AN EXPERIMENTAL STUDY OF THE CLASSIFICATION OF SOUNDS IN CONTINUOUS SPEECH ACCORDING TO THEIR DISTRIBUTION IN THE FORMANT 1 – FORMANT 2 PLANE

J. N. SHEARME and J. N. HOLMES

In 1950 Potter and Steinberg[1] published their well-known paper in which they showed that areas on the formant 1–formant 2 plane correspond to vowels spoken in isolated monosyllables with the initial consonant /h/ and the final consonant /d/. This work was followed by the much more detailed experiments of Peterson and Barney,[2] using the same speech material, and the results they published have been widely used by other workers as standard formant data for American English vowels. Potter and Steinberg suggested that in connected speech these vowel areas are only approximated and the present paper is concerned with a further investigation into the movements of formants during normal continuous speech. The most important use of speech is for conversation, and therefore this would be the most interesting type of material to examine. It is, however, difficult to obtain recordings of large amounts of natural conversation, and so continuously read speech was used as the subject for this investigation.

The measurements described in both of these previous papers were made from spectrogram sections, but the labour involved in taking a sufficient number of these measurements from continuous speech would have been so great that the possibility of using an automatic method of formant-frequency measurement was investigated. Automatic formant tracking apparatus was available which had been developed for research in analysis-synthesis telephony, and it appeared to be sufficiently accurate to use for measuring vowel area diagrams (errors were normally less than 50 c/s). The method adopted used the formant-frequency analogue signals from the formant trackers to control the deflection plates of a cathode-ray tube, and the positions of the formants on the plane were sampled by pulses on the intensity modulation electrode. The sampling pulses were arranged to be applied during vowels and vowel-like sounds only, and the resultant display was a distribution of bright spots on the cathode ray tube face. Photographs were taken using a 2-minute exposure while speech recordings were being played into the formant trackers.

[1] Potter, R. K., and Steinberg, J. C., *J. Acous. Soc. Am.*, 22, 807 (1950).
Peterson, G. E. and Barney, H. A., *J. Acous. Soc. Am.*, 24, 175 (1952).

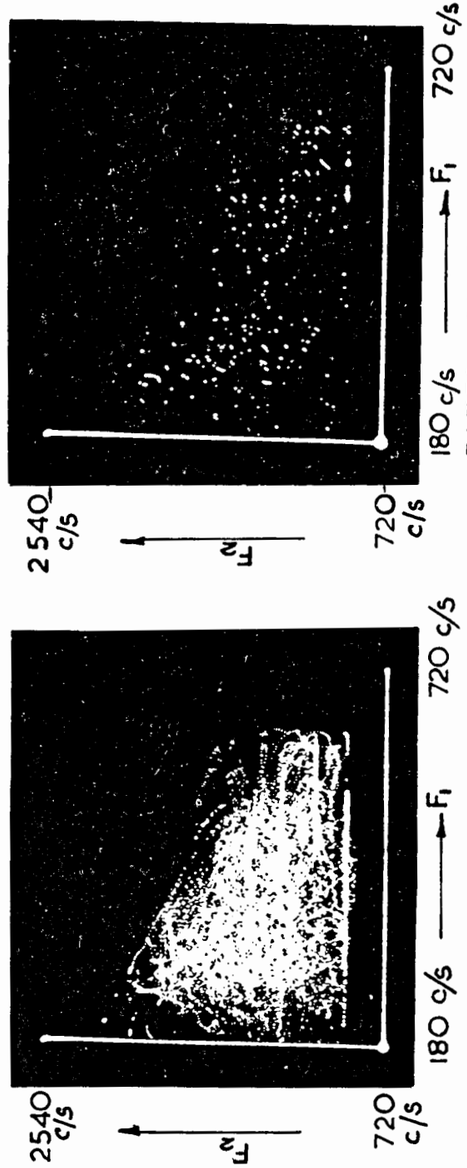

Fig. 1. Distribution of formant 1 and formant 2 during vowels and vowel-like sounds for one speaker.

Because of the presence of vowel-like consonants, transition regions, and diphthongs it was expected that the photographs would show a continuous distribution of formant position on the F_1 F_2 plane, but that there would be obvious concentrations in the regions of the steady vowels, which could be assigned phonemic values. It was found, in practice, that the spots did not appear to cluster into phoneme areas. In an attempt to remove the effects of the rapid formant movements in transitions and diphthongs the F_1 F_2 measurements were repeated with the restriction that points were only recorded when the rate of change of formant frequency was small. The diagrams so obtained still did not reveal clustering, and the distribution of points on the plane was not in fact appreciably altered even when the rate of change was restricted so much that the density of points was reduced to less than one tenth of the density without rate of change restriction. This effect is illustrated in Fig. 1.

It was concluded from these automatic measurements that really steady vowels only occur for a very small proportion of the time in connected speech, and that there was no way of associating the formant tracker outputs with phonemic values. To overcome this last limitation it proved necessary to measure formant positions from spectrograms, and to make phonetic transcriptions of the speech samples which could be correlated with the spectrograms sound by sound. The formant measurements were made directly from wide-band spectrograms, which, for the male speech examined, gave an estimated error never exceeding 50 c/s.

On these spectrograms there was no obvious way to determine precise beginning and end points for each vowel, and therefore to ensure that the entire duration of each vowel was taken into account the complete F_1 F_2 tracks of the vowels and their associated transitions were plotted. All occurrences of one vowel sound by one speaker during 30 seconds of speech were plotted on one diagram, and a set of such diagrams was made to cover several vowels for each of three male speakers. One typical track diagram is shown in Fig. 2.

It is not obvious from such tracks what particular attribute of them is associated with the perceived vowel. One might expect that all tracks of any one vowel would point towards a target area, but if such areas exist it was certainly not obvious how to locate them. It was found, however, that for each vowel a fairly small area could be chosen which contained at least some part of all tracks for that vowel. Such an area is marked on Fig. 2. Areas chosen in this way constitute a set of vowel regions for continuous speech, somewhat similar to the Peterson and Barney vowel areas.

For comparison the three speakers were also asked to produce the vowels in isolated monosyllables, and the formant data for these were plotted on the same diagrams. A typical result is shown by the cross in Fig. 2, which can be seen to be well outside the vowel region as defined above, and is even outside the region of the complete tracks including all transitions. This effect was found to exist generally, and Figs. 3, 4 and 5 show the areas and corresponding monosyllable points for several vowel sounds of each of the three speakers. It can be seen that the vowel areas are considerably displaced from the monosyllable points towards the neutral-

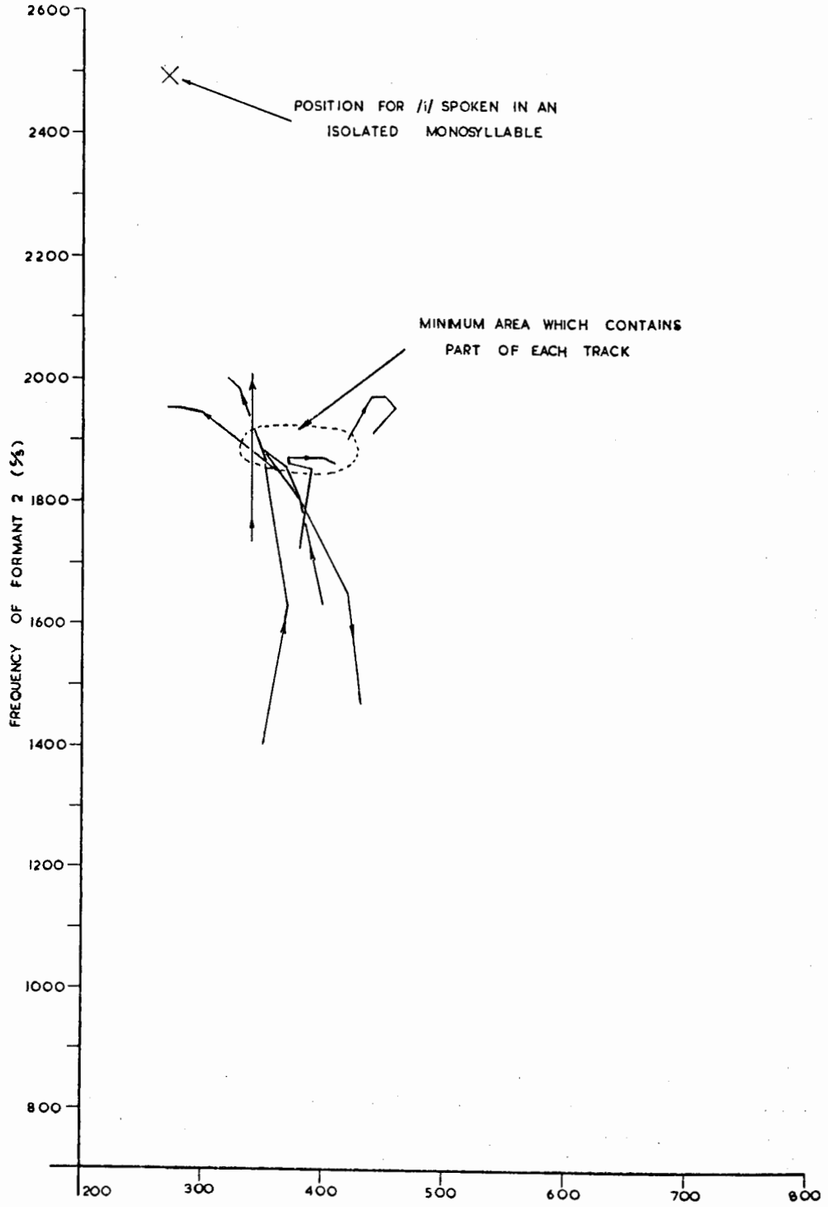

Fig. 2. F_1 F_2 tracks for the sound /i/ of speaker JSG

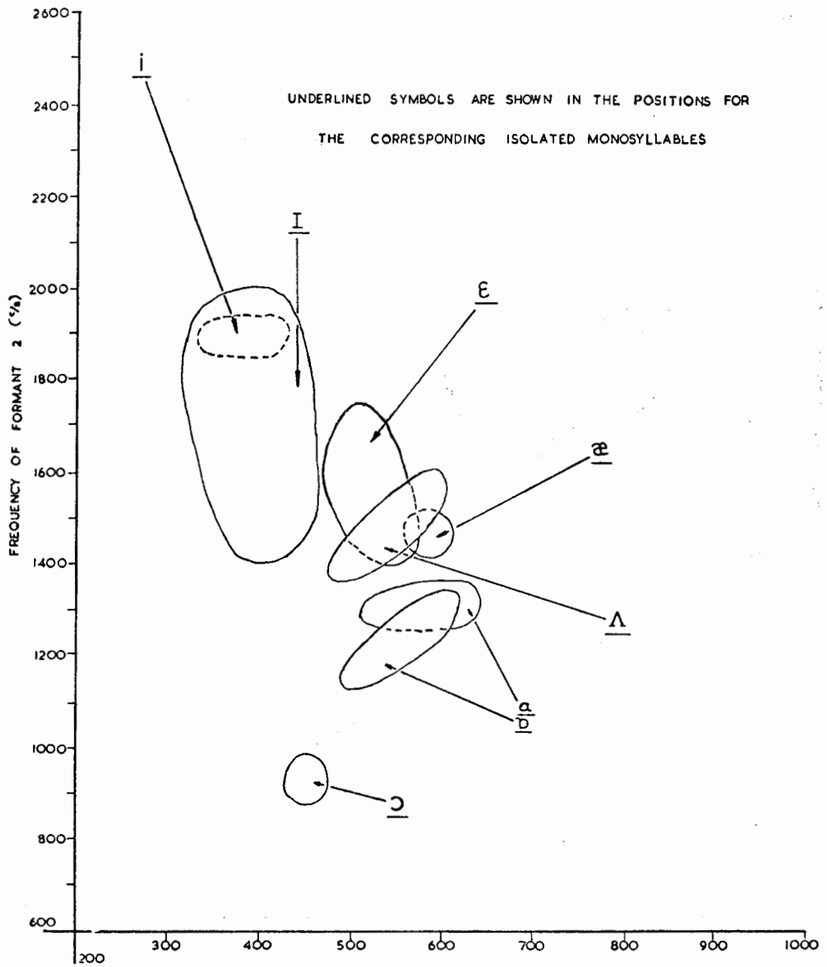

Fig. 3. Vowel areas for speaker JSG

vowel position; the displacements are so large that the deviations of the regions from the neutral position are only about a half of the deviations for isolated monosyllables.

Although the work does not cover a wide range of speech material, the results show quite clearly that the presently accepted formant frequencies of vowels (which

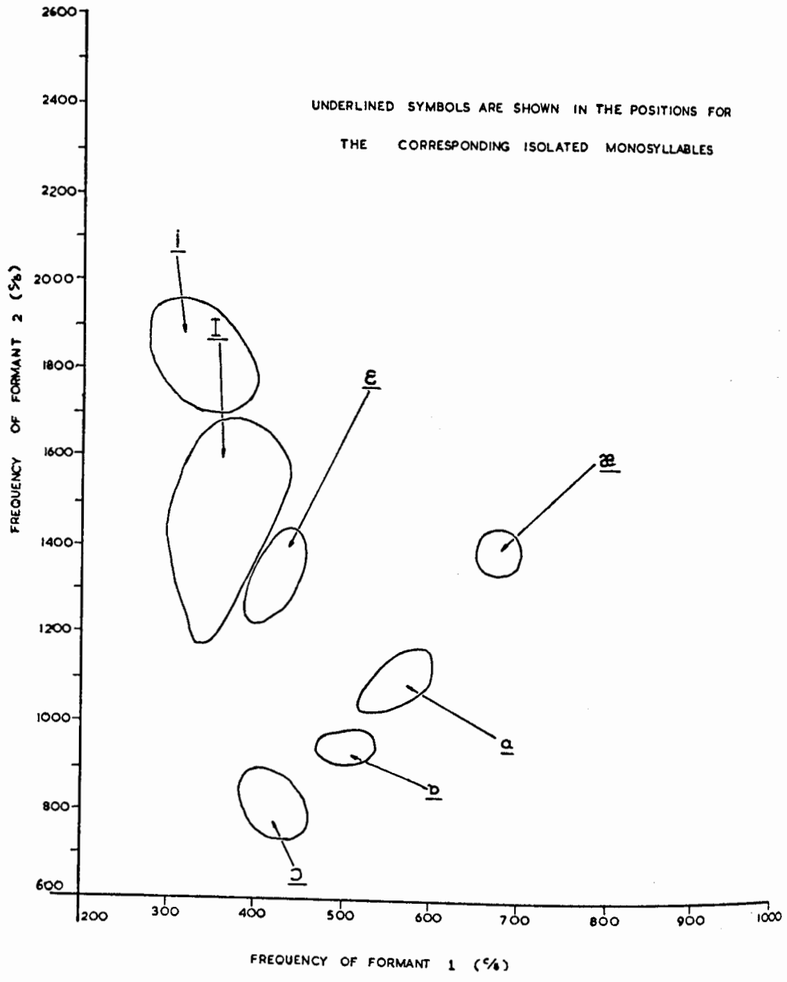

Fig. 4. Vowel areas for speaker PJE

have been measured from isolated monosyllables) are not applicable to connected speech. It is thus apparent that it may be very misleading to use data obtained from carefully spoken non-typical speech material as the basis for design of any apparatus intended for operation on ordinary continuous speech.

The vowel areas obtained in this work overlap to some extent, and even if this

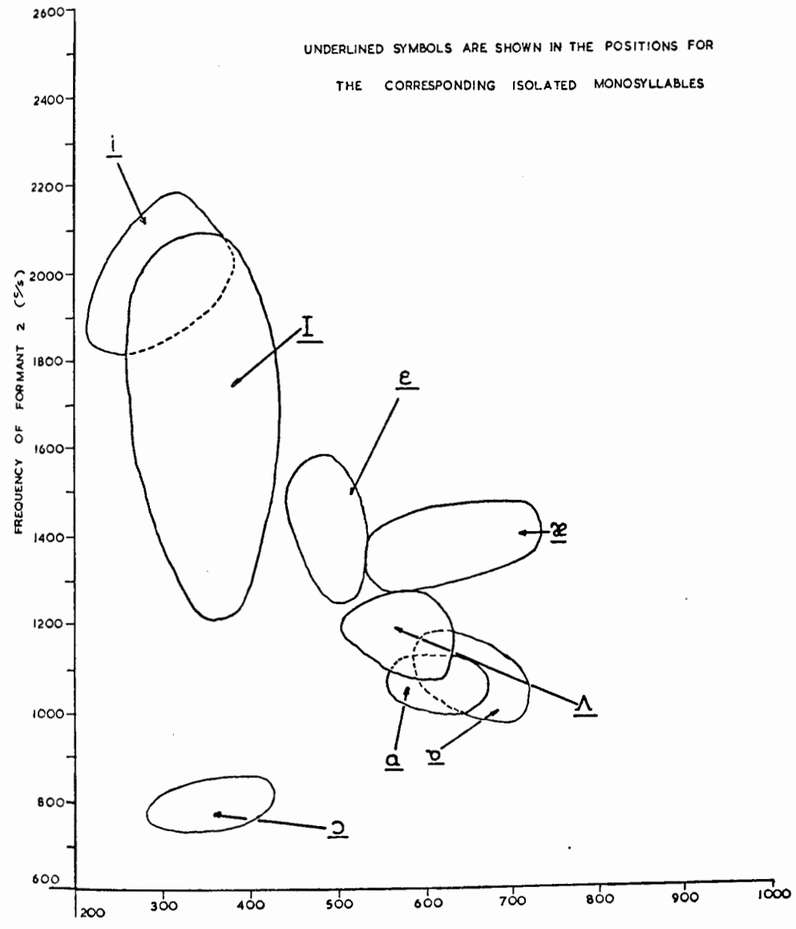

Fig. 5. Vowel areas for speaker LCK

overlap did not exist there would not be a simple correspondence between formant position and phonemic value, because the areas drawn on Figs. 3, 4 and 5 are only occupied for a small proportion of each vowel duration, and also because parts of the tracks of any one vowel frequently pass through the areas of other vowels. It is possible, however, that further investigation of formant positions and movements,

taking into account such additional factors as environment and stress, may eventually lead to a definite relationship between acoustic properties and phonemic values, but it seems certain from these preliminary results that it would not be a simple relationship such as might have been expected from the results of Peterson and Barney.

Joint Speech Research Unit
Ruislip, Middlesex, U.K.