

XV. SPEECH COMMUNICATION*

Prof. K. N. Stevens
Prof. J. B. Dennis
Prof. M. Halle
Dr. A. S. House

Dr. T. T. Sandel
Jane B. Arnold
J. M. Heinz
W. L. Henke

S. Inomata†
A. P. Paul
J. R. Sussex
E. C. Whitman

A. AN AUTOMATIC SPECTRUM MATCHING PROGRAM

The automatic analysis technique that will be described in this report is an extension of a computer analysis procedure that has been reported previously.^{1, 2} This program is the same as that devised for manual matching of speech spectra³ except that the decision functions, i. e., the strategy, are carried out by the program. The strategy used by the automatic program was designed to parallel, as far as possible, that of experimenters who were experienced in the use of the manual program.

In matching vowel spectra, in general, two situations can arise, and a separate strategy has been developed to accommodate each situation. The first, which is the most difficult but the rarest, occurs when no previous data about the spectrum are available. In this case, the strategy attempts to obtain a rough approximation to the formant frequencies (within 200-300 cps). The second situation exists when approximate values or the results of a previous match of a neighboring sample are available to the program.

When the first strategy is used, the approximate locations of the formant frequencies are found by assembling comparison spectra from a small set of stored curves. A given comparison spectrum is constructed from a stored set of first-formant (F1) curves with resonances 100 cps apart and six F2 curves with resonances 200 cps apart. The latter curves include the effect of a fixed set of higher poles. All 36 possible pairs of F1 and F2 combinations are compared with the speech spectrum, and the values yielding the lowest squared-error score are chosen to be the initial approximation. The frequencies of F3 and F4 are assigned values that depend on F₂, and average bandwidths are assigned to each formant. Once such a set of values has been determined, the analysis program will make use of the second strategy.

When the second strategy is in effect, there are eight parameters that can be varied by the automatic program: the frequency and bandwidth of F1, F2, F3, the frequency of F4, and the value of a real-axis zero, corresponding to a "tilt" or glottal-spectrum correction curve. The formant parameters are changed simultaneously at every trial, while the tilt is changed every third trial. The program carries out the following

*This research was supported in part by the U. S. Air Force (Electronic Systems Division) under Contract AF19(604)-6102; in part by the National Science Foundation (Grant G-10800 and Grant G-16526); and in part by the National Institutes of Health (Grant MH-04737-02).

†On leave from the Electrotechnical Laboratory, Ministry of International Trade and Industry, Tokyo, Japan.

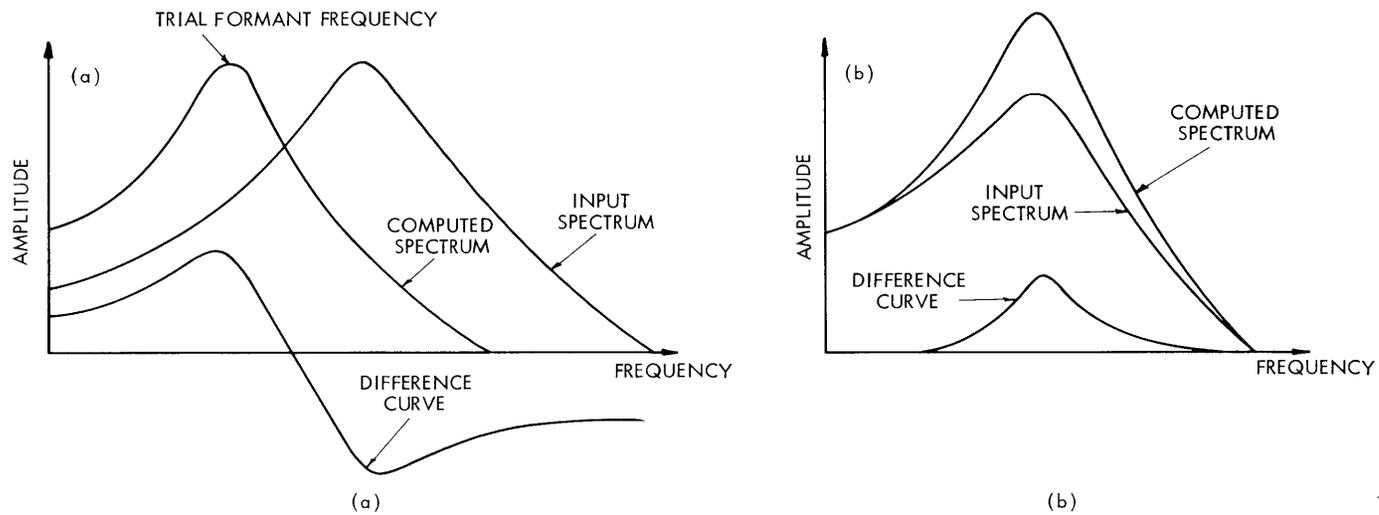


Fig. XV-1. Comparison of input spectrum and internally generated spectrum for the case in which there is (a) an error in formant frequency, and (b) an error in formant bandwidth.

operations: (a) compute the frequency increment, δF_i , and the bandwidth increment, δb_i , for the i^{th} formant; (b) constrain the increments to prevent unreasonable behavior; (c) tilt; (d) generate trial spectrum; (e) repeat steps 1-5 until a best fit has been achieved.

The basic part of the program is the evaluation of the two functions for computing the frequency and bandwidth increments. The frequency-increment function is based on an examination of the error curve in the vicinity of the formant frequency that characterizes the trial spectrum. Idealized curves corresponding to a frequency error and a bandwidth error are shown in Fig. XV-1. If we let SL = weighted sum of difference-curve values for a specified frequency range to the left of the trial formant frequency, and SR = the same weighted sum to the right of the trial formant frequency, then the frequency increment is given by

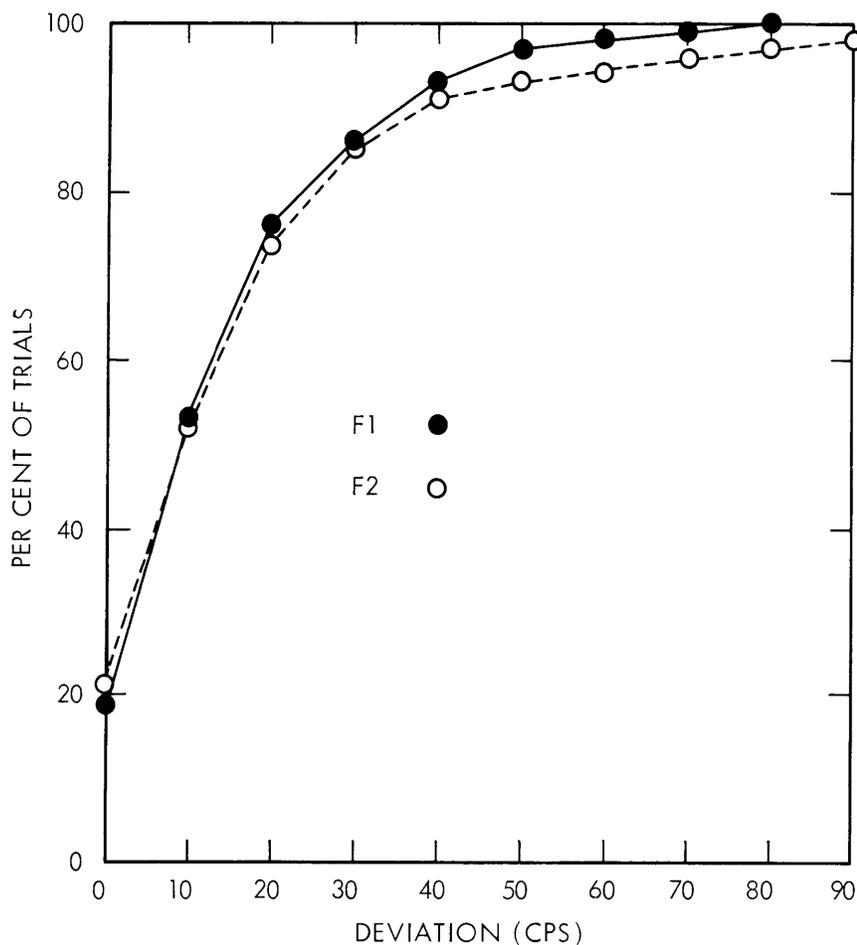


Fig. XV-2. Cumulative distribution curve of difference between formant frequencies obtained by automatic and manual analysis procedures.

(XV. SPEECH COMMUNICATION)

$$\delta f_i = \frac{SL - SR}{K} \cdot bw_j,$$

where K is an empirically derived constant that is the same for all formants, and bw_j is the bandwidth of the analyzing filter nearest the trial formant frequency. The bandwidth increment is obtained from a similar function, except that a table look-up is used to specify values of a function $\delta\beta_i = f(SL+SR)$. Constraints are placed on the δf_i so that each δf_i tends to zero with successive trials. The best fit is for the case in which the squared-error score has a minimal value; recognition of the best match is simplified by placing a limit on the number of trials.

The new automatic program (MATCH IV) has been used to specify the formant frequencies for three adjacent spectral samples taken from each of a large number (360) of vowels spoken by three male talkers. These data were compared with data on the same spectral samples obtained as part of an earlier study⁴ by the manual matching procedure. Some idea of the performance of the automatic program as compared with the manual matching procedure is given in Fig. XV-2. This figure shows, for example, that 75 per cent of the time the values of F_1 and F_2 as obtained by the automatic methods are within ± 20 cps of those obtained by the manual procedure. Results such as these suggest strongly that the automatic procedure leads to valid data on formant locations.

A. P. Paul

References

1. A. P. Paul, Strategy for the automatic matching of speech spectra, S.M. Thesis, Department of Electrical Engineering, M.I.T., August 1961.
2. A. P. Paul, Automatic matching of vowel spectra, Quarterly Progress Report No. 63, Research Laboratory of Electronics, M.I.T., October 15, 1961, p. 137.
3. C. G. Bell, H. Fujisaki, J. M. Heinz, K. N. Stevens, and A. S. House, Reduction of speech spectra by analysis-by-synthesis techniques, J. Acoust. Soc. Am. 33, 1725-1736 (1961).
4. K. N. Stevens, H. Fujisaki, and A. S. House, Analysis of vowel spectra, Quarterly Progress Report No. 60, Research Laboratory of Electronics, M.I.T., January 15, 1961, pp. 177-179.

B. DATA DERIVED BY AN AUTOMATIC SPECTRUM MATCHING PROGRAM

The automatic analysis program described in Section XV-A has been used to extract formant frequencies throughout the vowel portions of a number of utterances. The speech materials were drawn from a library of utterances that have been described previously.¹ The measurements were made in eight stressed vowels occurring in 15 different symmetrical non-nasal consonant environments. The words were spoken by three male talkers. The computer was instructed to determine the formant frequencies for all

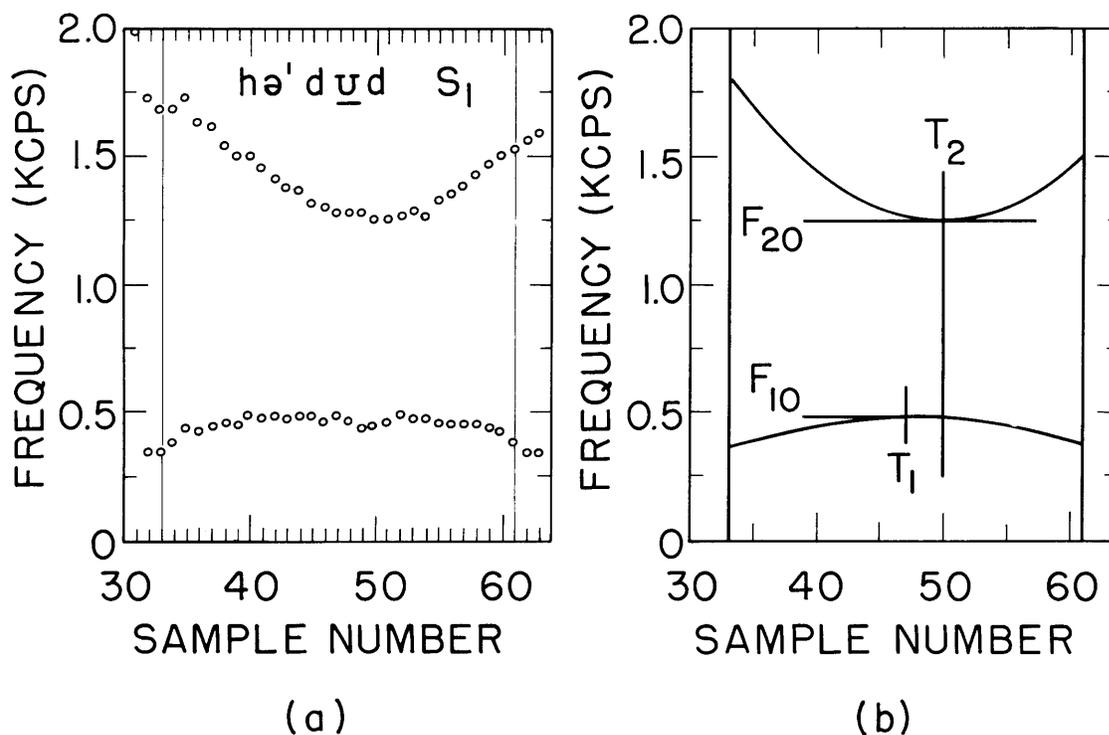


Fig. XV-3. (a) Example of data on the frequencies of first two formants obtained by the automatic analysis program. Spectral samples occurring 120 per second. The light lines at samples 33 and 61 represent vowel boundaries obtained from visual examination of spectrograms. (b) Approximations to the data in (a) obtained by fitting parabolic curves to the plotted points within the vowel boundaries. The two curves are completely described by the parameters F_{10} , F_{20} , T_1 , and T_2 , together with the curvature values for the parabolas.

spectral samples within the vowel boundaries. The time locations of these boundaries had been established earlier from visual examination of spectrograms.²

Typical results of the automatic analysis, showing the frequencies of the first two formants for one utterance, are given in Fig. XV-3a. The curves depicting F_1 and F_2 as a function of time for the symmetrical consonant environments used in this study usually have a form similar to that shown in this figure. Each formant begins at some value that is presumably dependent upon the vocal-tract configuration for the initial consonant. The curve then proceeds toward a maximum or minimum value associated with some target configuration, and finally tends toward a value determined by the vocal-tract configuration for the final consonant.

Examination of data from a number of vowels has suggested that the plotted points for each formant of each vowel might be approximated by two parabolic segments, one for the initial portion of the vowel and the other for the final portion. Each segment is of the form

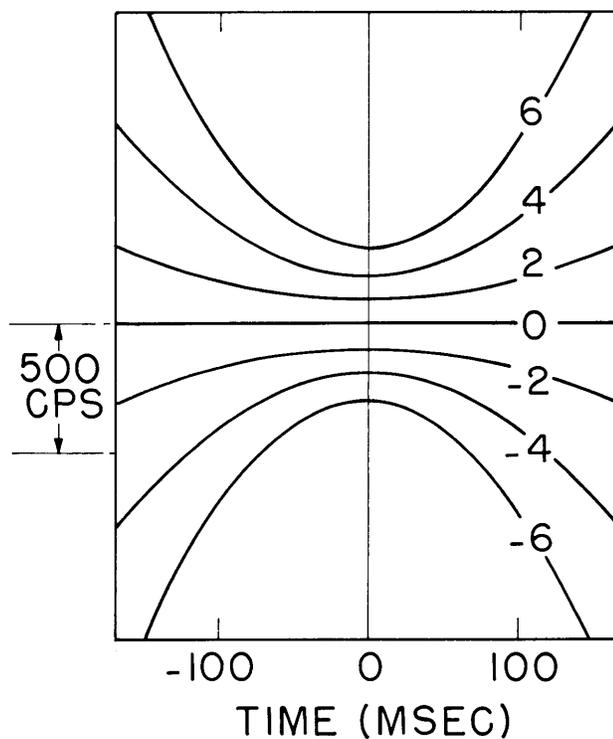


Fig. XV-4. Example of parabolic curves used to specify vowel formant data. The parameter is the "curvature value."

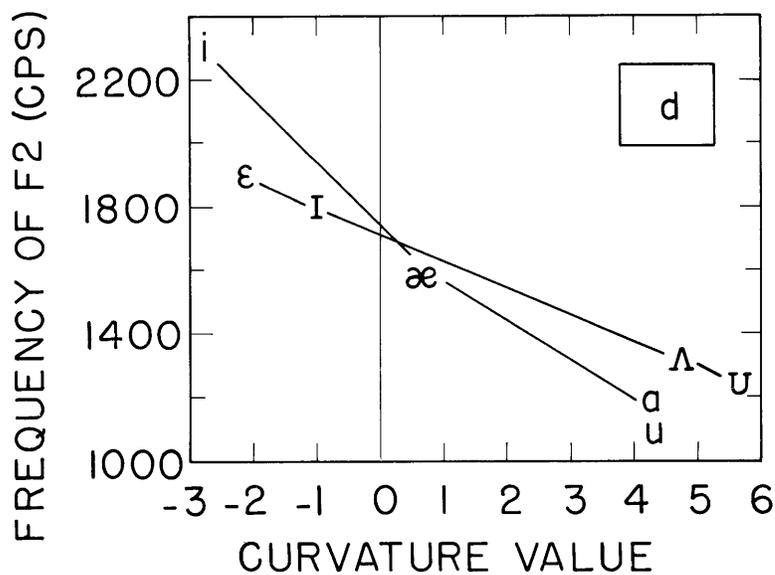


Fig. XV-5. Frequency F_{20} vs curvature value for various vowels preceded and followed by the consonant /d/. Each point represents average data for three talkers.

$$F_i = a_i(t-T_i)^2 + F_{i0},$$

where F_i represents formant frequency, t is time, and a_i , T_i and F_{i0} are parameters that are selected to give a best fit; ($i = 1$ for the first formant and $i = 2$ for the second formant). The parameter F_{i0} represents the maximum or minimum value reached by the formant, and T_i indicates the point in time at which this maximum or minimum value occurs. The parameter a_i is a measure of the curvature of the formant, and can be positive or negative. Figure XV-3b shows the parabolic curves that give good fits to the data points plotted in Fig. XV-3a. The parabolic curves used in Fig. XV-3b were selected from a set of curves similar to those shown in Fig. XV-4. The parameter values on these curves will be called arbitrarily "curvature values."

Curvature values and frequencies of F1 and F2 can be related either by examining data from various vowels in a given consonantal context or by examining data from a given vowel in various contexts. These two ways of organizing the data are illustrated in Figs. XV-5 and XV-6.

In Fig. XV-5 the maximum or minimum frequency of F2 is plotted against the curvature value of F2 for each of eight stressed vowels preceded and followed by /d/. There are systematic changes in curvature values from negative to positive as the frequency of F2 decreases from high to low values, and this change differs slightly for the long and short vowels. The point of intersection of the two functions with the point of zero curvature value is approximately 1700 cps; when the frequency of F2 for a vowel is less than 1700 cps its curvature value is positive, and conversely it is negative when F_2 is greater than 1700 cps. In other words, the value 1700 cps describes the articulatory

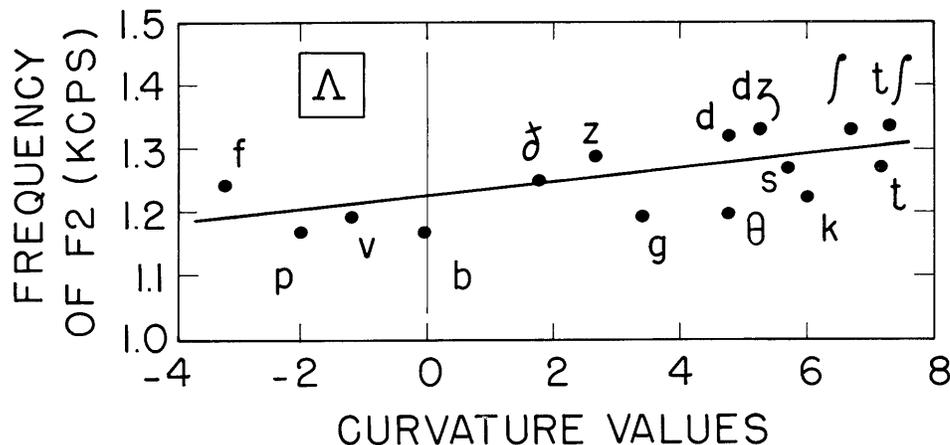


Fig. XV-6. Frequency F_{20} vs curvature values for one vowel in various symmetrical consonant contexts as indicated. Each point represents average data for three talkers. The straight line was fitted visually to the data.

(XV. SPEECH COMMUNICATION)

context provided by the postdental stop consonant /d/, that is, the so-called F2 locus of /d/ is at approximately 1700 cps.^{3, 4}

In Fig. XV-6 the maximum or minimum frequency of F2 for a given vowel is plotted against curvature values associated with various consonantal contexts. The data display a wide variation in curvature value, and there is a small but apparently systematic increase in F2 as the curvature values become more positive. If the straight line can be interpreted as representing a best fit to the plotted points, then the frequency 1220 cps can be interpreted as the F2 value associated with the unperturbed vowel articulation, that is, 1220 cps is the so-called F2 target value of this vowel. When the curvature value is negative, the maximum value of F₂ is less than the target value; for positive curvature values, the minimum value of F₂ exceeds target value. In articulatory terms, this interpretation suggests that there is an effective inertia in the system that gives rise to undershoot in the articulatory processes.

K. N. Stevens, A. S. House, A. P. Paul, Jane B. Arnold

References

1. K. N. Stevens, H. Fujisaki, and A. S. House, Analysis of vowel spectra, Quarterly Progress Report No. 60, Research Laboratory of Electronics, M.I.T., January 15, 1962, pp. 177-179.
2. A. S. House, On vowel duration in English, J. Acoust. Soc. Am. 33, 1174-1178 (1961).
3. P. Delattre, A. M. Liberman, and F. S. Cooper, Acoustic loci and transitional cues for consonants, J. Acoust. Soc. Am. 27, 769-773 (1955).
4. K. N. Stevens and A. S. House, Studies of formant transitions using a vocal-tract analog, J. Acoust. Soc. Am. 28, 578-585 (1956).

C. ERRATA

In the report entitled "Reduction of Speech Spectra to Descriptions in Terms of Vocal-Tract Area Functions" by J. M. Heinz, published in Quarterly Progress Report No. 64 (pages 198-203), the numbers referred to in the legend of Fig. XXII-7 were omitted from the figure. They read as follows:

Fig. XXII-7a	244	(frequency)
	-1	(interval size = 2 ⁻¹ cm)
	32	(number of interval)
Fig. XXII-7b	852	
	-1	
	32	
Fig. XXII-7c	2192	
	-1	
	32.	

J. M. Heinz