THE PERCEPTION OF OBSTRUENT CLUSTERS

Z.S. BOND

A number of studies, dealing with the perception of order of non-speech sounds, indicate that correctly perceiving the order of sounds of short duration is quite problematic. Hirsch (1959) reported that, after considerable practice, subjects could perceive the order of two sounds correctly if the onset of the sounds was separated by 15 to 20 msec. Hirsch concluded that the minimal temporal interval required for perception of order is independent of the quality of the sound and of the duration of the sound (within the limits of the experiment).

Broadbent and Ladefoged (1959) found that, at first, subjects could not perceive the order of sounds correctly unless the onset of the sounds was separated by 150 msec; with considerable training, a 30 msec separation became adequate for accurate perception of order.

Both of these experiments required the subjects to judge the order of only two stimuli. However, the task is much more difficult when the subjects have to determine the order of three or more stimuli. Several experiments involving the perception of order of more than two sounds are reported by Warren and Warren (1970). In the first experiment, subjects were asked to determine the order of three sounds, each lasting 200 msec, which were repeated over and over without pauses. The subjects performed no better than chance. When the order of four sounds was to be judged, the duration of each item had to be increased to between 300 and 700 msec for half of the subjects to identify the sequence correctly.

These experiments indicate that perceiving the order of acoustic events correctly is quite difficult. However, listeners to speech seem to have no comparable difficulty perceiving the order of the elements of speech, even though, for consonant clusters, the duration of the component elements comes close to the minimal separation which seems to be required for accurate perception of order. The problem, therefore, is why the perception of order of speech sounds is not more difficult for listeners than it appears to be.

The traditional view of speech perception is that speech is perceived in terms of phoneme-like units. If the perception of consonant clusters is considered from this point of view, the expectation is that a listener perceives a consonant cluster such as *sp*

by first identifying the *s* and then identifying the *p*; that is, the traditional view implies that consonant clusters are perceived 'phoneme-by-phoneme'. The accuracy with which listeners normally perceive speech can then simply be attributed to practice.

However, it is also possible, and has been suggested by a number of theorists, that some special mechanisms are employed in the perception of consonant clusters. On the basis of the Broadbent and Ladefoged (1959) and the Hirsch (1959) experiments, Neisser (1967) suggests that a listener gradually learns to distinguish a consonant cluster such as *ps* from a cluster like *sp*, rather than perceiving a sequence of two consonants in a certain order. He implies that consonant clusters are perceptual units, not normally analyzed into their components.

Recently, Wickelgren has suggested that speech may be perceived in terms of context-sensitive "allophones" (1969a, 1969b). Essentially, Wickelgren argues that speech is coded in sub-phonemic units, each unit being marked for what precedes and follows it. Thus, the order of the elements can be inferred from the elements themselves. Wickelgren's idea is readily applicable to the perception of consonant clusters: a listener would code a consonant cluster as an unordered sequence, with each element marked for what precedes and follows it. These elements would be assembled in the correct order at some further point in speech processing, and the listener would arrive at the intended sequence.

The perception of consonant clusters is an interesting problem for empirical study, particularly since it is related to the widely accepted notion that the typical unit in speech perception is the phoneme. By observing the pattern of perceptual confusions obtained for obstruent clusters, it is possible to make some inferences about the mechanisms underlying the perception of these clusters.

For this experiment, fifteen pairs of English words ending in obstruent clusters were selected to serve as stimuli. Five pairs of words ended in the obstruent clusters *ps* or *sp*; five ended in the clusters *ts* or *st*; five ended in the clusters *ks* or *sk*. A typical set of words employed in the experiment is *apse*, *asp*; *mats*, *mast*; *Max*, *mask*. Each stop-fricative cluster occurred with and without a morpheme boundary. Three lists, employing each of the words two times, were recorded at three different signal-to-noise ratios: O d.b., + 12 d.b., and − 6 d.b.

Nineteen subjects participated in the experiment. The subjects listened to the stimulus tape, and wrote what word they heard. In addition, five subjects listened to the tape a second time, giving spoken responses.

As is to be expected, the less intense the signal is, in relation to noise, the more errors occur. Generally, a stop-fricative cluster is more accurately reported than a fricative-stop cluster. Of more interest, however, are the resultant confusion patterns. Only confusion patterns obtained when the signal-to-noise ratio is − 6 d.b. will be presented.

Figure 1 shows the confusion patterns obtained for the six consonant clusters for both spoken and written responses. For all but one consonant cluster, the most common error is a simple reversal of the two consonants in the cluster. Errors that

involve only substitution are much less common than errors involving either simple reversal or reversal plus substitution, for all six consonant clusters.

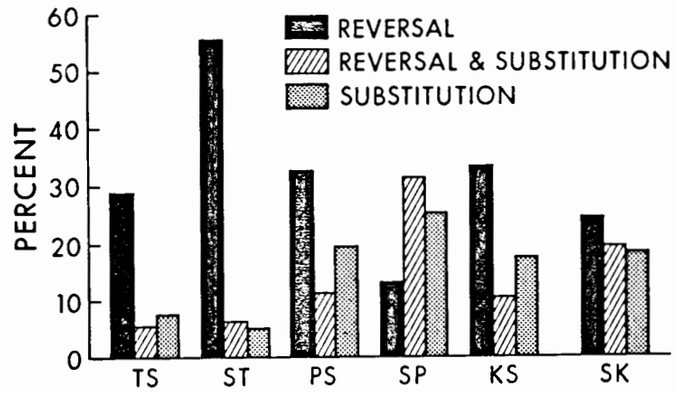

Fig. 1. Confusion Errors for All Responses.

Figure 2 shows the confusion pattern obtained for spoken responses. The confusion patterns are essentially the same as when all responses are considered together: errors involving reversal of elements predominate. Furthermore, there is no advantage for spoken responses; the percent of correct responses is approximately the same: 39 % for all responses, 41 % for written responses, and 33 % for spoken responses.

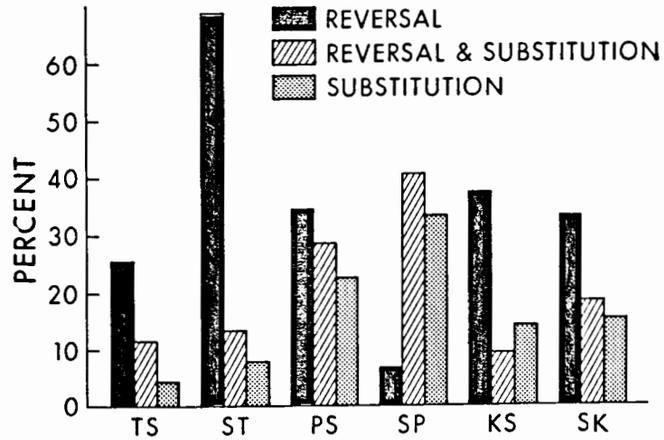

Fig. 2. Confusion Errors for Spoken Responses Only.

Figure 3 shows the confusion patterns obtained for words with a morpheme boundary intervening between the stop and the fricative. For two of these clusters,

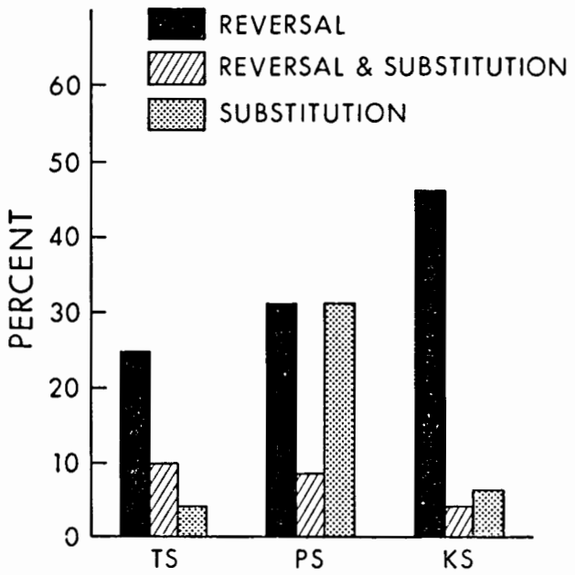

Fig. 3. Confusion Errors in the Presence of Morpheme Boundary.

ts and *ks*, reversal is still the most common perceptual error. For the *ps* cluster, however, substitution and reversal errors are almost equally likely.

The finding that has the most bearing on the perception of consonant clusters is that reversal errors are the most common errors. This finding is not consistent with the idea that the phoneme is the typical perceptual unit. If a listener perceives consonant clusters 'phoneme-by-phoneme', then, given that the segments occur in a particular order, there is no reason for the listener to reverse that order. Granted, he might on occasion forget the order, but there is no reason to suppose that he would be more likely to forget the order of the segments than to forget one of the segments; thus, reversal errors would be no more common than substitution errors. However, that is clearly not the case: reversal errors are much more common. This finding implies that some special perceptual mechanisms must be postulated for the perception of consonant clusters.

Neisser's suggestion that a consonant cluster is a perceptual unit, and Wickelgren's suggestion that a consonant cluster is coded in terms of some element very much like an allophone, are both compatible with the data.

If consonant clusters are perceptual units, then a *ps* cluster is most similar to an *sp* cluster. When the signal is degraded by the addition of noise, the items that are most similar to each other will be confused most; thus, reversal errors will be most likely.

If a consonant cluster is coded in terms of context-sensitive 'allophones', then the allophone of *s* before *p* will be slightly different acoustically from the allophone *s* after *p*. This difference, however, will be the most subtle part of the signal; particularly, it will be more subtle than the acoustic information differentiating consonants from each other. These small acoustic differences will be the first to disappear when the signal is degraded by noise; consequently, reversal errors will be the most common. Thus, either Neisser's or Wickelgren's suggestion will account for the observed result.

The essential point is that the sequence of consonants is not processed 'phoneme-by-phoneme', but that some other perceptual processing mechanisms must be postulated, dealing with at least a sequence of two consonants, without regard for the order of the consonants. Furthermore, the method of processing the consonants is unaffected by the presence of a morpheme boundary.

Department of Linguistics
Ohio State University, Columbus, Ohio, and
Department of Linguistics
The University of Alberta
Edmonton, Alberta, Canada

REFERENCES

Broadbent, D.E. and P. Ladefoged
1959 "Auditory Perception of Temporal Order", *Journal of the Acoustical Society of America* 31:1539.
Hirsch, I.J.
1959 "Auditory Perception of Temporal Order", *Journal of the Acoustical Society of America* 31:759-767.
Neisser, Ulric
1967 *Cognitive Psychology* (New York, Appleton-Century-Crofts).
Warren, R.M. and R.P. Warren
1970 "Auditory Illusions and Confusions", *Scientific American* (December):30-36.
Wickelgren, W.A.
1969a "Context-sensitive Coding, Associative Memory, and Serial Order in (Speech) Behavior", *Psychological Review* 76:1-15.
1969b "Context-Sensitive Coding in Speech Recognition, Articulation, and Development" in *Information Processing in the Nervous System*, K.N. Leibovic, ed. (New York, Springer-Verlag).

DISCUSSION

KIM (Urbana, Ill.)
I have only a few minor comments.

1. I don't think that anyone has claimed explicitly that the unit of speech perception is a phoneme. My feeling is that such a notion was implicit in everybody's mind due to the fact that phoneme was the cornerstone of structural linguistics.

2. You showed that there were differences in perceptual behavior among different consonant clusters. This is very interesting, because in the history of English, the

clusters behaved differently. For example, the initial consonant of *ps*, *kn*, etc., in such words as *psyche*, *knife*, was dropped but the dropping did not happen to other clusters. While one can speculate about the articulatory factors that caused this dropping, e.g., near simultaneous oral closures, perhaps the perceptual factors had something to do with the dropping as well.

3. You referred to Wickelgren's paper as a model of your data. If you were referring to his recent paper in *Psychological Review* 76.1-15 (1969), his context-sensitive associative-chain model that he describes there is a production model, not a perceptual model.

BOND

The assumption that a phoneme-like unit is the basic unit in speech perception seems to be common in psychology, particularly.

For the perception model, I am referring to Wickelgren "Context-Sensitive Coding in Speech Recognition, Articulation, and Development" (1969b, not 1969a).

MOL (Amsterdam)

I am sorry to say that I might cite the names of many authors who have openly stated that speech recognition comes down to identifying one phoneme after the other. They even declared this at phonetic congresses. I even fear that this view may be held by the majority of linguists.

BOND

I am very pleased to hear Prof. Mol's comment.

Particularly in the psychological literature on speech perception, the phoneme is considered to be a perceptual unit, but some phonetic literature can give the same impression; for example, the attempt to find invariant acoustic correlates of individual phonemes can be interpreted to imply that a hearer must process individual phonemes, in sequence.