SPEECH & NON-SPEECH: WHAT HAVE WE LEARNED?

Anthony E. Ades[1], Max-Planck-Gesellschaft, Projektgruppe für Psycholinguistik, Nijmegen, Netherlands.

There have been two strands of research in the speech/non-speech controversy. Firstly there are experiments where speech is compared to non-speech signals that have critical acoustic properties of speech. (See Wood, 1976, for references). This work has shown that there is no real difference: the perceptual properties of speech arise from its acoustics, not from its "speechlikeness".

This paper is concerned with the second strand, where series of speech sounds are compared to stimuli that differ along simpler dimensions like pitch and intensity. I shall summarise the arguments presented in a recent theoretical article (Ades, 1977), and extend them to other paradigms. The conclusion I shall draw is that speech/non-speech differences, as well as consonant/vowel differences, do not result from any inherent property of the sounds themselves, such as their speechlikeness, or degree of "encodedness" (Liberman, Mattingly and Turvey, 1972), but instead depend on a property of the ensembles of stimuli used in these experiments.

This property is the range, or width of context, of the ensemble. One may think of it as the number of just-noticeable-differences across the series. This analysis is borrowed from Durlach and Braida's (1969) quantitative theory of intensity resolution. They and their colleagues, in the course of testing this theory, have obtained results for intensity that are quite analogous to results commonly obtained for speech.

There has been a consistent failure to control for the range variable when making comparisons between vowels, consonants, speech, and non-speech.

Identification and Discrimination

It all started, I think, with Miller's observation (1956) that we can discriminate far better than we can identify. Consider an intensity discrimination experiment where the subject is asked to decide if two sounds are the same or different. This can be done reliably if they are about one dB apart. Now suppose that there are 15 sounds, evenly spaced along a continuum spanning 25 dB. About

---

(1) This paper was prepared while the author held a Fellowship under the Royal Society European Science Exchange Programme.

25 discriminations could be made in this space. But when the subject is asked to label the stimuli with a number between 1 and 15 (give as much practice and feedback as possible), only seven plus or minus two categories can be used accurately. The subject is distinguishing between adjacent stimuli in identification less than half as well as (s)he would in discrimination. How can this be? After all, sensitivity to acoustic signals and to differences between them must be the same in both situations.

The answer must lie in the memory requirements. In discrimination there are two or more stimuli: the subject must store their sensory traces, compare them (perhaps by subtraction), and pronounce on the difference if any. Call this the trace mode. In the identification case, a single sensory trace must be compared to some representation of the entire stimulus series. Where, the subject must decide, does this stimulus fit, given all the others I have heard. This is the context coding mode. Presumably, the representation of the series is not in the form of traces, but is in some verbal or numerical code.

Now consider what happens to identification when the range of the ensemble is increased from 25 to 50 dB. A reasonable guess is that "accuracy", defined as the ability to place the current stimulus in context, expressed as a percentage of the size of the context, will remain constant. Of course, the absolute size of errors, expressed in j.n.d. terms, will now be larger. An archer who remains a constant 3 degrees off centre will show a larger absolute error at 50 yards than at 25.

So far, then, we assume that in discrimination (in its ideal form), the only factor affecting performance is the noisiness of the representation of the acoustic traces. In identification there will be trace noise too, but there will also be context coding noise when the subject attempts to locate a stimulus in its context. This will increase as the range increases.

The critical prediction is that as long as range is small, identification performance will be as good as discrimination. For, context noise will be minimal, and trace noise will be the only determinant in both tasks.

The theorising above is an informal statement of Durlach and Braida's (1969) quantitative theory for intensity resolution. The

above prediction, that as range decreases, identification improves, and finally approximates discrimination, was confirmed by Pynn, Braida, and Durlach (1972), for stimuli differring in intensity.

And now to speech. The classic result (Studdert-Kennedy et al., 1970) is that for series of consonant - vowel stimuli, discrimination is scarcely better than identification. Given Miller's paper on how identification is relatively weak in non speech, it was natural to see the speech results as evidence for a speech-specific mode of processing. The alternative I propose is simply that CV series are not unusual by virtue of being speech, but simply have relatively small ranges. I have shown elsewhere (Ades, 1977) that the best estimates of the range of CV series make them comparable in j.n.d. terms to the small ranges used by Durlach, Braida, and their colleagues for intensity resolution experiments.

Typically, a series of synthetic speech sounds from /ba/ to /da/, or from /ba/ to /pa/, spans between 3 and 5 j.n.d.s. A series of vowels, on the other hand must stretch across about 10 j.n.d.'s to reach from one category to another. We thus expect that discrimination on vowels will far excede what would be predicted from identification. This has been consistently found. Generally, though, it has been incerpreted to mean that vowels are somehow less "speech-like" that consonants (as if one could have stop consonants without vowels!). It should be clear that I am trying to replace this rather mystical theorising with the idea that speech, non-speech, vowels and consonants are all the same. The observed differences are due to the range variable. The number of j.n.d.'s across the series, can be used as a stimulus-free approximation of the size of the range.

More complex experiments

In certain cases, the range has an effect in discrimination experiments, not just in identification. This is because certain variations in the task parameters may make it profitable for the subject to operate in the context mode, i.e. to do identification: as, for instance, when the procedure adds noise or interference to the sensory trace mode. We can predict that any manipulation that makes comparison of traces harder will only worsen performance if the range is large! For, if the range is small, the subject can escape the trace noise, slip into the context-coding mode and not suffer too much from context noise.

In these cases it is important how discrimination is tested: if the pair to be discriminated randomly changes from trial to trial, then the effective range is the range of the entire series. But if the same pair is tested many times before another part of the series is tested, the effective range will obviously be very small. It turns out that in speech research the "roving level" method is always used. Thus procedures that cause trace comparison to be harder, such as increasing the time interval between the two stimuli, or by forcing the subject to compare three traces at a time rather than two, such procedures will, in roving level testing, make the range variable critical.

Experiments of this type have been done with vowels and consonants (Pisoni, 1973, 1975). As we predict from the Durlach and Braida model, manipulations that worsen discrimination have a stronger effect on vowels than on consonants, because, according to the range hypothesis, the small range of consonants makes escape into context-coding possible without running into context memory noise. In intensity resolution, Berliner and Durlach (1973) have shown that increased time delay between stimuli to be discriminated worsens resolution only if the range is large.

The "anchor" effect and RT Experiments

The same ideas can be applied to other paradigms where speech and non-speech have been contrasted. In the two areas that follow I confess to being less certain of my argument, because I do not know of research where the range variable has been systematically studied.

Firstly, the "anchor effect". A series of sounds varying in pitch is constructed and the subject asked to identify them as "High" or "Low". If an endpoint stimulus (the anchor), say the highest pitched one, is presented two or three times as often as the others, the entire identification curve is shifted towards to the anchor. However, such shifts do not occur in stop-consonant series (Sawusch and Pisoni, 1973; Simon and Studdert-Kennedy, 1978). Again, we might expect that the different ranges of pitch and consonant series are involved. We may assume a 5 j.n.d. range for the speech series. The pitch series went from 114 Hz to 15o Hz: assuming a difference limen of 0.5 Hz for pitch (Klatt, 1973), this

series would span over 50 j.n.d.s. Both Simon et al., and Sawusch et al. (1974) also found a strong anchor effect in a series varying in intensity. This covered 18 dB in one experiment and 24 in the other, about 2o j.n.d.s.

Certainly, then, the range differences between the speech and non speech series were marked. But why should the range determine the anchor effect? I have no formal answer to this, but it is clear that anchor effects cannot be located in the trace mode. Also, Berliner, Durlach and Braida (1977) have shown that the "edge effect", whereby resolution in identification is better at the ends of a continuum than in the middle, and which is identified in their model as a perceptual anchoring effect in the context coding mode, is enhanced by increased range.

A second paradigm is a Reaction Time task where the subject must press one of two buttons depending on whether the stimulus is /ba/ or /da/, or whether it has high or low pitch. The point here is that if the subject is responding to the speech distinction, irrelevant variation in pitch slows the RT. However, irrelevant variation in place of articulation has a much smaller effect on RT to the pitch distinction (Day and Wood, 1972). Wood (1973) also showed that there was mutual interference between pitch and intensity, and also between place of articulation and voicing. This was interpreted as revealing two separate systems: such that there was interference within each, speech with speech, non-speech with non-speech; but no interference between.

The alternative is that both pitch and intensity discriminations are easy, while both place and voicing are harder. Interference will occur if the irrelevant variation is as salient or more salient than the distinction being tested. The situation where interference is least is precisely the one where the discrimination (pitch) is much more salient than the interfering dimension (place).

Finally, let me add that the point I have been trying to make for discriminations vs identification, anchor effects, and RT experiments has already been forcefully made for experiments on the Precategorical Acoustic Store (PAS), and on the hemispheric lateralisation of speech. The fact that sets of stop-consonant-vowel syllables produce no recency effect in PAS, whereas sets of vowels do, has been taken to mean that consonants and vowels are differentially "encoded" (Liberman et al., 1972). But Darwin and Baddeley (1974) have shown that the vowel/consonant distinction here is irrelevant: what controls the recency effect is, again, the discriminability of the items within the ensemble. Similarly, the same factor is critical in determining the degree of hemispheric lateralisation for vowels (Godfrey, 1974).

Conclusions

At the very least it must be conceded that explorations of speech/non-speech and vowel/consonant differences might be meaningless unless factors corresponding to discriminability across the stimulus ensemble are controlled. It is obvious that the range variable is all-important in the experiments briefly reviewed here. In addition, once range is controlled for, a single unified theory for all stimuli seems well within reach. And this is surely preferable to one theory for non-speech, a second theory for consonants, (and an in-between theory for vowels).

Whether or not the above proposals are correct, the entire speech/non-speech issue seems to have acquired a life of its own, which it fights for against all odds. However, according to the views expressed here, it has taught us very little, and has simply served to direct out attention from the real problems of speech perception, exemplified for example in automatic recognition (Klatt, 1977, for a review), where the psychological contribution remains slight and engineering solutions prevail.

References

Ades, A. E. (1977): "Vowels, Consonants, Speech, and Nonspeech", Psych. Rev. 84, 524-530.

Berliner, J. E., and N. I. Durlach (1973): "Intensity Perception IV: Resolution in Roving Level Discrimination", JASA. 53, 127o-87.

Berliner, J. E., N. I. Durlach, and L. D. Braida (1977): "Intensity Perception VII. Further Data on Roving Level Discrimination and the Resolution and Bias Edge Effects", JASA, 61, 1577-85.

Darwin, C. D., and A. D. Baddeley (1974): "Acoustic Memory and the Perception of Speech", Cogn. Psych, 6, 41-6o.

Day, R. S., and C. C. Wood (1972): "Interaction between Linguistic and Nonlinguistic Processing", JASA, 51, 79(A).

Durlach, N. I., and L. D. Braida (1969): "Intensity Perception I. Preliminary Theory of Intensity Resolution", JASA, 46, 372-83.

Godfrey, J. J. (1974): "Perceptual Difficulty and the Right-Ear Advantage for Vowels". Brain and Language, 4, 323-36.

Klatt, D. H. (1973): "Discrimination of Fundamental Frequency Contours in Synthetic Speech: Implications for Models of Pitch Perception", JASA, 53, 8-16.

Klatt, D. H. (1977): "Review of the ARPA Speech Understanding Project", JASA, 62, 1345-66.

Liberman, A. M., I. G. Mattingly, and M. T. Turvey (1972): "Language Codes and Memory Codes". In A. W. Melton and E. Martin (Eds.) Coding Processes in Human Memory, Washington, D. C.: Winston.

Miller, G. A. (1956): "The Magical Number Seven, Plus or Minus Two: Some Limits on our capacity for Processing Information", Psych. Rev., 63, 81-97.

Pisoni, D. B. (1973): "Auditory and Phonetic Codes in the Discrimination of Consonants and Vowels", Perc. Psych., 13, 253-60.

Pisoni, D. B. (1975): "Auditory Short-Term Memory and Vowel Perception", Memory and Cognition, 3, 7-18.

Pynn, C. T., L. D. Braida, and N. I. Durlach (1972): "Intensity Perception III. Resolution in Small-Range Identification", JASA, 51, 559-66.

Sawusch, J. R., and D. B. Pisoni (1973): "Category Boundaries for Speech and Nonspeech Sounds", JASA, 54, 76(A).

Sawusch, J. R., D. B. Pisoni and J. E. Cutting (1974): "Category Boundaries for Linguistic and Nonlinguistic Dimensions of the Same Stimuli", JASA, 55, S55(A).

Simon, H. J., and M. Studdert-Kennedy (1978): "Selective Anchoring and Adaption of Phonetic and Nonphonetic Continua", JASA, 64, 1338-57.

Studdert-Kennedy, M., A. M. Liberman, K. S. Harris, and F. S. Cooper (197o): "Motor Theory of Speech Perception: A Reply to Lane's Critical Review", Psych. Rev., 77, 234-49.

Wood, C. C. (1973): "Levels of Processing in Speech Perception. Neurophysiological and Information Processing Analyses". Unpublished Doctoral Dissertation, Yale University.

Wood, C. C. (1976): "Discriminability, Response Bias, and Phoneme Categories in Discrimination of Voice Onset Time", JASA, 6o, 1381-89.