it is not easy to determine the sound units that are functional in speech perception, the question can be addressed simultaneously with the study of acoustic features in speech perception.

In our model, features are evaluated and matched to those features which define units in long-term memory. A unit is represented in long-term memory by a prototype which consists of a list of acoustic features. We assume that perceptual recognition of speech is mediated by vowel, consonant-vowel, or vowel-consonant syllable units in long-term memory. This assumption contrasts with the more commonly accepted notion of phonetic or phonemic prototypes in which phonetic or phonemic decisions mediate speech perception. Although it is only natural to say that a particular acoustic property cues voicing, the perception of the phonetic feature of voicing does not mediate syllable recognition in our model. Experiments that have evaluated the acoustic properties that are responsible for phonetic contrasts ask listeners to distin-

guish among speech segments of, at least, syllable length. These experiments do not necessarily mean that speech perception of the syllables was mediated by the phonetic contrasts defined by the experimenter.

In addition to the problem of the lack of acoustic invariance for some consonants, phoneme units cannot easily account for the finding that the vowel sometimes provides direct acoustic information about the consonant portion of a syllable. Vowel duration has a large effect on the voicing contrast of a vowel-consonant syllable in word-final position. Experimental and theoretical work in our laboratory supports the idea that acoustic features of the vowel portion and consonant portion are perceived independently, integrated together, and evaluated against syllable units in memory.

Stimulus versus process descriptions

Researchers are converging on the belief that there exists a plethora of potential acoustic features in speech perception. In contrast to the relatively small number of linguistic distinctive features, the potential candidates for acoustic features seem endless. Faced with this army of potential features, what might be the most valuable tack to take? Rather than attempting to define and catalog the large family of features, it might be more worthwhile to design prototypical experiments to assess how a small number of acoustic features are evaluated and integrated together in speech perception. The goal would be to develop a testable description of the process of speech perception rather than a complete stimulus description of all acoustic features. Needless to say, good judgment on the part of the speech researchers will allow a gradual accumulation of a stimulus description in their quest for understanding speech perception processes.

Acoustic versus contextual determinants

Speech perception research has been characterized by the study of speech perception as a function of acoustic changes in speech sounds. The researchers have not denied that other sources of information may also be exploited in perceiving natural speech. Not long after the investigator begins to understand how acoustic features are evaluated and integrated together in speech perception, it becomes necessary to assess how the processes work when

contextual influences are also available. As an example, feature evaluation and integration could be studied as a function of both acoustic changes in the speech signal and contextual constraints in terms of how likely a given sound may occur in a given context. A quantitative description of analogous experiments in reading supports the idea that contextual constraints simply provide an independent source of information exactly analogous to what would be provided by an additional feature.

Speech perception versus speech recognition

Upon reflection, it is apparent that speech recognition does not mirror speech perception. I recognize (and classify) two sounds as the same without necessarily perceiving them as identical. I believe that the idea of perceptual constancies has misled researchers in not only areas of visual perception but also in speech. The receding object is recognized as the same object even though the retinal input undergoes drastic changes. But the perception of the object also changes as is easily demonstrated by a little perceptual scrutiny. Following in the behavioristic tradition, researchers usually ask listeners to identify or classify sounds and take performance as an index of perception. Are we asking observers to make the stimulus error as the early introspectionists would claim or are there experimental tasks and performance measures that provide good indices of speech perception? This issue may help illuminate the general area of categorical perception by asking to what extent categorical perception is not categorical perception but simply categorical recognition.

To more directly tap perception, experimenters might employ continuous rather than discrete response alternatives. A discrete judgment may not be sensitive to the continuous changes in perception produced by continuous changes in an acoustic property of the speech sound. As an example, small increases in voice onset time for a velar stop might be perceived as making the sound more like /ki/. However, if the sound is still perceived as more like /gi/ than /ki/, the listener may always respond with /gi/. If the listener's judgments are consistent, the different sounds would be responded to equivalently even though they are perceived as different. By asking the observer to make a judgment on a continuum between the discrete alternatives, the responses may

more directly mirror perception. We have obtained orderly data from observers marking off a line in order to place the percept somewhere between discrete alternatives.

Speech perception versus speech understanding

It is easy to forget that speech perception does not necessarily entail speech understanding and that accurate understanding does not demand accurate speech perception. Consider a lexical decision task in which a listener indicates whether each test is a word or a nonword. The nonwords, such as "prust" and "mantiness", are perceived correctly and could be repeated even though no understanding takes place. I don't think that it would be profitable to argue that nonwords are not perceived. Our last noisy party reminds us that a significant amount of speech understanding can occur without perfectly accurate speech perception. In many highly constrained sentence contexts, the listener understands exactly some of the message before he perceives it. In fact, a few recent studies have provided some support for the idea that understanding can actually modify perception. A more convincing demonstration is how the perceived clarity of the words of a song is enhanced when the listener simultaneously reads them. In any case, it is necessary to distinguish between the case in which the listener resolves a piano sound sufficiently to distinguish it from adjacent sounds on the musical scale and the case in which the sound is also identified as middle C.

In our model, perception and understanding occur at two different stages of information processing. The primary recognition process evaluates and integrates acoustic features and outputs a perceptual experience of a speech sound. The secondary recognition process operates on the perceptual information to impose meaning and, therefore, a relatively abstract encoding. Although these are highly analogous processes, they utilize different categories of information in long-term memory and may be influenced by different properties of higher-order contextual constraints.

Speech versus nonspeech

It seems appropriate to close with this issue (or nonissue) since it is the topic of this symposium. Although speech represents language and nonspeech does not, it is important to know to what extent perception of speech is analogous to perception of

nonspeech. Does nonspeech perception derive from an evaluation and integration of acoustic features defined with respect to segments of sound? Remarkable parallels between speech and nonspeech have been reported in recent years. Rather than concluding that serious investigators should return to psychophysical studies of nonspeech in order to understand basic auditory processes, it seems more productive to assume that speech offers so much more for experimental study and that the most direct route to an understanding of auditory perception is to be found in the study of speech perception.[2]

References

Derr, M.A. and D.W. Massaro (1978): "The contribution of vowel duration, F_0 contour, and frication duration as cues to the /juz/-/jus/ distinction." WHIPP Report #8.

Massaro, D.W. (1978): "Letter information and orthographic context in word perception." Technical Report No. 453.

Massaro, D.W. (1975): (Ed.) Understanding language: An information processing analysis of speech perception, reading, and psycholinguistics. New York: Academic Press.

Massaro, D.W. and M.M. Cohen (1976): "The contribution of fundamental frequency and voice onset time to the /zi/-/si/ distinction", JASA, 60, 704-717.

Massaro, D.W. and M.M. Cohen (1977): "The contribution of voice-onset time and fundamental frequency as cues to the /zi/-/si/ distinction", Perc. Psych. 22, 373-382.

Massaro, D.W. and G.C. Oden (1978): "Evaluation and integration of acoustic features in speech perception", WHIPP Report #9.

Oden, G.C. (1978): "Integration of place and voicing information in the identification of synthetic stop consonants," JPh, in press.

Oden, G.C. and D.W. Massaro (1978): "Integration of featural information in speech perception", Psych. Rev. 85, 172-191.

---

(2) The research reported in this paper was carried out with the collaboration of Michael M. Cohen, Marcia A. Derr, and Gregg C. Oden and was supported in part by National Institute of Mental Health Grant MH 19399 and in part by the Wisconsin Alumni Research Foundation.