

A Direct Magnitude Scaling Method to Investigate Categorical Versus Continuous Modes of Speech Perception

M.D. Vinegrad⁺

Haskins Laboratories, New Haven

Summary. The perception of synthetic steady-state vowels, synthetic consonant-vowel syllables, and pure tones was investigated using a psychophysical scaling procedure involving direct magnitude estimation. In each of the three stimulus classes, there were thirteen members equally spaced along a physical continuum; the investigation was designed to measure the degree to which the stimulus members appeared to be evenly spaced along a corresponding perceptual continuum. The experimental technique required the subject to judge the members of each set of stimuli in terms of their similarity to each other. The results suggested that for stops, the perceptual spacing depended upon phoneme identification but that for vowels, the spacing was relatively independent of phoneme identity. The vowel data, in fact, approximated to the tone data (included to provide a nonspeech comparison). The results are interpreted in terms of the notion of categorical versus continuous modes of speech perception.

⁺Currently, Goldsmith's College, University of London

Acknowledgement. The author is indebted to Dr. A.M. Liberman for his encouragement, suggestions, and helpful criticism through the course of this research.

The experiment reported here is a psychophysical study of synthetic speech perception. The primary aim was to investigate quantitatively the tendency for stop consonants to be perceived categorically and steady-state vowels to be perceived continuously. The method used was a stimulus scaling technique that has proved useful in other areas of perceptual research. In the experiment, a sequence of three stimuli was presented to a subject; the stimuli were spaced along a physical continuum and the subject was required to judge how the stimuli appeared to be spaced along a perceptual continuum. The subject indicated his judgment by spacing points along a line.

The technique is both simple and direct. A variety of studies using nonspeech stimuli have shown the reliability and usefulness of this approach (Torgerson, 1958). Speech stimuli have not previously been studied in this way, partly because of their complexity and multidimensionality. Most of the studies found in the scaling literature are limited to situations where changes in both stimulus and perception are unidimensional. One aim of this investigation, therefore, was to test the suitability of the method for synthetic speech stimuli. At a practical level, scaling has the advantage of being relatively easy to carry out and, in addition, directly reflects the way stimuli are perceived by a subject. Other psychophysical procedures (for example, discrimination measurements) lead only to inferences about the mode of perception. Results from investigations using discrimination techniques (Liberman et al., 1967; Stevens et al., 1969) suggest that there may be different modes of perception for vowels and stop consonants. The present experiment is intended to be a further examination of this question. The stimuli used were steady-state vowels, stop consonants, and pure tones; the tones were included to provide a nonspeech comparison.

Description of Stimuli

The speech sounds were thirteen steady-state vowels and thirteen consonant-vowel syllables. They were synthesized by Stevens, Liberman, Studdert-Kennedy, and Öhman (1969) on the OVE II speech synthesizer at the speech transmission laboratory at the Royal Institute of Technology at Stockholm. A full description of the stimuli is given in Stevens et al. (1969). In each set, the stimuli (numbered 1 through 13, for reference) were evenly spaced along a physical continuum. Listened to in order, the vowels appeared to form a smooth series running through the American English vowels /i/, /I/, and /ε/. The physical

spacing of the thirteen vowels corresponded to changes in the frequencies of the first three formants of a five-formant pattern. Moving along the continuum from stimulus 1 to stimulus 13, the frequency of the first formant rose in approximately equal steps while the frequencies of the second and third formants decreased in approximately equal steps. For stimulus 1, the respective values of the first three formants were 270.5, 2300, and 3019 cps and for stimulus 13, the respective values were 530.5, 1858, and 2492 cps. Small deviations from even spacing existed because formant frequencies could only be set within a few cps. The bandwidths of the first three formants were fixed at 60, 80, and 100 cps. The frequencies of the fourth and fifth formants were constant throughout. The duration of each stimulus was 300 msec. The consonant-vowel syllables consisted of a stop consonant followed by the vowel /ɛ/. The stimuli (again labeled 1 through 13), when listened to in order, seemed to form a series broken into three segments each characterized by a change in stop. To most North American English listeners, the transition seemed to be /g/---/d/---/b/. Each of the thirteen stimuli was 300 msec in duration. The final 260 msec corresponded to the vowel portion of the syllable; it was a steady-state vowel with the first three formants fixed at 700, 1550, and 2600 cps. Before reaching these fixed values, the three formants underwent transitions along parabolic contours. The transitions for the second and third formants started at different points for different stimuli, and it was in terms of these starting points that the stimuli were spaced along the physical continuum. Going along the continuum from stimulus 1 to 13, the starting frequency for the second formant decreased in equal steps, while the starting frequency for the third formant increased in equal steps as far as stimulus 7 and then decreased in equal steps over the remainder of the range. This variation was designed to parallel the change in speech sound to be expected if the place of consonantal articulation were to be moved in thirteen equal stages from velar to alveolar to labial position. The set of thirteen tones was recorded directly from an audiogenerator. The frequencies of stimuli 1 and 13 were 250 and 298 cps, respectively, with the eleven intermediate stimuli evenly spaced in steps of 4 cps. The output of the audiogenerator was constant and was recorded on a Roberts tape recorder.

Experimental Procedure

In preparing stimuli for the experiment, multiple copies of the

original recordings of the speech stimuli were made on the Roberts recorder. Magnetic tape segments of each vowel and consonant-vowel syllable were then cut and spliced to form sequences of stimuli suitable for scaling. Similar sequences of tones were made by splicing magnetic tape segments of each frequency recorded from the audiogenerator.

The experiment was divided into three parts: (1) vowels, (2) stops, (3) tones. A single scaling procedure was used throughout. All subjects did the three parts in the same order and each part was completed before the subject had any experience with the stimuli of a later part.

The scaling procedure required the subject to listen to two sequences of stimuli and then to make a judgment. The first sequence always contained the same seven stimuli, viz., 1-3-5-7-9-11-13. The second sequence always contained three stimuli, viz., 1-x-13, where x is any of the thirteen stimuli. The subject was required to make a judgment about stimulus x; he was required to indicate how similar or close stimulus x seemed to be to stimulus 1 (or stimulus 13). He was not required to identify x in absolute terms but merely to indicate the position of x by marking a point on a line such that the point bore the same position in relation to the ends of the line as stimulus x bore to stimuli 1 and 13. The stimulus presentation is more cumbersome than is usual for direct magnitude estimation. The procedure was devised by trial and error and was designed to eliminate context effects. These are discussed in more detail below.

The subject was given a straight line, seven inches in length, and told that the left-hand end of the line was to be taken as representing the first stimulus and the right-hand end as representing the third stimulus. If the middle stimulus sounded exactly like the first, the subject was instructed to place a point at the extreme left-hand end of the line; if the middle stimulus sounded exactly like the third stimulus, the subject was instructed to place a point at the extreme right-hand end; if the middle stimulus sounded slightly different from the first, the instruction was to place a point slightly in from the left; and so on. Demonstrations and practice were provided until the subject had mastered the task. In addition, the subject was told that, although there were a number of different stimuli that might occur in the middle position, any one of them might be repeated. Subjects were discouraged from trying to identify the middle stimulus; they were urged to respond according to how similar the stimuli sounded. The two sequences were presented repeatedly but with random rotation in the choice of stimulus to fill the x position.

The subject was provided with a sheet with six seven-inch lines horizontally spaced out between two verticals. There were no labels or other marks to guide the subject in his judgment. The subject was told to use one line per judgment and to turn over to new sheets as necessary. No description of the stimuli was given; subjects were left to form their own frames of reference.

For part one of the experiment, a tape was made containing thirty-nine replications of the two sequences of stimuli; the middle position was filled by each of the thirteen stimuli three times; the order was random. There were two versions of the tape, one a partial rerandomization of the other. For parts two and three, each tape contained fifty-two replications of the stimulus sequences; the middle position was filled by each of the thirteen stimuli four times. The randomization was restricted so that each of the thirteen stimuli occurred twice in the first twenty-six presentations and twice in the second set of twenty-six. (The second set of twenty-six was, in fact, a partial rerandomization of the first twenty-six.) The vowel data were collected over eight experimental sessions, and the stop and tone data were each collected over six sessions. In an experimental session, the subject listened to a tape once through, making 39 judgments in the case of the vowels and 52 in the case of the stops and tones. In total, each subject made 312 judgments on each of the three kinds of stimulus.

The timing of the stimulus presentations was as follows: there was a three-second pause between the end of the first sequence and the beginning of the second and a five-second period for the subject to make his judgment, and then the cycle began again with the repetition of the first stimulus sequence. In each sequence, there was a half-second pause between one stimulus and the next. The five-second judgment period appeared to be optimum; longer periods left too much time for doubt, and subjects found the pace helped them form a suitable set for responding.

The tapes were played on a Hewlett Packard tape deck through a loudspeaker in a language laboratory. Subjects worked simultaneously but were not able to see each other's responses. Practice trials were given at the beginning of each session.

Context Effects

The experimental procedure required the subject to make a succession of judgments, and because there was no absolute standard to judge by, the subject

tended to make judgments in relation to each other. The practical consequence was that one judgment was partly determined by the preceding one. In the preliminary experiments, this was a serious source of error. With the vowels and tones, different sets of judgments were obtained for the same triplets presented in different orders, but with the stops, the effect was either negligible or absent. It was to overcome the context effect that each triplet was preceded by the longer (constant) sequence of stimuli. The constant sequence seemed to reset the subject's frame of reference and to eliminate, or drastically reduce, interference from the preceding judgment. In order to provide a uniform procedure throughout the experiment, the constant sequence was also used when scaling the stops. The fact that the context effect occurred with the vowels and tones but not with the stop consonants is an important difference between these classes of stimuli.

Subjects

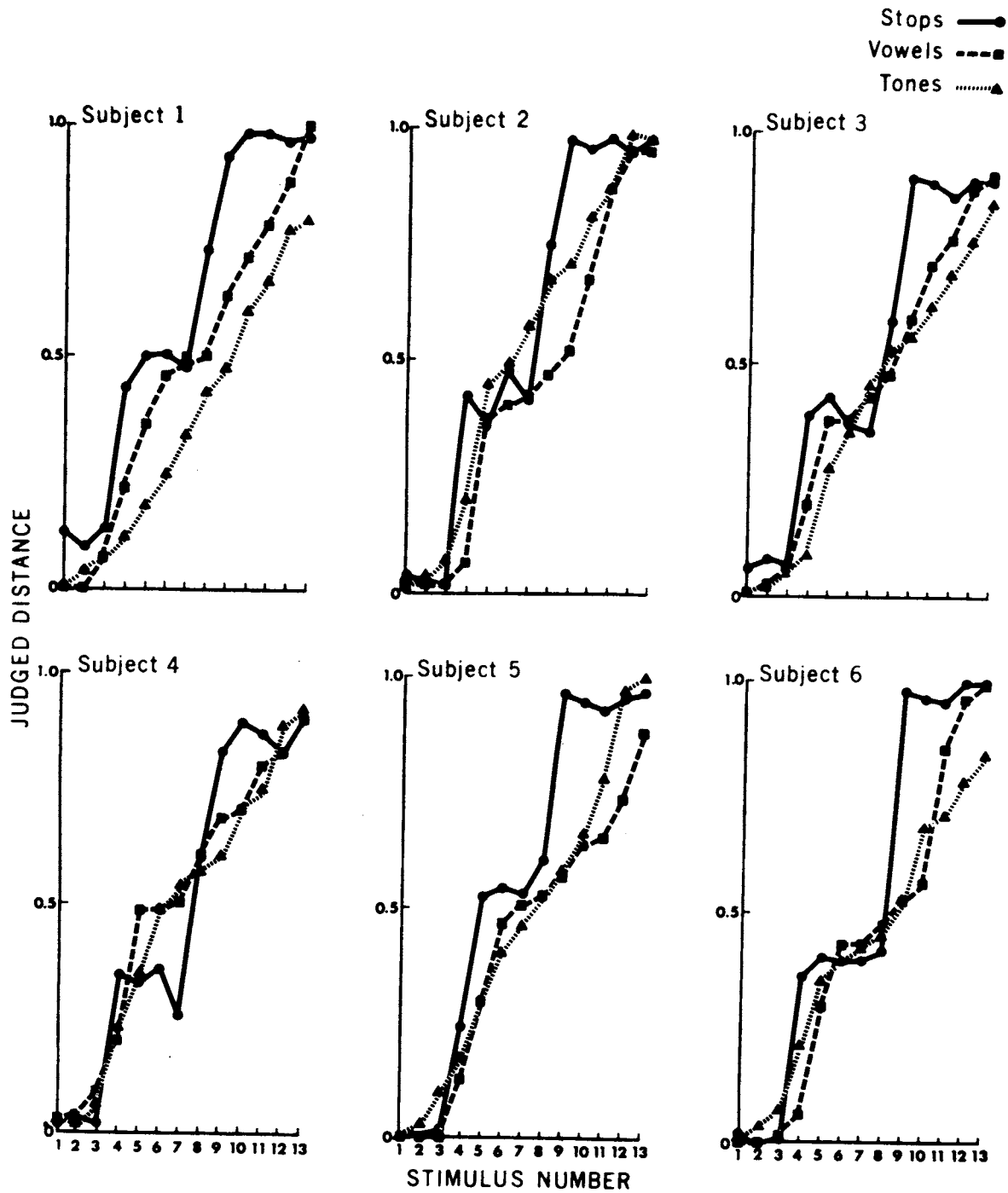
Six Canadian English-speaking undergraduates at a university in Ontario served as subjects. Their ages were between 18 and 21; three were male and three female.

Results and Discussion

In each part of the experiment, the subject judged each of the thirteen stimuli twenty-four times. The judgments were tabulated by measuring the distance of each point marked by the subject from the left-hand end of the line. The mean distance (i.e., judgment) for each stimulus is shown on the ordinates in Figure 1. Results are shown separately for each subject. The ordinates are calibrated so that 0 corresponds to a point marked at the extreme left-hand end of the line and 1.0 to a point marked at the extreme right-hand end of the line. There are clear individual differences, but the curves tend to exhibit certain common features from subject to subject. For stops, the curves seem to be broken into three segments; for tones, they seem essentially continuous; while for vowels, there is a tendency for the curves to be intermediate in form to the other two. Figure 2 shows mean curves calculated over the six subjects.

The clear trends and the consistency from subject to subject seem an adequate answer to one question posed by the study: the speech stimuli are scalable by the method used.

The judged distance along a perceptual continuum
of each member of a series of thirteen stimuli.



Note: 0 and 1.0 represent points on the continuum corresponding to the first and last members of the series. There were three series of stimuli: vowels, stops, and pure tones.

FIG. 1

Pooled Data for Six Subjects

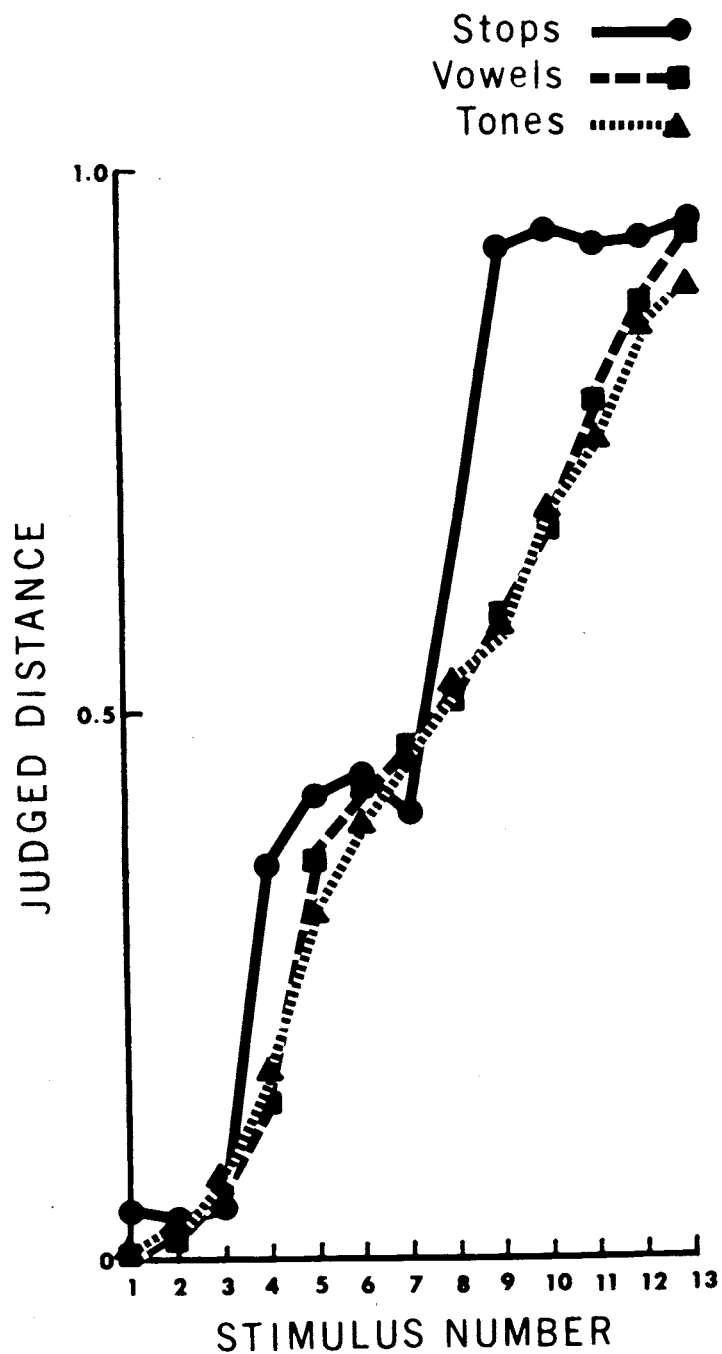


FIG. 2

The speech stimuli were spaced along a physical continuum covering three phonemes so that each sequence cut across phoneme boundaries. Adjacent stimuli either fell within a single phoneme category or fell in a region of transition from one phoneme category to the next. How did these phoneme boundaries affect the judgments? The effect of phoneme boundaries on perception has been reported by a number of workers: Eimas (1963); Griffith (1958); Liberman et al. (1957); Studdert-Kennedy et al. (1963, 1964); Fry et al. (1962); Stevens et al. (1963). The synthetic speech stimuli of the present investigation were previously used in a study by Stevens, Liberman, Studdert-Kennedy and Öhman (1969). These workers established phoneme boundaries for both the vowels and stops. The stimuli in each set showed some overlapping, but the general picture was clear. For stops, the preponderance of identification responses placed stimuli 1, 2, and 3 in phoneme category /g/; stimuli 4, 5, 6, and 7 in phoneme category /d/; and stimuli 9, 10, 11, 12, and 13 in phoneme category /b/. For vowels, a preponderance of responses placed stimuli 1, 2, 3, and 4 in phoneme category /i/; stimuli 5, 6, 7, 8, and 9 in phoneme category /I/; and stimuli 10, 11, 12, and 13 in phoneme category /ε/.

In the present experiment, these phoneme boundaries can be seen to have a marked effect upon the judgment of the stops and a small, perhaps negligible, effect upon the judgment of the vowels. In general, for stops, there was little change in judgment from one stimulus to the next when the stimuli fell in the same phoneme category but a marked change in judgment when the stimuli crossed a phoneme boundary. For vowels, the change in judgment seemed to be more a continuous function of stimulus variation along a continuum and was little affected by the phoneme boundaries. The nonspeech stimuli, the tones, gave results like the vowels, only the curves are somewhat smoother.

These results tend to confirm the findings of Stevens et al. (1969), who also obtained discrimination functions for the stimuli. They showed that discrimination of adjacent members of the stimulus series was poorest when the pair fell at the center of a phoneme category and best when the pair fell in different phoneme categories. This phenomenon was far more marked in the case of the stops than of the vowels. In the case of the stops, discrimination was little better than would have been the case if the subject had been able to discriminate about as well as he could identify. In the case of the vowels, however, discrimination was considerably

better than could be predicted from the identification data; as in the present experiment, the vowel data tended to bear more of a continuous relation to the stimulus variation.

Thus, the two experiments, using very different psychophysical procedures, agree in making a distinction between steady-state vowels and stop consonants that may amount to a difference in mode of perception. The stimuli for the stop consonants tend to be perceived in terms of category of identification to a much greater extent than the stimuli for the steady-state vowels; the difference amounts to a greater limitation on the perception of stimulus differences for stops than for vowels. The vowels seem to be closer to the tones in mode of perception, but it is possible that vowels presented in a context of other speech sounds would behave more like the consonants. The present scaling technique seems to provide a fairly suitable method of comparing the perceptual characteristics of different classes of speech sound.

References

- Eimas, P.D. (1963) The relation between identification and discrimination along speech and non-speech continua. *Language and Speech* 6, 206.
- Fry, D.B., A.S. Abramson, P.D. Eimas, and A.M. Liberman. (1962) The identification and discrimination of synthetic vowels. *Language and Speech* 5, 171.
- Griffith, B.C. (1958) A study of the relation between phoneme labelling and discriminability in the perception of synthetic stop consonants. Unpublished Ph.D. dissertation, University of Connecticut.
- 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. Exptl. Psychol.* 54, 358.
- Liberman, A.M., F.S. Cooper, D.P. Shankweiler, and M. Studdert-Kennedy. (1967) Perception of the speech code. *Psychol. Rev.* 74, 431.
- Stevens, K.N., S.E.G. Ohman, and A.M. Liberman. (1963) Identification and discrimination of rounded and unrounded vowels. *J. Acoust. Soc. Amer.* 35, 1900 (Abstract).
- Stevens, K.N., A.M. Liberman, M. Studdert-Kennedy, and S.E.G. Ohman. (1969) Cross language study of vowel perception. *Language and Speech* 12, 1.
- Studdert-Kennedy, M., A.M. Liberman, and K.N. Stevens. (1963) Reaction time to synthetic stop consonants and vowels at phoneme centers and at phoneme boundaries. *J. Acoust. Soc. Amer.* 35, 1900 (Abstract).
- Studdert-Kennedy, M., A.M. Liberman, and K.N. Stevens. (1964) Reaction time during the discrimination of synthetic stop consonants. *J. Acoust. Soc. Amer.* 36, 1989 (Abstract).
- Torgerson, W.S. (1958) Theory and Methods of Scaling. (John Wiley & Sons, New York).