SYMPOSIUM NO. 8: THE PERCEPTION OF SPEECH VERSUS NONSPEECH

(see vol. II, p. 431-489)

Moderator: David B. Pisoni[1]

Panelists: Anthony E. Ades, Pierre L. Divenyi, Michael F. Dorman, Dominic W. Massaro, and Quentin Summerfield

Chairperson: Arthur S. Abramson

DAVID B. PISONI's INTRODUCTION

Historically, the study of speech perception may be said to differ in a number of ways from the study of other aspects of auditory perception. First, the signals used to study the functioning of the auditory system were simple and discrete, typically varying along only a single physical dimension. By contrast, speech signals display very complex spectral and temporal relations. Although speech signals have also been varied along single physical dimensions, the perceptual consequences of such manipulation have not always followed from "equivalent" stimulations of a nonspeech nature. Alternatively, we may presume that the complexity of the spectral and temporal structure of speech and its variation is one additional source of perceptual differences between speech and nonspeech signals. Second, most of the research dealing with auditory psychophysics over the last thirty years has been concerned with the discriminative capacities of the sensory transducer and the functioning of the peripheral auditory mechanism. In the case of speech perception, however, the relevant mechanisms are assumed to be centrally located and intimately related to the more general cognitive processes that involve the encoding, storage and retrieval of information in memory. Moreover, experiments in auditory psychophysics have typically focused on experimental tasks and paradigms that involve discrimination rather than identification or recognition, processes thought to be most relevant to speech perception. All in all, it is generally believed that a good deal of what has been learned from research in auditory psychophysics and general auditory perception is only marginally relevant to the

---

1) David Pisoni could not be present at the congress and Michael Studdert-Kennedy acted as moderator at the meeting. David Pisoni is author of the introduction below.

study of speech perception and to an understanding of the underlying perceptual mechanisms. This situation has changed for the better in recent years as shown by the work of Dr. Divenyi and other psychophysicists who have become concerned with questions of speech perception. Despite these obvious differences, investigators have been interested in the differences in perception between speech and nonspeech signals. That such additional differences might exist was first suggested by the report of the earliest findings of categorical discrimination of speech by Liberman and his colleagues (1957). And it was with this general goal in mind that the first so-called "nonspeech control" experiment was carried out by Liberman and his colleagues (1961) in order to determine the basis for the apparent distinctiveness of speech sounds. In this study the spectrographic patterns for the /do/ and /to/ continuum were inverted producing a set of nonspeech patterns that differed in the onset time of the individual components. The results of perceptual tests showed peaks in discrimination for the speech stimuli replicating earlier findings. However, there was no evidence of comparable discrimination peaks for the nonspeech stimuli, a result that was interpreted at the time as further evidence for the distinctiveness of speech sounds and the effects of learning on speech perception. Numerous speech-nonspeech comparisons have been carried out over the years since these early studies, including several of the contributions to the present symposium. For the most part, these experiments have revealed results quite similar to the original findings of Liberman et al. Until quite recently, research reports have confirmed that performance with nonspeech control signals failed to show the same discrimination functions that were observed with the parallel set of speech signals (Cutting and Rosner, 1974; Miller et al., 1976; Pisoni, 1977). Subjects typically responded to the nonspeech signals at levels approximating chance performance. In more recent years, such differences in perception have been assumed to reflect two basically different modes of perception--a "speech mode" and an "auditory mode". Despite attempts to dismiss this dichotomy, additional evidence continues to accumulate as has been suggested by several of the new findings summarized in the papers included in this symposium.

The picture is far from clear, however, because the problems inherent in comparing speech and nonspeech signals have generated several questions about the interpretation of results obtained in earlier studies. First, there is the question of whether the same psychophysical properties found in the speech stimuli were really preserved in the parallel set of nonspeech control signals. Such a criticism is appropriate for the original /do/--/to/ nonspeech control stimuli which were simply inverted patterns reproduced on the pattern playback. The same remarks also apply to the well-known "chirp" and "bleat" control stimuli of Mattingly et al. (1971) which were created by removing the formant transitions and steady-states from the original speech context. These stimuli were presented in isolation to subjects for discrimination. Such manipulations, while nominally preserving the phonetic "cue" obviously result in marked changes in the spectral context of the signal which no doubt affects the detection and discrimination of the original formant transition. Such criticisms have been taken into account in the more recent experiments comparing speech and nonspeech signals as summarized by Dr. Dorman and Dr. Liberman, in which the stimulus materials remain identical across different experimental manipulations. While these more recent studies relieve some of the ambiguities of the earlier experiments, problems still remain in drawing comparisons between speech and nonspeech signals. For example, subjects in these experiments rarely practice with the nonspeech control signals to develop the competence required to categorize them consistently. With complex multidimensional signals it is quite difficult for subjects to attend to the relevant attributes that distinguish one signal from others presented in the experiment. A subject's performance with these nonspeech signals may therefore be no better than chance if he/she is not attending selectively to the same specific criterial attributes that distinguished the original speech stimuli. Indeed, not knowing what to listen for may force a subject to attend selectively to an irrelevant or misleading attribute of the signal itself. Alternatively, a subject may simply focus on the most salient auditory quality of the perceived stimulus without regard for the less salient acoustic properties which often are the most important in speech such as burst spectra or formant transitions. Since almost all of the nonspeech experiments conducted in the past were

carried out without the use of discrimination training and feedback to subjects, an observer may simply focus on one aspect of the stimulus on one trial and an entirely different aspect of the stimulus on the next trial. Without training experience to help the subject identify the criterial properties, the observed performance may be close to chance, a result that has been reported quite consistently in the literature. Setting aside some of these criticisms, the question still remains whether drawing comparisons in perception between speech and nonspeech signals will yield meaningful insights into the perceptual mechanisms deployed in processing speech. In recent years, the use of cross-language, developmental and comparative (i.e., cross-species) designs in speech perception research has proven to be quite useful in this regard as a way of separating out the various roles that genetic predispositions and experience play in speech perception. On the other hand, these types of investigations provide needed information about the course of learning and perceptual development since spoken language must be acquired in the local environment through social contact. On the other hand, comparative studies with both speech and nonspeech stimuli are useful in defining the lower limits on auditory system function. However, there are serious limitations in studies of this kind. For example, while it is cited with increasing frequency that chinchillas categorize synthetic stimuli differing in VOT in a manner quite similar to English-speaking adults, little if anything is ever mentioned, however, about the chinchilla's failure to carry out the same task with stimuli differing in the cues to place of articulation in stops, a discrimination that even young prelinguistic infants can make (Eimas, 1974). Should we then conclude that the English voicing contrast is purely sensory in origin, while place of articulation or voicing in Thai is somehow more "linguistic", brought on by inheritance or very early experience? With a little reflection, I think the answer must surely be negative. Such comparative studies are useful in speech perception research but only to the extent that they can specify the lower-limits on the sensory properties of the stimuli themselves. However, these findings are incapable, in principle, of providing any further information about how these signals might be "interpreted" or coded within the context of the experience and history of the organism.

Animals simply do not have spoken language and they do not and cannot recognize, as far as I know, differences between phonetic and phonological structure, a fundamental dichotomy in all natural languages. Cross-language and developmental designs have also been quite useful in providing new information about the role of early experience in perceptual development and the manner in which selective modification or tuning of the perceptual system takes place. Although the linguistic experience and background of a listener was once thought to control his/her discriminative capacities in speech perception experiments, recent findings strongly suggest that the perceptual system has a good deal of plasticity for retuning and realignment, even into adulthood. The extent to which control over the productive abilities remains plastic is still a topic to be explored. To what extent is it then useful to argue for the existence of different modes of perception for speech and nonspeech signals? Some investigators such as Dr. Ades would simply dismiss the distinctions drawn from earlier work on the grounds of parsimony and generality. He has argued recently (Ades, 1977) and in his contribution to this symposium that differences in perception between speech and nonspeech or consonants and vowels can be accounted for simply by recourse to the notion of "range" or the width of the context expressed in terms of the number of JNDs. As long as the range is small, absolute identification performance will be as good as differential discrimination. When the range is large, however, discrimination will be better than identification. Thus according to the account offered by Ades, a consonant continuum should display a smaller range than a vowel continuum. But as shown in Fig. 1 the facts are quite the reverse of his predictions.

In this figure we have reproduced the identification data collected by Perey and Pisoni (1977) in a magnitude estimation task. On each trial subjects had to respond to a stimulus with a rating on a scale from 1 to 7. One group of subjects received a consonant continuum differing in VOT, another received a vowel continuum. Through various transformations of the obtained stimulus-response matrix, scale scores were derived and an estimate of the perceived psychological spacing between stimuli was obtained. Scale scores are expressed in this figure in terms of

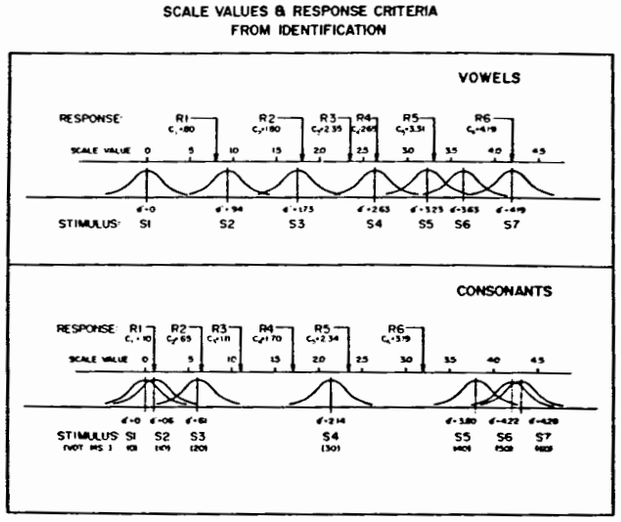

Figure 1. Scale values showing the perceived psychological space for consonants and vowels. Data were taken from Perey and Pisoni (1977) who required subjects to use a rating response in identification.

d's and by summing these individual values, an estimate of the total range or spacing of the stimuli was obtained. The cumulative d' is shown on the far right of each panel. Notice that the cumulative d' for the vowels shown on the top is 4.19 while the value for the consonants shown on the bottom is 4.28. If stimulus range were the correct explanation of the differences in perception between consonants and vowels as Dr. Ades would have us believe, the consonants should have displayed the smaller range. Obviously, this is simply not the case. However, what is of interest in this figure is the psychological spacing of signals within each panel. For the consonants, the spacing between adjacent stimuli is clearly unequal with a grouping close to the endpoints of the series. For the vowels, the spacing is more nearly equal across all the test stimuli suggesting the possibility of better resolution in discrimination, a result that has been known for

many years. Thus, Dr. Ades' argument that the range of stimuli can account for differences in perception between consonants and vowels or speech and nonspeech would seem to be incorrect, despite his attempts to generalize the Durlach and Braida (1969) model to speech perception. Moreover, this is a curious position to maintain anyway as it is commonly recognized, not only in speech perception research but in other areas of perceptual psychology, that "nominal" stimuli may receive differential amounts of processing or attention by the subject, that subjects may organize the interpretation of the sensory information differently under different conditions and that the sensory trace of the initial input signal may show only a faint resemblance to its final internal representation resulting from encoding and storage in memory. It is hard to deny that a speech signal elicits a characteristic mode of response in a human subject--a response that is not simply the consequence of an acoustic waveform leaving a meaningless sensory trace in the auditory periphery. Nevertheless, there is a great deal to learn about how the auditory system codes complex acoustic signals such as speech. Dr. Dorman, in summarizing work on the perception of transitions in speech and nonspeech context, has tried to establish the need for a specialized speech processor to account for differences in labelling of sine-wave stimuli when heard as either speech or nonspeech. Such explanations seem to me entirely premature at this time as the relevant psychophysical experiments with nonspeech signals have simply not been carried out yet. To remedy this state of affairs we have begun to collect labelling data in our laboratory recently using brief FM stimuli followed by a constant frequency (CF) steady-state. Schematized spectrograms of the test stimuli are shown in Fig. 2.

The left panel of this figure shows an idealized set of stimuli differing in the initial starting frequency of the FM. Three steady-state (CF) frequencies were selected, 850, 1500 and 2300 Hz. For each set we generated 21 test signals which spanned a range of 500 Hz above and below the CF of the steady-state component. In Experiment I the three sets of signals consisted of an isolated single component as shown on the left. In Experiment II we added an additional 500 Hz component to each of the original three sets of stimuli. Subjects were required to identify the

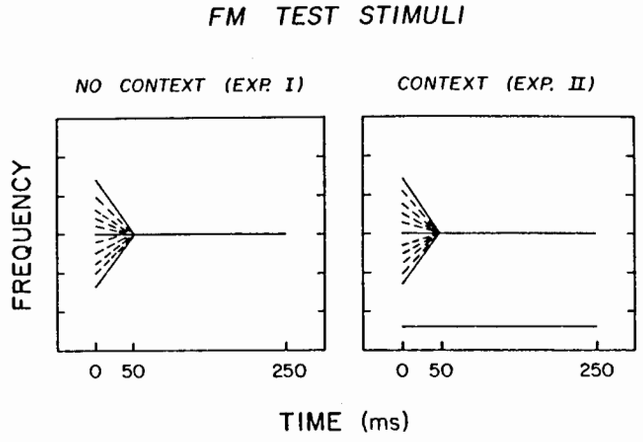

Figure 2. Schematized patterns showing the time course of the non-speech FM stimuli: The panel on the left illustrates the test stimuli without spectral context, the panel on the right shows the addition of a low frequency component to the same signals.

stimuli as "rising", "level" or "falling" after a brief training period with good exemplars selected from each category. The results of both experiments are shown in Fig. 3.

The labelling functions shown at the top for the three CF conditions reveal that the middle or "level" category response increases slightly in size as the CF of the steady-state increases from 850 Hz to 1500 Hz, a result that is consistent with what is known about frequency resolution in the auditory system. Over a wide range of frequencies, discrimination follows Weber's law. Thus, the level category should widen as the frequency of the steady-state increases for the same difference in initial starting frequency. Note that we have plotted starting frequency on a linear rather than log scale. The results for Experiment II in which an additional steady-state component was added are shown in the lower panel of the figure. Notice that for the 850 Hz condition the "level" category is now slightly larger than in the top panel suggesting the strong possibility of some interaction between the individual components. However, the other two condi-

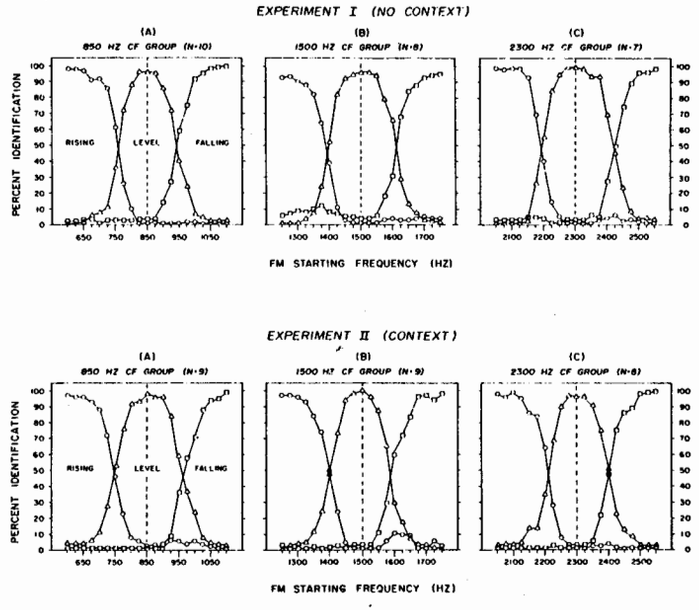

Figure 3. Identification data for FM stimuli obtained with three different steady-state CF's, 850 Hz, 1500 Hz and 2300 Hz. The top panel shows the identification data collected for FM's without context, the lower panel shows the data for test signals with the additional steady-state context present.

tions in Experiment II show a somewhat narrower range for the "level" category compared to the top panel indicating better resolution of frequency in the presence of another signal, a well known fact in auditory psychophysics. These recent findings were not originally intended to refute the arguments of Dorman and his colleagues who favor the postulation of some specialized perceptual mechanism for processing speech signals. Rather, I simply wanted to illustrate by way of example that the location of perceptual categories observed with nonspeech signals is not rigidly controlled by some simple physically defined invariant such as the direction of the frequency change. Moreover, as Dr. Divenyi has pointed out so well in his paper, we need to know much more about how the

basic constraints of the auditory system affect the way speech is initially coded for subsequent processing. Thus, in the present case several basic facts about frequency discrimination are sufficient to account for changes in our subjects' perceptual categorization of nonspeech FM's that are similar to speech. Whether it will be possible to generalize such psychophysical explanations to more complex signals such as speech remains to be seen from future research currently in progress in our laboratory and elsewhere.

In summary, there still appears to be good evidence for distinguishing between speech and nonspeech signals and for recognizing the existence of two distinct modes of perception, one associated with the sensory or psychophysical correlates of acoustic signals and the other with the interpretation and coding of acoustic signals as speech. Recent work has attempted to make these differences more precise by subjecting them to experimental test and searching for common underlying explanations. Taken together such results suggest to me that, just as in the case of "species-typical responding" observed in the behavior of other animals, the notion of a "speech mode" of perception captures certain aspects of the way human observers typically respond to speech signals that are highly familiar to them. We still do not know if it is simply a matter of familiarity as with music or whether there is something deeper and more closely related to biological survival of the organism. Nevertheless, such a conceptualization does not, at least in my view, commit one to the view that human listeners cannot respond to speech signals in other ways more closely correlated with the sensory or psychophysical attributes of the signals themselves. To deny the speech mode, however, is to ignore the fact that acoustic signals generated by the human vocal tract are used in a distinctive and quite systematic way by both talkers and listeners to communicate linguistically, a species-typical behavior that is restricted, as far as I know, to Homo sapiens.

Past experiments comparing the perception of speech and nonspeech signals have been quite useful in characterizing how the phonological systems of natural languages have, in some sense, made use of the general properties of sensory systems in selecting an inventory of phonetic features and their acoustic correlates (Stevens, 1972). The relatively small number of distinctive features and their acoustic correlates that can be observed across a wide variety of diverse languages implies that there is a common sensory basis for language perception, a common means of controlling the mechanisms of speech production and a common cognitive definition of linguistic structure. Whether these facts are causally related will no doubt be a matter of much debate, speculation and new research in the years to come. It is clear, nevertheless, that the distinctions drawn in perception between speech and nonspeech signals still remain fundamental, setting apart research on speech perception from the study of auditory psychophysics and the field of auditory perception more generally.

Acknowledgements

The preparation of this paper was supported, in part, by NIMH grant MH-24027 and NINCDS grant NS-12179 to Indiana University in Bloomington. I am grateful to Peter Jusczyk and Jim Sawusch for comments on an earlier draft of the paper. Robert Remez discussed many of the theoretical issues summarized in the paper with me at length and provided helpful editorial comments that improved the overall exposition and quality. His help is greatly appreciated.

References

Ades, A.E. (1977): "Vowels, consonants, speech and nonspeech", Psych. Rev. 84, 524-530.

Cutting, J.E. and B.S. Rosner (1974): "Categories and boundaries in speech and music", Perc. Psych. 16, 564-570.

Durlach, N.I. and L.D. Braida (1969): "Intensity perception I. Preliminary theory of intensity resolution", JASA 46, 372-383.

Eimas, P.D. (1974): "Auditory and linguistic processing of cues for place of articulation by infants", Perc. Psych. 16, 513-521.

Liberman, A.M., K.S. Harris, H.S. Hoffman, and B.C. Griffith (1957): "The discrimination of speech sounds within and across phoneme boundaries", J.Exp.Psych. 54, 358-368.

Liberman, A.M., K.S. Harris, J.A. Kinney, and H.L. Lane (1961): "The discrimination of relative onset time of the components of certain speech and non-speech patterns", J.Exp.Psych. 61, 379-388.

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

Miller, J.D., C.C. Wier, R. Pastore, W.J. Kelly, and R.J. Dooling (1976): "Discrimination and labeling of noise-buzz sequences with varying noise-lead times: An example of categorical perception", JASA 60, 410-417.

Perey, A.J. and D.B. Pisoni (1977): "Dual processing versus response-limitation accounts of categorical perception: A reply to MacMillan, Kaplan and Creelman", JASA 62, S1, 60-61.