DUPLEX PERCEPTION AND INTEGRATION OF CUES: EVIDENCE THAT SPEECH IS DIFFERENT FROM NONSPEECH AND SIMILAR TO LANGUAGE

Alvin M. Liberman, Haskins Laboratories, New Haven, Connecticut

The observations relevant to a comparison of speech and nonspeech -- the subject of our symposium -- can be divided, according to one's purposes, into several classes. As my contribution to our discussion, I would call your attention to two. In the one class are those research findings that can enlighten us about speech as a putative mode of perception, different phenomenologically from other aspects of audition. In the other class are those findings that permit us to see one of the possibly unique characteristics of speech as an instance of a correspondingly unique characteristic of language.

Phonetic perception as a mode. Perhaps the most direct way to observe the characteristic differences between perception in the phonetic and auditory modes is to contrive to have the same stimulus patterns perceived, either alternatively or simultaneously, as speech and nonspeech. The first reported example was by Lane and Schneider (1963). Using more or less unnatural synthetic speech that sampled the continuum of voice-onset-time from voiced to voiceless syllable-initial stops, they undertook to obtain discrimination functions under two conditions: in one, they told the subjects they would hear speech; in the other, arbitrary sounds of a complex sort. In the event, the discrimination functions were different: those obtained in the "speech" condition showed the usual peak at the phonetic boundary, while those from the other condition did not.

A more recent and thoroughgoing experiment of this general type has been carried out by Bailey, et al. (1977). Inasmuch as it is to be reported at this symposium, I will say no more about it.

In both of these studies, stimulus patterns were heard as speech or nonspeech, but not at the same time. Let us turn now to those cases in which the two percepts are experienced simultaneously. Such duplex perception was accomplished first by Rand (1974), and applied by him to perception of the transition cue for syllable-initial stops. Into one ear he put just the brief second- and third-formant transitions that are sufficient, in the appropriate context, to produce the perceived difference in place among [b], [d], and [g]. By themselves, these transitions sound more like a nonspeech 'chirp' than anything else. Into the other ear, he put the remainder of the pattern. Let us call it the 'base'. By itself, the base sounds more or less like a syllable. To some listeners, indeed, it appears to be a stop-vowel syllable but, having a fixed acoustic structure, it can, of course, produce only one stop, not all three. The interesting effect occurs when the transition cue and the base are presented dichotically (and in approximately the right time relationship), for then the listener fuses the two inputs so as to perceive the three coherent stop-vowel syllables he would have perceived had the two inputs been mixed electronically (and presented binaurally), while at the same time perceiving the 'chirp' he would have perceived had the transition cues been presented in isolation. (The chirp is heard in the ear to which the isolated transitions are presented; the syllable is heard in the other ear.) Thus, the same brain perceives the same stimulus input in two phenomenologically different ways, as speech and nonspeech, at the same time. This provides, at the very least, an excellent way to gain an impression of what the difference between phonetic and auditory modes sounds like.

Being interested in the possibility of using Rand's technique for the purpose of further and more nearly precise comparisons of the phonetic and auditory modes, I succeeded in reproducing the phenomenon, but experienced difficulty in getting it to be sufficiently stable to permit the further investigations I had in mind. Recently, however, David Isenberg and I (1978) have produced a duplex percept like that of Rand, but easier to hear, we think, and also, perhaps, more stable. We followed Rand's procedure, but changed the stimulus pattern. In particular, we thought it advisable to make the critical (and isolated) cue be the third-formant transition. The advantage, in that case, might be that the remainder of the pattern -- all of the first and second formants, plus the steady-state portion of the third formant -- would be quite full and speechlike. Accordingly, we chose the contrast [r] vs [l], putting the critical third-formant transition cue into one ear and the remainder (the base) into the other. It was quickly apparent that this arrangement did make it relatively easy to obtain the duplex percept. Indeed, tests with a number of listeners confirmed that they could "correctly" hear [r] or [l] (depending on which transition cue was presented), while simultaneously hearing the transition cue as a chirp.

Having thus found that the simultaneous fusion and separation of the two parts of the pattern (base and isolated transition) seemed to occur easily and consistently, we undertook further tests. The one I would briefly describe here was designed to assess the effects on the duplex percept of separately varying the intensity of the two stimulus components. We observed, first, that changing the intensity of the transition cue caused changes in the perceived loudness of the chirp, but no such changes in the fused [ra] or [la]; changes in the perceived loudness of the fused [ra] and [la] seemed to occur only as a result of variations in the intensity of the base.

To test this manifestation of duplexity more systematically, we carried out the following experiment. On each trial, we presented to the listener a sequence of dichotic pairs in which the transition component (in one ear) varied between the [r] cue and the [l] cue according to some predetermined order. Also, on each trial we varied the intensity of (1) the transition cue, or (2) the base, or (3) neither. The listener's task was to tell the order of [ra] and [la] syllables he heard, and, in addition, which loudnesses, if any, had changed. The results indicated that our listeners were, in fact, fusing the dichotic inputs to hear the 'correct' syllable, while at the same time dissociating the loudnesses by assigning the intensity of the transition cue to the chirp and the intensity of the base to the fused speech percept. We find this phenomenon interesting in its own right, but also as a basis for further investigation into the properties of the two components -- speech and nonspeech -- of the duplex percept. Consider, in that connection, an earlier study by Mattingly et al (1971) that compared the discrimination of a transition cue when, in the one case, it was presented in isolation and perceived as a chirp, and when, in the other, it was in its proper place in the speech pattern and cued a phonetic distinction. That study found differences in the discrimination pattern under the two conditions, but the interpretation of that finding

was subject to the reservation that the transition cue was, after all, in different contexts. By taking advantage of the duplex percept, we can, perhaps, obtain results that will avoid the need for that reservation and thus speak more straightforwardly to the difference between speech and nonspeech.

The integration of cues in speech and syntax. One of the most general characteristics of speech is that the information appropriate to a phonetic segment is typically contained in a numerous variety of cues; moreover, these are widely distributed through the signal and sometimes overlapped with cues for other phones. This is so because of the nature of articulation and coarticulation (Cooper, 1963; Fant, 1973): the various components of an articulatory gesture, distributed as they are in time, spread their acoustic consequences through the signal. Thus, the closing and opening gestures appropriate to an intervocalic stop affect the duration of the preceding syllable and also its offset, the occurrence and duration of an intervocalic silence, and the temporal and spectral characteristics of the onset of the following syllable. Conversely, and as a result of coarticulation, information about successive segments is often collapsed into a single acoustic segment and conveyed simultaneously, as in the case of most consonant-vowel syllables. Yet the speech processor somehow sorts the cues, as it were, assigning each to the appropriate part of the perceived phonetic structure. More to the point of our present purpose, it "integrates" into a unitary percept all the cues for a particular phone, no matter how various and widely distributed the cues may be (Liberman and Studdert-Kennedy, 1977; Repp et al, 1978; Bailey and Summerfield, 1978; and Dorman et al, 1978). It is difficult to see how this can be accomplished by ordinary auditory mechanisms, so we assume phonetic processes specialized for the purpose.

Consider, for example, the above-mentioned experiment by Repp et al. It dealt with perception of the utterance: "Did you see the gray (great) ship (chip)?" The variables of interest were (1) the nature of the next-to-last word, which was biassed either toward gray or great; (2) the duration of the silent interval between gray (great) and ship (chip), and (3) the duration of the fricative noise in ship (chip).

Let us now look first at the "forward" action of an earlier-occurring cue on a later-occurring one: given a perceptual boundary between ship and chip that varied according to the duration of the fricative noise and also the duration of the preceding silent interval, there was a further variation that depended, other things equal, on whether the preceding word was biassed toward gray or great. Now consider an effect in the opposite direction -- the "backward" action of a later-occurring cue on the perception of an earlier-occurring one. This was exemplified by the finding that the listener perceived gray or great depending, all else equal, on the duration of the fricative noise in ship; with other cues properly set, the listeners perceived 'gray' when the duration of the fricative noise (in the next syllable) was relatively short, but great when it was relatively long. Thus, the perception of gray could be changed to great by adding fricative noise in the syllable that followed the target word.

Apparently, the listeners in that experiment integrated into a unitary phonetic percept a variety of acoustic cues that stretched over at least two syllables and overlapped completely with cues relevant to other phones. But how does the listener do this? More specifically, how does he know when to stop integrating? Looking at the variety of cases of this type, we conclude that the integration period is marked neither by a temporal criterion (integrate every x msec), nor by an acoustic one (integrate every time a particular kind of sound is heard). Rather, the integration seems to occur over any stretch of the signal that contains the acoustic consequences of just those articulatory maneuvers that are the peripheral reflections of the speaker's intent to produce a particular phonetic segment. We must wonder, then, how the listener delimits the proper span over which to integrate, in what form he holds the pre-integrated cues, and what he does while waiting.

Consider now, though briefly, how analogous this is to what happens in the decoding of syntax. Surely, the meaning of a syntactic structure (e.g., a sentence) cannot be had except as the listener takes account of the words the structure comprises. As in the phonetic case, the size of this structure is not defined by a temporal criterion, nor by an acoustic one. Rather, it appears to be any number of words that are relevant to the

syntactic structure, and that depends, in turn, on the nature of the message the speaker means to convey. Here, too, then, we must wonder how the listener knows when the structure is complete, in which form he holds the words pending completion, and what he does while waiting.

References

Bailey, P.J., Q. Summerfield, and M. Dorman (1977): "On the identification of sine-wave analogues of certain speech sounds", Haskins Laboratories Status Report on Speech Research, SR-51/52, 1-25.

Bailey, P.J. and Q. Summerfield (1978): "Some observations on the perception of [s] + stop clusters", Haskins Laboratories Status Report on Speech Research, SR-53, Vol. 2, 25-60.

Cooper, F.S. (1963): "Speech from stored data", 1963 IEEE International Convention Record, Part 7, p. 139.

Dorman, M., L. Raphael, and A.M. Liberman (1978): "Some experiments on the sound of silence in phonetic perception", (submitted for publication).

Fant, G. (1973): "Descriptive analysis of the acoustic aspects of speech", Speech Sounds and Features, Ch. 2, 25-6. (Article based on a paper by Fant presented at a Wenner-Gren Foundation Research Symposium held at Burg Wartenstein, Austria, 1960, which appeared originally in Logos, 5, 3-17 (1962).)

Isenberg, D. and A.M. Liberman (1978): "Speech and nonspeech percepts from the same sound", JASA 64, Suppl. No. 1, J20.

Lane, H.L. and B.A. Schneider (1963): "Discriminative control of concurrent responses by the intensity, duration and relative onset time of auditory stimuli", unpublished report, Behavior Analysis Laboratory, University of Michigan.

Liberman, A.M. and M. Studdert-Kennedy (1977): "Phonetic perception", in Handbook of Sensory Physiology, Vol. VIII, "Perception." ed. by R. Held, H. Leibowitz, and H.L. Teuber; Heidelberg: Springer-Verlag, Inc.

Mattingly, I.G., A.M. Liberman, A.K. Syrdal, and T. Halwes (1971): "Discrimination in speech and nonspeech modes", Cogn. Psych. 2, 131-157.

Rand, T.C. (1974): "Dichotic release from masking for speech", JASA 55, 678-680.

Repp, B.H., A.M. Liberman, T. Eccardt, and D. Pesetsky (1978): "Perceptual integration of temporal cues for stop, fricative, and affricate manner". J. Exp. Psych.: Human Perception and Performance (in press).