Vol. 5, Part 4, October-December 1962, pp. 171 - 189

THE IDENTIFICATION AND DISCRIMINATION OF SYNTHETIC VOWELS*

D. B. FRY, ARTHUR S. ABRAMSON, PETER D. EIMAS,** and ALVIN M. LIBERMAN**
University College, London and Haskins Laboratories, New York

A series of thirteen two-formant vowels was synthesized and used as the basis of labelling and discrimination tests with a group of English-speaking listeners. The sounds varied only in F1/F2 plot and the resulting vowel qualities were such that listeners found no difficulty in assigning each sound to one of three phonemic categories, those of the vowels in *bid, bed* and *bad*. The results of the tests were compared with those previously obtained in experiments involving the consonant phonemes /b, d, g/.

It appears from the data that the phoneme boundaries in the case of the three vowel phonemes are less sharply defined than in the case of the stop consonants. The labelling functions for the vowels show a gradual slope and the discrimination functions do not show any marked increase in sensitivity to change in the region of the phoneme boundaries. It is clear also that the listeners were able to discriminate differences very much smaller than would need to be distinguished simply in order to place vowels in the appropriate category. The results show further that the effect of sequence or acoustic context in the perception of vowels is very considerable.

In all the aspects examined in these experiments, the perception of synthetic vowels is found to be different from that of synthetic stop consonants. These differences lend some support to the hypothesis that the degree of articulatory discontinuity between sounds may be correlated with the sharpness of the phonemic boundaries that separate them.

In previous studies of the acoustic cues that may be used in the identification of English consonants, it has been found that, at least for certain classes of consonant, there is a strong tendency for listeners to hear speech sounds in a categorical fashion. The evidence for this has been gained by presenting to listeners a series of synthetic speech sounds, differing from each other by some very small step on a single acoustic dimension, asking them to label each sound in the series with an appropriate phonemic label and also measuring their sensitivity to change in this acoustic dimension. For the consonant group /b, d, g/, for example, a set of stimuli was used in which the transition of the second formant (F2) was systematically varied (Liberman, Harris, Hoffman, and Griffith, 1957; Griffith, 1957). The subjects' labelling of these stimuli placed the sounds in three well-defined classes, corresponding to the three phonemic units, with a sharply marked boundary between each class and the neighbouring one.

* *This work was supported in part by the Carnegie Corporation of New York.*

The measures of discrimination showed that subjects were in fact more sensitive to change in the region of the phoneme boundary than at other points in the continuum of change. The labelling and the discrimination data taken together suggested very strongly that subjects were in fact hearing this series of sounds categorically and the discrimination data were therefore re-examined on the assumption that listeners were able to discriminate only to the extent that they were able to identify the sounds as belonging to different phonemes. It was found that the major variations in the ability to discriminate which would be predicted on the basis of this extreme assumption were actually present in the data but that the general level of discrimination was somewhat higher than purely categorical hearing of the sounds would require.

Later experiments dealing with a variety of acoustic dimensions, but most of them concerning at least in part the broad class of stop consonants, have reinforced this view of consonant perception and lent support to the theory that a listener's own articulatory habits may be an important factor in determining the way in which he perceives speech sounds. In the case of /b, d, g/ there is clearly a discontinuity at the articulatory level between /b/ and /d/ and between /d/ and /g/; it is not in fact possible for a human speaker to produce a series of sounds that changes smoothly from /b/ to /d/. The tendency to hear such sounds categorically may well be connected, therefore, with the existence of such articulatory discontinuities. If this were so, then we should expect that there will be marked differences between various classes of speech sound as to the way in which they are perceived, since there are certainly differences in the degree of articulatory discontinuity.

We are of course concerned here only with groups of sounds in which change with respect to a single acoustic dimension is a sufficient cue for phonemic differences. Within this limitation, there are cases such as change in the mid-point of a noise and as a cue to the difference between /s/ and /ʃ/, and change in third formant transition as a cue to /l/ and /r/, where the articulatory discontinuity is very much less obvious than in the change from /b/ to /d/. We might therefore expect to find that in these instances perception might also be less categorical in character. There is, however, a class of sounds, the vowels, in which continuous articulatory change from one member of the class to another is possible (at least insofar as English vowels are concerned) and vowels therefore provide an excellent testing-ground for the hypothesis that categorical hearing of speech sounds and articulatory discontinuity are related to each other.

It is well-established that the frequencies of the first and second formants (F1 and F2) taken together are a sufficient cue for vowel differences and that these frequencies are very largely dependent on tongue articulation. Analytical acoustic studies of vowels have shown that F1 and F2 vary over a wide range in the case of a single vowel uttered by many different speakers and that there is considerable overlap in the plots of F1 and F2 when data for all the vowels in the English system are taken together (Peterson and Barney, 1952) Since the correlation between F1 and F2 and tongue articulation is close, this overlapping in the acoustic data indicates that vowel articulation not only can be changed continuously but that it in fact is so varied

if we take into account utterances from different speakers in different contexts. The experiments reported in this paper represent a preliminary attempt to explore the perception of synthetic stimuli in which F1 and F2 are varied systematically in such a way as to cover the range corresponding to several vowel phonemes. The stimuli were used as the basis for labelling and discrimination experiments similar to those carried out for /b, d, g/ and the results constitute in the first place a contribution to the study of the relation between articulation and perception.

There are, however, a number of other reasons for attempting this kind of perceptual experiment with vowels. The consonant-vowel dichotomy has appeared in discussions of language from the most ancient times and has persisted down to the present despite some questioning in recent years of the validity and the necessity of the distinction. If the two classes of sound fulfil different functions in speech we may expect to find differences correlated with their occurrence at a number of different levels, among them the perceptual level. Experimental results showing that listeners perceive vowels and consonants in different ways would form at least contributory evidence to suggest that the two classes are functionally distinct.

One important respect in which vowels and consonants differ seems to be in their informational loading. This is indicated in a qualitative way by the observation that English speech in which all vowel distinctions have been artificially eliminated is many times more intelligible than English in which all consonant differences have been removed. The same principle is recognized in the alphabetic spelling of such languages as Hebrew where all the essential information is conveyed by the consonant letters. Reliable quantitative studies of this difference would need to be based on computations of the number of possible choices available at succeeding points in actual phonemic sequences and a comparison of the numbers for vowel and consonant phonemes, having regard to the total information content of the sequence. Data concerning this aspect of English will shortly be available as a result of computer work on phonemic transcriptions of English speech (Denes, unpublished). Meanwhile one could argue theoretically that if the loading of consonants were high, then the most important part of consonant reception by the listener would consist in placing the sound in its appropriate phonemic category; it would be essential that this operation should be done accurately and also quickly, requirements which would be efficiently met by the kind of categorical hearing for which there is already a good deal of evidence in the case of consonants. It is quite clear that vowels, whatever their loading from the phonemic point of view, carry other kinds of information. They are the principal vehicle for rhythm and intonation, they carry the voice quality of the speaker, convey his emotional state and, in English especially, provide most of the information about dialect. All these kinds of information are borne by relatively long time segments and hence are delivered at rates very much slower than the phonemic rate. Rhythm and intonation patterns occupy a span usually equivalent to that of a number of phones; the listener's appreciation of a speaker's dialect is the cumulative effect of hearing a number of vowels over and over again, a process which may take a matter of minutes. For these purposes, rapid and accurate

categorisation of the vowels is not important and if it should turn out that the loading of the vowels with respect to phonemic information is relatively low, then there would be good grounds for expecting that vowels should be perceived rather differently from consonants.

The differences in vowel quality that must be perceived in comparing the vowels of one dialect with those of another are very much smaller than the differences between the constituent vowels of one speaker's vowel system so that in making judgments of dialect a listener is discriminating sub-phonemic vowel differences which would be imperceptible to him if vowels were perceived in the categorical way that has been found to hold for stop consonants. Something much more in the nature of continuous hearing seems to be called for in the case of the vowels where in general it is possible to trade speed for fineness of discrimination. In comparing the perception of stop consonants and vowels we may, therefore, be dealing with opposite ends of a scale ranging from rapid, categorical and hence relatively coarse hearing at one extreme to relatively slow, continuous but highly discriminating hearing at the other.

One further difference is implied in the contrast between categorical and continuous hearing. In order to provide a basis for quick and accurate decoding, the former must deal in categories which are relatively fixed and independent of context. If context exerts a great influence on the perception of sounds, this must lead either to a great number of errors in decoding or to the need for more time in which to take in and allow for the nature of the neighbouring sounds. The results of some experiments with stop consonants (Eimas, 1962), which we shall have occasion to discuss later in this paper, indicate very little influence of context. In labelling stimuli as /b, d, g/, for example, listeners' judgments of a given stimulus were not much affected by the preceding stimulus; the categories they were using seem to be rather sharply defined and most stimuli fell clearly into one or another of them. In any case where the listener is not functioning in this way, where he is discriminating comparatively fine differences and dealing with a continuum rather than discrete classes, that is to say in what we have referred to as continuous hearing, we should expect that context might play an important part. There will be a tendency for the subject's judgment to be determined largely by the fact that x is 'light' compared with y, or that z is 'dark' compared with x rather than by a longer-term conviction that x and y are in the class 'light' and z in the class 'dark'.

There is already some evidence that the perception of vowels is greatly dependent on context. Experiments by Ladefoged and Broadbent (1957) have shown that subjects' identification of the vowel in an English monosyllable can be influenced by the formant patterns used in a preceding carrier sentence. These results not only demonstrate an effect of context on vowel perception but also support the view which has been generally held for a very long time that in dealing with vowels uttered by a particular speaker, listeners rapidly form an appropriate reference frame against which they judge the quality of and identify the sounds which occur. The reference frame is readily changed when utterances from another speaker are received and it

is clearly dependent on judgments of the relations between vowel qualities. Essentially, this is a matter in which context is bound to exert considerable influence and the categories that the listener is using will be shifting classes determined by the interrelations within a system rather than well-defined absolute categories. This is not to say, of course, that listeners experience serious difficulty in placing vowel sounds into phonemic categories. In the experiments by Broadbent and Ladefoged, for example, the subjects showed no hesitation in selecting the syllable that they heard, and in general listeners are able to assign vowel sounds to phonemes. The important point is that the particular phonemic category selected is dependent on context, that is more specifically on the vowel reference frame which is operative for the listener at the time of reception.

In these circumstances, then, we expect the identification of vowels in a labelling test to be rather dependent on context and if this effect is strong enough there will be trends in the experimental data to indicate that the sequence in which the stimuli are presented to the subjects is a factor of some weight. A later section of this paper gives an account of a method of treating the data so as to find out whether this factor is important.

Procedure

The purposes of this experiment required, first, that we have as stimuli a series of synthetic vowels that vary along an articulatory and acoustic continuum from one phoneme to another. The relevant data are obtained, then, by presenting these vowels to listeners (1) for identification as phonemes and (2) for discrimination on any basis whatsoever. In this way we determine whether or not there are peaks in discrimination at the phoneme boundaries and, also, to what exent the listener can or cannot hear intra-phonemic differences. In general we were at pains to make the procedures of this experiment correspond as closely as possible to those of earlier studies on consonant perception; this was done in order that the results of the several studies might the more easily be compared.

Stimuli

We chose to synthesize /ɪ/, /ɛ/, and /æ/, and to divide the space between them so as to have a total of 13 stimuli. Our aim is best described by asking the reader to imagine a two-dimensional acoustic space whose coordinates represent the frequencies of the first and second formants, scaled logarithmically, and then to consider that our stimuli would, ideally, lie at equal distances along a straight line drawn through the points at which /ɪ/, /ɛ/, and /æ/ are located.

The vowels were synthesized on a machine called "Alexander", a formant-type terminal analogue synthesizer designed and built at the Haskins Laboratories.[1] It can

[1] *No technical description has been published as yet, but the general design is similar in many respects to other formant-type synthesizers in use elsewhere. For a discussion of them, see Fant (1958).*

TABLE 1

Formant Frequencies of the Synthetic Vowel Stimuli

STIMULUS NUMBER	FIRST FORMANT	SECOND FORMANT
1	330	1980
2	380	1970
3	410	1960
4	460	1930
5	490	1910
6	500	1890
7	550	1880
8	580	1860
9	650	1860
10	700	1820
11	780	1820
12	830	1780
13	890	1760

be controlled manually for steady-state sounds or by means of a pattern on an optically scanned acetate loop for the synthesis of running speech. It has four formant generators connected in parallel. Only two of them, excited by buzz pulses of variable repetition rate, were used for the present study. Smooth onsets and offsets of vowel amplitude and fundamental frequency envelopes were obtained by circuitry that gave an exponential rise and decay.

Appropriate formant frequencies for the three vowels were obtained by reference to data available in the literature (e.g., Peterson and Barney, 1952), supplemented by the results of our own exploratory work. There is, of course, some error and uncertainty in the control of the synthesizer, just as there is in the measurement of the sounds the synthesizer has produced. After synthesizing the thirteen vowels that were to constitute the stimuli of the experiment, we measured the formant frequencies by inspection of wide- and narrow-band spectrograms as well as narrow-band sections made on the Kay Sonagraph. A 1200 cps./inch scale was used for better visual resolution than the standard 2000 cps. can give. The harmonics of a complex wave of 400 cps. were used for frequency calibration. After arriving at formant frequencies this way, we then repeated the procedure with another set of spectrograms and sections for each of the variants. Table 1 gives the averages of the two sets of measurements of the formants of the synthetic vowels. These numbers have been rounded to the nearest 10 cps. as a realistic estimate of the attainable precision. We estimate the formant band-widths to be about 100 cps. throughout. The average difference in intensity between F1 and F2 in the same stimulus in 8db. The difference in over-all intensity between stimuli occurring in one triad does not

The spectrographic examination of the stimuli revealed that we did not always succeed in coming as close to the intended frequencies as we might have wished. In particular, the difference between numbers 5 and 6 is considerably smaller than it was supposed to be.

Measurement of discrimination

A forced-choice ABX method was used to determine how well the listeners could discriminate the synthetic vowels. In this method the stimuli are arranged in triads, the first (A) and second (B) being always different, and the third (X) being always identical with the one or the other. The listeners' task is to determine whether X is identical with A or with B and to guess if necessary.

We undertook to measure discrimination between each stimulus and those that were one, two, and three steps away from it on the stimulus scale. This made a total of 33 A-B pairs. Each ABX triad was arranged in all possible permutations (ABA, ABB, BAA, and BAB) to counterbalance series or order effects. Given 33 A-B pairs and four ABX permutations of each one, there was, then, a total of 132 ABX triads. These triads were presented to the subjects in random order, the number of presentations being such as to provide 20 judgments per stimulus comparison (A-B pair) for each subject. Three of the subjects worked longer and made a total of 40 judgments of each stimulus pair.

Phoneme identification

To determine how the various stimuli were assigned to the three phoneme classes (/ɪ/, /ɛ/, and /æ/), we presented to the subjects the same ABX arrangements of stimuli that had been used in the discrimination tests, but instead of asking whether X was identical with A or with B, as we had in measuring discrimination, we instructed the subjects to label each stimulus as /ɪ/, /ɛ/, or /æ/. No other responses were accepted, and the subjects were asked to guess if necessary. The triads were presented in random order. Obviously the stimuli near the middle of the continuum will appear more often than those near the ends; the number of judgments per stimulus varied accordingly from 84 to 184.

Subjects

Eight paid volunteers attending the University of Connecticut summer school served as subjects. They had been selected from a group of 17 on the basis of a special pre-test in which they had been found to be most consistent in applying phoneme labels to the synthetic vowels.

RESULTS

In presenting the results of the vowel experiments it is necessary to stress once more that these are preliminary attempts at studying the discrimination and labelling of vowel sounds and the results too must be considered as being preliminary. We

shall have occasion to compare the vowel data with those for stop consonants, particularly /b, d, g/, and it will be immediately apparent that the former are very much less tidy than the consonant data. There are several reasons for the greater amount of noise in the vowel data. The first is to be found in the experimental stimuli themselves since past experience has shown that a great deal of experimentation is necessary before one can produce the best synthesised sounds for a given set of discriminations. The vowel-like sounds used in the present series of experiments were adequate but no more, and it should be possible in future work to effect some improvement in the stimuli. More exact specification and closer control of formant frequencies, intensities and band-widths is needed, the signal to noise ratio in the test tapes should be improved and it is advisable to examine the desirability or otherwise of introducing constant third or higher formants and to determine the most suitable time course for over-all intensity and fundamental frequency variation.

A second reason for the scatter of the results is also connected with the stimuli but is one which is inherent in the nature of the experiment. A pre-requisite for this type of measurement is that we should find a single dimension such that variation with respect to it is a sufficient cue for phonemic differences. In the case of the vowels the F1/F2 plot may justifiably be considered as forming a single dimension and thus satisfies the conditions but the situation is nonetheless different from that encountered in the consonant experiments. The variation of F2 transition in the /b, d, g/ case gave rise to a set of stimuli which cued the recognition of these three consonants but which nowhere in the progression suggested to the listener some other English consonant; it was, as it were, a linear sequence. An F1/F2 plot, on the other hand, is a point on a plane on which it is possible to define an area bounded by extreme values of F1/F2 either for all vowels occurring in a given language or, more generally, for all vowels that can be produced. Such an area represents a physical projection of a space containing a great range of vowel qualities perceived by listeners. The correlation between the physical and the perceptual is, of course, not simple and when we set out to find a series of F1/F2 plots forming a progression from /ɪ/ through /ɛ/ to /æ/ we cannot take it as axiomatic that these values will lie on a straight line in the physical space, nor even that they will necessarily lie on a smooth curve. We know already from analytical data that the sounds corresponding to a given phoneme will cover a considerable area in the F1/F2 space so that there will clearly be a number of paths which will form a progression from one vowel to another and to determine the best path for the purposes of labelling and discrimination experiments is an empirical matter. Further, even assuming that we have found the optimum progression, there is the added difficulty that any deviations from this through small errors in setting up the stimuli may produce sounds which suggest to the listeners some vowel other than the three which are the basis for the labelling.

Such difficulties as these were in fact encountered in the preparation of the test [illegible] in the experiments reported here. They represent in themselves a particular

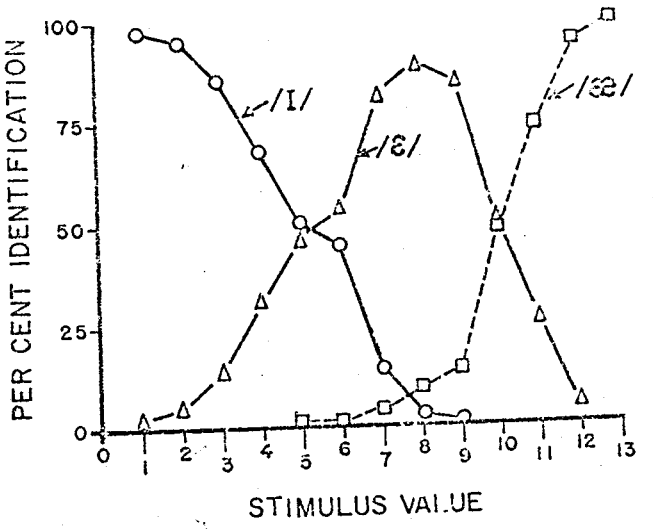

Fig. 1. Identification of the synthetic vowel stimuli as English /ɪ/, /ɛ/, or /æ/ as a function of first and second formant-frequency combinations as shown in Table 1. The data are the pooled responses of all eight subjects.

facet of the fundamental reason for the scatter of the data which is simply that subjects do find it difficult to label vowel sounds consistently. The variability of the vowel sounds produced by speakers, reflected in the large area of the F1/F2 space occupied by the sounds belonging to a single vowel phoneme, is matched by a corresponding inconsistency on the part of listeners when they are asked to label vowel sounds. As we shall see from the results and in the subsequent discussion, vowel categories are not as sharply defined nor as absolute as some consonant categories and vowel judgments are very highly susceptible to the effects of context. In these circumstances, improvements in the stimuli in future work may be expected to get rid of some of the variability in the data but there will remain that part of it which is, it seems, inherent in the judgments we have been trying to investigate.

Vowel identification results

Fig. 1 shows the pooled responses of eight subjects to the vowel identification test. The stimuli evoked a considerable number of identifications in each of the phonemic categories, though there were in all rather fewer judgments in the /æ/ category than in either of the other two. Stimuli 1 and 13 set an artificial boundary to the /ɪ/ and /æ/ categories and we cannot, of course, make a valid comparison of the extent of all three categories unless the range of stimuli is enough to take listeners into a fourth and fifth category at either end. The range used appears to be quite satisfactory for the purpose of these experiments, that is to show boundaries between the /ɪ/ and /ɛ/ and the /ɛ/ and /æ/ phonemes, and the judgments are not unduly weighted in favour of any one phoneme.

It is, perhaps, worth noting that the labelling functions of Fig. 1 are less sharp than those that have been obtained when comparable procedures are carried out with synthetic stops (Eimas, 1962). Such a comparison reveals another difference in that in the case of the consonants the degree of agreement or consistency reaches a high level in all three categories whilst in the vowels the middle category does not produce as nigh a level as the other two. These are indications of the effect of context on vowel judgments, which will be discussed in a later section. It will be enough to point out here that the context effect in vowels works by *contrast* which means, if we express it in terms of phonetic classification, that a given vowel sound will appear more open when preceded by a sound closer than itself and more close after a sound more open than itself. In the case of the stimuli used in these tests, every sound in the series strikes the listener as being more open than stimulus No. 1 so that the effect of context here is to increase the number of judgments that No. 1 is /ɪ/ ; similarly, all sounds appear closer than stimulus No. 13 and this increases the judgments that No. 13 is /æ/. For all the stimuli labelled as /ɛ/, however, there are included in the test some sounds that make them appear closer and others that make them appear more open. Hence these stimuli are labelled less consistently and this fact is reflected in the labelling function for /ɛ/. More generally, of course, all the vowel stimuli other than those at the extreme of the continuum will tend to be labelled according to the context in which they are presented ; this will reduce the consistency with which the labels are applied and thus produce the sloping functions of Fig. 1 rather than the more nearly quantal functions found with the stops.

Vowel discrimination results

The curves of Fig. 2 show the percentage of correct responses in discriminating between stimuli which differ by one, two and three steps. The series of stimuli are set off on the horizontal axis in arbitrarily equal steps and in the same order as in the case of the labelling data. The continuous curve in each part of the figure indicates the pooled results for all subjects. It has been noted above that the step between stimuli 5 and 6 was considerably smaller than other steps and this accounts for the fact that at this point in the graph the level of discrimination falls to nearly 50%. The mean level of discrimination for the one-step differences is however very close to 75%, which would normally be taken as a threshold criterion. When the difference between test items is as large as two steps on the stimulus scale, the percentage of correct responses is very near to 100%, and there is very little room for improvement on this in the case of the three-step differences.

In the course of previous work on the discrimination of consonants a model has been developed which enables us to consider to what extent discrimination is influenced by categorical perception and to predict from labelling data the level of discrimination to be expected if subjects were able to discriminate only to the extent that they could place the same stimuli consistently in phonemic categories. In the case of a variety of consonantal discriminations it has been found that the discrimination

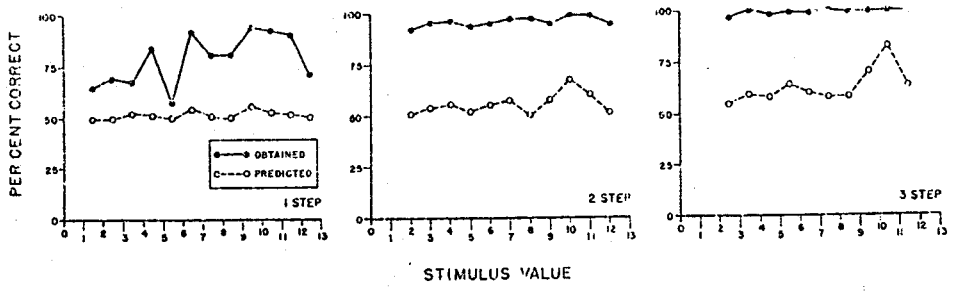

Fig. 2. Obtained and predicted discrimination for one-, two-, and three-step differences among the synthetic vowel stimuli. The data are the pooled responses of all eight subjects.

functions obtained experimentally lie quite close to the functions predicted on the basis of labelling data. Usually the subjects' level of discrimination is slightly better than the predicted level but the major inflections in the predicted functions, which appear at the region of the phoneme boundaries, are found also in the experimental data. The broken lines in Fig. 2 represent the discrimination function for the vowels predicted in the same way, that is to say on the assumption that subjects can discriminate only as well as they can label. In the case of the one-step discriminations, the predicted scores do not differ materially from the chance score of 50%. For the two-step differences, the level of the predicted scores rises a little and there is a pronounced maximum for the discrimination of stimulus 9 from stimulus 11, the sounds which lie closest to the phoneme boundary indicated in the labelling data. This maximum is, of course, still more marked in the predicted scores for the three-step differences but for other parts of the range the predicted scores remain quite low. In the obtained scores for the one-step discriminations, the low level of discrimination at stimulus 5 rather confuses the picture, but there is no doubt that the obtained level is far above the predicted level. At two and three steps the difference between the obtained and the predicted scores is even more striking and the discrimination is so good throughout the range of stimuli that there is no room for improvement in the region of the phoneme boundaries. On the basis of these data we have to say that the perception of the vowels is continuous rather than categorical. There is no evidence of discontinuities in the discrimination functions at phoneme boundaries. More generally, it is clear that discrimination is much better than that which is predicted on the extreme assumption that the listeners can only hear phonemically (i.e., categorically).

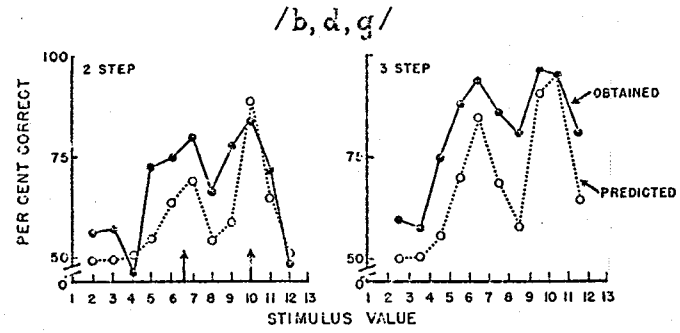

Fig. 3. Group obtained and predicted discrimination functions for the /b, d, g/ stimuli at two and three steps. Reproduced from Eimas (1962), Fig. 4.

There is a marked difference between this result and the discrimination data for /b, d, g/ obtained by Eimas (1962) and shown in Fig. 3.[2] We should remark here that the comparison between the vowel data and the results of Eimas' study is a reasonable one. In both cases there were three phoneme classes; these three classes were divided into fourteen stimulus steps in the case of the stops and thirteen in the case of the vowels. It is clear from Fig. 3 that the perception of the stops tends to be categorical. There are obvious peaks in discrimination in the regions of the phoneme boundaries, and in general discrimination is very little better than is predicted on the basis of the extreme assumption that the listeners can only discriminate as well as they can apply the three phoneme labels. Thus, in the case of the stops the subjects take very little notice of anything but phonemic distinctions, while in the vowels, on the other hand, they discriminate quite well between sounds within the same phonemic category. It was suggested earlier that in order to recognize dialectal differences a listener would in fact need to take account of differences in vowels which are considerably smaller than the phonemic differences within the system of a single speaker. The experimental results provide strong evidence that listeners are well able to do this.

The effect of context on identification

In discussing the identification data we have already mentioned that the sequence in which sounds were presented for labelling was a factor which influenced the results. The same test tapes were used in both the labelling and discrimination experiments so that many of the stimuli for identification were heard in quick succession (at intervals

[2] *This part of Eimas' study, which was intended to provide a basis for certain other comparisons between continuously and categorically perceived stimuli, was much like the earlier experiments of Liberman, Harris, Hoffman, and Griffith (1957) and Griffith (1957). Eimas' procedures were more like those of the present study, however, in that the stimuli were presented for labelling in ABX triads and the results for all subjects were pooled.*

of one sec.) and in groupings (ABA and ABB) which would tend to maximise the effect of context. The rather gradual slope of the identification functions, the high level of consistency in labelling the end-points of the range of stimuli, the relatively inconsistent labelling of the middle phoneme category, it has been suggested, may all be signs that context is playing an important part in identification. In the present section we shall consider further treatment of these data intended to give us some idea of the magnitude of the effect.

Let us consider first the case in which a listener hears a sound, *x*, paired with another sound, *y*, that is, either followed or preceeded by *y*. If context or sequence has some effect on what he perceives, it must operate in one of two directions: either *y* will seem more unlike *x* because of its proximity in time, or it will seem more like *x*. In the first case we should say that context was working in the direction of *contrast* and in the second, of *assimilation*. Previous work on the perception of vowels suggests that the effect is more likely to be one of contrast. In the experiments reported by Ladefoged and Broadbent, for example, the lowering of F1 in the carrier sentence made the test item sound more open in quality, that is as though it had a higher F1.

The stimuli for the vowel experiments, which are numbered from 1 to 13 in Table 1, are ranged in order from the most /ɪ/-like to the most /æ/-like vowel, that is from the closest sounding to the most open sounding. If the effect of context was in the direction of contrast, then any stimulus which in the test sequence was paired with a stimulus of lower number than itself would tend to sound more open; when paired with a stimulus of higher number it would sound closer because of this proximity. We can obtain a very simple and crude measure of the context effect therefore if we examine the labelling responses for each stimulus in the range and break the judgments down into two groups, those made when the stimulus was paired with another of higher number and those made when paired with one of lower number. The greatest effect that context could possibly have would be exemplified when for a given stimulus, *x*, all the labelling judgments were swung in a particular direction when *x* was paired with higher numbered stimuli and in the opposite direction when paired with lower numbered stimuli. The way in which this simple measure was applied can best be seen from hypothetical examples such as those shown in Table 2.

In the first example we suppose that of all responses to stimulus *x*, 50% labelled it as /ɪ/ and 50% as /ɛ/. We now divide the responses into those made when *x* was paired with a stimulus having a higher number and those when it was paired with one with a lower number. Assume now that every time that *x* was paired (in the ABX triads) with a stimulus having a higher number, the label applied to *x* was /ɪ/ and that every time *x* was paired with a stimulus having a lower number, the label was /ɛ/. This would represent the maximum contrast effect, expressed in our example by subtracting the percentage of /ɪ/ judgments when *x* was paired with higher valued stimuli from that when *x* was paired with lower valued stimuli, giving the value 100 (the positive sign has been chosen arbitrarily to denote contrast). In

TABLE 2

Hypothetical Examples of Maximum Context Effect

	Example 1		Example 2	
	/ɪ/	/ɛ/	/ɪ/	/ɛ/
Total response distribution	50%	50%	50%	50%
Response distribution when paired with higher numbered stimuli	100%	0%	0%	100%
Response distribution when paired with lower numbered stimuli	0%	100%	100%	0%
Maximum context effect	100%		−100%	

	Example 3		Example 4	
	/ɪ/	/ɛ/	/ɪ/	/ɛ/
Total response distribution	60%	40%	60%	40%
Response distribution when paired with higher numbered stimuli	100%	0%	20%	80%
Response distribution when paired with lower numbered stimuli	20%	80%	100%	0%
Maximum context effect	80%		−80%	

the second example we have the same total distribution of responses, but the effect of context is now reversed since *x* is labelled /ɛ/ whenever paired with stimuli above it and /ɪ/ when paired with stimuli below it. By the same procedure as before this gives the value of −100 for the context effect, which represents maximal assimilation.

It will more frequently happen that /ɪ/ and /ɛ/ are not equally divided in the total distribution of responses and examples 3 and 4 show maximum context effects computed for the case of some other division of the total. Stimulus *x* is paired equally with higher and lower numbered stimuli. If now 60% of the total responses are /ɪ/, the maximum contrast effect would mean that *x* was labelled /ɪ/ in all instances when paired with higher numbered stimuli and also in 20% of the pairings with lower valued stimuli. We arrive at the value for maximum context effects in the same way as before, obtaining the value of 80 (contrast). In example 4, the effect is again reversed and maximum context effect is −80 (assimilation).

In dealing with the data, we have used this method to compute both the maximum possible and the actual effect of context for the various synthetic vowels.[3] The actual

[3] *Context effects of the kind described here could not be calculated for the stimuli at either end of the continuum, since these could be paired only with stimuli lying to one side or the other.*

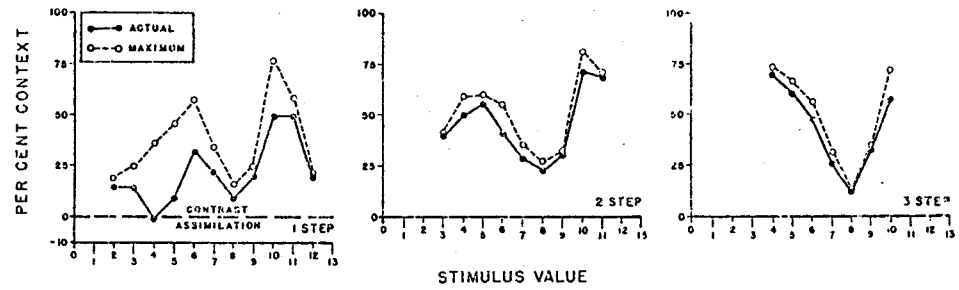

Fig. 4. Actual and maximum context effect for one-, two-, and three-step differences, averaged over all eight subjects.

effect can vary between zero and the maximum possible; the extent to which the actual effect approximates the maximum possible is a rough measure of the influence of context on perception.

In Fig. 4 the broken lines show the maximum, and the continuous curves the actual context effects for one-, two-, and three-step differences. Clearly there was an effect of context, and, just as clearly, the effect was consistently in the direction of contrast. By comparing the actual effect with the maximum one sees, further, especially in the two- and three-step cases, that the context effect was very large in relation to what it might have been. This is to say that wherever there was variation in the response to a vowel, that variation was not primarily random, but was rather determined almost entirely by the context in which the vowel was judged.

To determine how great these relative magnitudes of context effect are it is instructive to compare the results of this experiment on the vowels with results of similar studies of some consonants and certain simple non-speech stimuli. Eimas (1962) has made just such comparisons, using the voiced stops /b, d, g/, a set of cards of varying reflectance, and various durations of white noise. He found the effect of context in the perception of the stops to be much smaller than that which we have observed here with the vowels. In the case of the non-speech stimuli the context effect was rather variable, but in general it was less than the vowels and greater than the stops.

DISCUSSION

On the evidence of the results presented in this paper there are clearly grounds for believing that listeners perceive vowels in a way somewhat different from that in which they perceive the stop consonants. The chief basis for this conclusion is to be found in the fact that the ability to discriminate vowels is far superior to what would be required merely in order to assign them to [illegible]

appears to be no particular sharpening of sensitivity in the region of phoneme boundaries, whereas discrimination of /b, d, g/ was very little if at all superior to labelling and showed the predicted increase in sensitivity near the phoneme boundaries.[4] It seems that we may here be dealing with two extremes in speech perception, in the case of the stops with perception that is maximally categorical and in the vowels with the extreme of continuous perception. In experiments involving plosive consonants it has often been commented upon by both experimenters and subjects that when one listens to the stimuli there indeed seems to be a rather sharp switching over between categories. This was particularly noticeable in the /b, d, g/ experiment (Liberman *et al*, 1957), the /d, t/ experiments (Liberman, *et al*, 1961a), and the /sI, spI/ experiments (Bastian, Eimas and Liberman, 1961). If the vowels and the stops do represent extremes in this matter, then one would expect that some other classes of sound may be perceived in a manner which falls somewhere between the categorical and the continuous, and it would obviously be profitable to explore other cases of labelling and discrimination from this point of view. It might be particularly interesting in this connection to study the fricatives, nasals, and liquids.

In the introduction to this paper it was suggested that some association was to be expected between categorical perception and the effect of context on perception. Sharply defined and strongly marked categories would be rather in the nature of absolute categories and in assigning sounds to them a listener would be rather little influenced by temporal sequence. In this respect, too, the vowels present an extreme case, for the influence of context on the vowel judgments is almost as great as it can possibly be in the conditions of the experiments. Indeed, as we pointed out earlier, the effect of context is greater on the vowels than on simple non-speech stimuli. Thus it appears that listeners have a strong tendency to judge a vowel by comparing it with one they have just heard. This fact lends support to the idea which has often been expressed in the past (see, for example, Joos, 1948; Broadbent and Ladefoged, 1960) that in dealing with vowels a listener establishes for himself a frame of reference appropriate to a given listening situation. In particular he is likely to set up a reference system of vowel qualities for an individual speaker and to judge all vowel sounds occurring in this individual's speech by comparing them with the reference values. A listener who is decoding the speech of a total stranger usually manages in quite a short time to erect the reference frame by making use of the redundancy of the language, which in most cases will determine with a very high probability what vowel 'must' have occurred, and then judging a particular vowel quality by comparison with those that have preceded it. Nevertheless, when faced with speech showing very marked features of a dialectal pronunciation far removed from his own, a listener may require to hear some considerable stretch of speech before he is able with certainty to assign a vowel to the correct category in sequences where redundancy does not resolve all ambiguities. It is only by paying a good deal of attention to the sequence of vowel qualities and by remembering them

[4] *Related studies on phonemic tones (Abramson, 1961) and phonemic vowel duration (Bastian*

that he builds the frame of reference.

The situation with regard to the stop consonants is plainly different. Even when the /b, d, g/ stimuli were presented in triads, as in the experiment by Eimas (1962) referred to earlier, the effect of context on the judgments was very small and certainly nowhere near to the maximum possible effect. The categories here appear to be rather more absolute and classification is more or less independent of sequential effects. There can be no doubt that the series /b, d, g/ includes marked discontinuities on the articulatory plane and, as we have already said, the association of categorical perception and articulatory discontinuity in this class of consonants formed the starting point for the theory that perceptual and articulatory continuity or discontinuity might be linked. Vowels present clearly a case of articulatory continuity and the present results, as far as they go, indicate continuous perception[5] but this evidence is not of course critical for the theory. It would be quite possible for articulatory continuity to be associated with well-defined categories established on some other, perhaps purely perceptual, basis. On the other hand, if it could be shown that continuous perception in speech was possible for a series of sounds which included marked articulatory discontinuities, this would certainly be a contra-indication to the theory. Material for such critical experiments might be found in the fricative series in English since it contains varying degrees of articulatory continuity. The change from /ʃ/ to /s/ involves very little break, there is a rather more decided one between /s/ and /θ/ and an even greater one between /θ/ and /f/. The difficulty here is that no single cue has equal weight in all these discriminations, as has been shown in experiments with naturally produced speech (Harris, 1958). It would therefore be difficult to generate a completely satisfactory series of stimuli to cover the whole group of fricatives, but it might be possible in separate experiments to discover how closely articulatory discontinuity is linked with sharpness of the phoneme boundary.

It has been suggested in previous papers that many of the data hitherto obtained concerning the perception of speech sounds show that this is a rather special kind of perception. In all cases where there was evidence of well-marked categories, the discrimination function showed peaks in the region of the phoneme boundaries and the general level of discrimination was such as to suggest that listeners could discriminate between sounds only very little better than they could place these sounds in appropriate phoneme categories; they could in fact distinguish only about as many different sounds as they could identify. This result is very different from those obtained in psychophysical experiments with non-speech stimuli. It has generally been found that in judging stimuli varied with respect to a single dimension, subjects are able to discriminate many more stimuli than they can identify with certainty; they can distinguish more different pitches or brightnesses than they can label correctly. The comparison between speech and non-speech stimuli has been directly made in experiments already reported (Liberman *et al.*, 1961a, 1961b) in which the discrimination of speech stimuli was compared with that of non-speech stimuli which were

[5] *As do the studies on phonemic tones and phonemic vowel duration (fn. 4). These phonemic distinctions also lie along articulatory continua*

as physically similar to the speech stimuli as it was possible to make them. The results showed two important differences between the two types of perception : first, for the non-speech sounds there was not the degree of fluctuation in discriminability that was found with the speech sounds and second, discrimination was generally worse for non-speech stimuli than for comparable speech stimuli. This led to the conclusion that the sharpening of sensitivity to change at the region of phoneme boundaries must be the result of linguistic training in the listener.

Experiments with non-speech controls have so far been limited to the class of stop consonants (discrimination of /d, t/, /sl, spl/, /p, b/) but it is significant that in each case discrimination of speech sounds has proved to be superior to that for the non-speech controls. It is clear that listeners have learned to make the speech discrimination as a result of linguistic training. In the case of vowels, there are certain difficulties inherent in obtaining suitable controls in order to find out how far the discrimination of vowels may be the result of training, particularly where the physical variable is to be formant frequency. A transposition of the formant frequencies on the frequency scale is likely to result in sounds which are either still too speech-like or too musical. Discrimination of the latter might not be influenced by linguistic training but would be dependent on musical training and this factor might mask the effect that was being studied. It is quite possible that suitable control stimuli for vowel experiments may be found empirically but meanwhile little comment can be made about the effect of linguistic training on vowel discrimination except in one respect, that is with regard to the evidence provided by the measures of context effect.

It was pointed out in the section on results that the influence of context on the perception of stop consonants was very weak and certainly less than in the case of the non-speech stimuli that were examined. For the synthetic vowels, on the other hand, the effect of context was if anything greater than for the non-speech stimuli. This suggests the very interesting conclusion, if this observation proves, in the light of further experimental evidence, to be well-founded, that linguistic training in the case of vowel perception may include learning to make the maximum use of context. It is possible that English speakers, at least, learn both to make the necessary vowel phonemic distinctions and to appreciate the sub-phonemic differences involved in the recognition of different dialects, basing both operations on the maximum use of short term memory for vowel quality. In other words, they learn to shift the frame of reference for vowels very frequently and very rapidly under the influence of sounds heard in the immediate past. If this were indeed so, vowel perception would not only present a very special case of perception but would also appear to be quite justifiably considered as being different from consonant perception.

References

Abramson, A. S. (1961). Identification and discrimination of phonemic tones. *J. acoust. Soc. Amer.*, 33, 842 (Abstract).

Bastian, J., Eimas, P. D., and Liberman, A. M. (1961). Identification and discrimination of a phonemic contrast induced by silent interval. *J. acoust. Soc. Amer.*, 33, 842 (Abstract).

Bastian, J. and Abramson, A. S. (1962). Identification and discrimination of phonemic vowel duration. *J. acoust. Soc. Amer.*, 34, 743 (Abstract).

Broadbent, D. E. and Ladefoged, P. (1960). Vowel judgments and adaptation level. *Proc. Roy. Soc. B.*, 151, 384.

Eimas, P. D. (1962). A study of the relation between absolute identification and discrimination along selected sensory continua. Ph.D. Dissertation (Connecticut).

Fant, G. (1958). Modern instruments and methods for acoustic studies of speech. Proceedings of the VIIIth International Congress of Linguists (Oslo), 282.

Griffith, Belver C. (1957). A study of the relation between phoneme labelling and discriminability in the perception of synthetic stop consonants. Ph.D. Dissertation (Connecticut).

Joos, M. (1948). Acoustic Phonetics. *Lang. Monogr.*, 23.

Ladefoged, P. and Broadbent, D. E. (1957). Information conveyed by vowels. *J. acoust. Soc. Amer.*, 29, 98.

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

Liberman, A. M., Harris, K. S., Kinney, J. A., and Lane, H. (1961a). The discrimination of relative onset-time of the components of certain speech and nonspeech patterns. *J. exptl. Psych.*, 61, 379.

Liberman, A., Harris, K. S., Eimas, P., Lisker, L., and Bastian, J. (1961b). An effect of learning on speech perception : the discrimination of durations of silence with and without phonemic significance. *Language and Speech*, 4, 175.

Peterson, G. E. and Barney, H. L. (1952). Control methods used in a study of the vowels. *J. acoust. Soc. Amer.*, 24, 175.