

The Scaling of Synthetic Speech

In a previous report it was suggested that scaling techniques might be applied to the study of the perception of synthetic speech. The report posed two questions:

- (1) Can reliable scaling methods be developed which work with stimuli such as synthetic speech sounds?
- (2) Will such methods be sufficiently sensitive to show up perceptual differences between different classes of speech sound?

The work to date suggests that the answer to both of these questions is yes. A number of different scaling methods have been tried. The data are reasonably reliable and there are systematic differences between steady state vowels and stop consonants.

In essence, the experiment requires a subject to estimate how much difference he perceives between stimuli. For example, stimuli are presented in random pairs and the subject judges the difference between the stimuli in each pair. The procedure is simple but contains a major methodological problem.

In the experiment, it is necessary to present a sequence of pairs rather than a single pair; the subject makes a corresponding sequence of judgments. The subject has no absolute standard to judge by and so tends to make the judgments in relation to each other. The practical consequence of this is that each judgment is affected by the one that precedes it. With the vowels, the effect is serious; different sets of judgments are obtained for the same vowel pairs presented in different orders. With the consonants, the effect is either negligible or absent. The fact

that this kind of context effect holds for the steady state vowels and not for the stop consonants is an important difference between these two classes of stimuli.

If meaningful comparisons are to be made between the two classes however, it is important to find a single suitable scaling technique. A number of different procedures have been tried. The aim has been to find a way of ensuring that the vowels will be judged independently of the order of presentation.

There are 13 vowels and 13 consonants. The vowels form a series which appears to run smoothly from 'EE' as in "beat" to 'e' as in "bet". The consonants form a series which seems to be broken into 3 segments, sounding rather like 'ga..da..ba'. Both series may be referred to in terms of: s13, s12, s11..... ..s2, s1.

In the initial scaling procedure, the vowels (or consonants) were presented in random pairs. Thirteen stimuli make ninety one possible pairs and each of these was presented twice. The method had to be abandoned because of the context effect with the vowels. Next, the stimuli were presented in pairs with s13 always serving as the first member of the pair, e.g., s13-s6; s13-s2; s13-s8. The hope was that the constant (extreme) vowel would serve as an 'anchor' and free each judgment from the influence of the last. It did not. A further modification consisted of preceding each of the pairs by a constant pair (made of the two extreme stimuli). For example,

s13-s1
s13-s6

s13-s1
s13-s2

s13-s1 could now function as a frame of reference and eliminate the context effect. However, the judgment made for s13-s2, for example, still depended on the judgment made for s13-s6.

A new procedure was tried. The stimuli were presented in triplets: s13-s6-s1; s13-s2-s1; s13-s8-s1; etc. In this experiment, the subject was presented with a line with one end representing s13 and the other end representing s1. The task was to dissect the line thus indicating the position of the middle stimulus in relation to the two 'anchors'. The context effect was still present but finally it was eliminated by preceding each triplet with a constant sequence:

s13-s11-s9-s7-s5-s3-s1
s13-s6-s1

s13-s11-s9-s7-s5-s3-s1
s13-s2-s1

etc.

Since the middle stimulus in the triplet may be any of the thirteen stimuli, there are thirteen possible combinations to present to the subject. For the vowels, these thirteen combinations have been presented in random order to a small group of subjects. Satisfactory results have been obtained. The context effect is no longer important and the data show a definite psychometric trend. A similar tape is now being made for the consonants.

The vowel data so far collected indicate a continuous function but one which is less than completely smooth. Additional data are required before any particular significance can be attached to this feature.

The immediate plan is to collect data for the vowels, for

the consonants and for a set of pure tones. The psychometric function for pure tones is well known and will provide a useful non-speech comparison.

Finally, it is hoped to extend this experimental method to other classes of speech stimuli so that a range of psychometric functions may be compared.

Michael Vinegrad