Reprinted from THE JOURNAL OF THE ACOUSTICAL SOCIETY OF AMERICA, Vol. 33, No. 12, 1725–1736, December, 1961

Reduction of Speech Spectra by Analysis-by-Synthesis Techniques

C. G. BELL,* H. FUJISAKI,† J. M. HEINZ, K. N. STEVENS, AND A. S. HOUSE
Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, Massachusetts
(Received September 6, 1961)

Procedures are described for reducing the speech wave to a specification in terms of the time-varying vocal-tract resonances and source characteristics. The basic method, which has been called analysis by synthesis, involves the comparison of speech spectra with a series of spectra that are synthesized within the analyzer. Each comparison spectrum is generated according to a set of rules based on an acoustical theory of speech production. The result of the analysis of each input spectrum is a set of parameters that describes the synthesized spectrum providing the best match. In one version of the method convergence, towards the best match is controlled by the experimenter; in another version convergence to a match is accomplished automatically without the intervention of the experimenter. All the operations have been programmed on a general-purpose digital computer and have been applied to the analysis of vowels and some consonants. The advantages of the analysis techniques are discussed.

THE problem of representing speech events in terms of low-information-rate signals that describe the essential features of the speech wave is one of the central problems in the area of speech communication. To the student of phonemics and phonetics it is important to be able to describe in a simple way the acoustical features associated with the various allophones of the phonemes. For the engineer concerned with problems of communication, an efficient description of speech signals is needed for the development of systems for speech bandwidth compression and for the realization of procedures for machine recognition and generation of speech.

The development of the sound spectrograph[1] represented a significant contribution to speech analysis, since it displays the speech events in a way that brings into clear focus certain of the essential features of the signal such as the formant movements. The three-dimensional intensity-frequency-time representation, together with the procedure for displaying spectral sections, provides a means for isolating significant features for certain classes of sounds, although the techniques are less successful for other classes, particularly certain types of consonants.

During the past few years, there have been two developments which suggest that it is now possible to bring more powerful techniques to bear on problems of speech analysis. The first has been the significant advance that has occurred in our understanding of the acoustics of speech production. The second has been the increasing availability of high-speed digital computers for applications such as speech analysis. Theoretical studies have led to a clearer understanding of the constraints imposed on the speech signal by the vocal mechanism, and have suggested means whereby speech signals can be represented in terms of parameters that have a definite and rigorous relation to articulation. Digital computers have made it possible to use the results of the acoustical studies in such a way that rapid and precise reduction of speech signals can be accomplished.

This paper describes an attempt to utilize the findings of the acoustical theory to develop procedures for the analysis and reduction of speech signals by computer techniques. Since the method is based on an acoustical theory of speech production, it is appropriate to outline the essential features of such a theory before proceeding to a description of the analysis techniques.

ACOUSTICAL THEORY OF SPEECH PRODUCTION

The generally accepted theory of speech production[2–4] views the speech wave as the result of acoustic excitation of the vocal tract by one or more sources. In the case of voiced sounds, there is a source at the glottis, and this glottal source is a quasi-periodic volume-velocity wave whose spectrum envelope decreases with increasing frequency at a rate of about 12 db/octave in the range 300–2500 cps. The characteristics of the glottal source are to a large extent independent of the vocal-tract configuration anterior to the glottis. For some classes of sounds there may be a source of excitation of the vocal tract as a result of a sudden pressure release or as a result of turbulent air flow through a constriction or past the teeth or other obstructions. Such a source can be considered as a differential-pressure source, usually located in the vicinity of a vocal-tract constriction, and this source generally has a relatively broad and smooth spectrum. The spectrum $P(s)$ of the sound pressure measured at a distance from the lips as a result of a source of excitation whose spectrum is given by $S(s)$ can be written

$$P(s)=S(s)\quad T(s)\quad R(s). \tag{1}$$

In this equation $T(s)$ is the transfer function of the vocal tract; for voiced sounds, $T(s)$ is the ratio of the volume

* Present address: Digital Equipment Corporation, Maynard, Massachusetts.
† Present address: University of Tokyo, Tokyo, Japan.
[1] W. Koenig, H. K. Dunn, and L. Y. Lacy, J. Acoust. Soc. Am. 17, 19 (1946).
[2] H. Dudley, Bell System Tech. J. 19, 495 (1940).
[3] T. Chiba and M. Kajiyama, *The Vowel, Its Nature and Structure* (Tokyo-Kaiseikan Publishing Company, Ltd., Tokyo, 1941).
[4] G. Fant, *Acoustic Theory of Speech Production* (Mouton and Company, 's-Gravenhage, 1960).

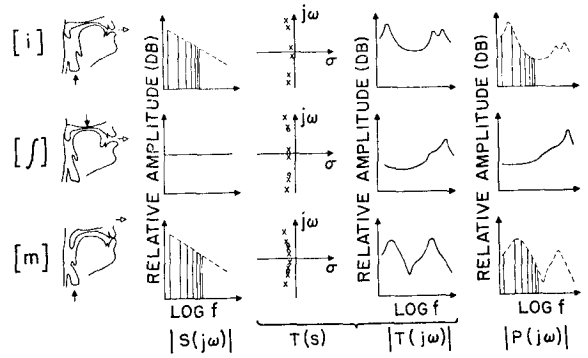

FIG. 1. A nonquantitative representation of the production and acoustic characteristics of speech sounds. In the left column are sketches of articulatory configurations in the midsagittal plane for three phones as indicated, together with source (solid arrows) and output (open arrows) locations. In addition each row of the figure shows the magnitude of the source spectrum $|S(j\omega)|$, the pole and zero locations in the complex frequency plane for the vocal-tract transfer function $T(s)$, the magnitude of the transfer function $|T(j\omega)|$, and the magnitude of the sound pressure at a distance in front of the face $|P(j\omega)|$, respectively, for each representative articulation. In cases with periodic (glottal) excitation, incomplete sets of harmonics are shown together with spectrum-envelope curves.

velocity at the mouth opening (and at the nostrils if there is coupling to the nasal cavities) to the source volume velocity, whereas for a noiselike or transient source at a constriction, $T(s)$ is the ratio of the volume velocity at the mouth opening to the sound pressure of the source. The radiation characteristic $R(s)$ is the ratio of the sound pressure at distance r in front of the talker to the volume velocity at the lips, and in the frequency range up to about 4000 cps is given approximately by the result for a simple source

$$R(s)=(s\rho/4\pi r)e^{-(sr/c)}, \tag{2}$$

where ρ=density of air and c=velocity of sound in air. In all of these relations, s is the complex frequency, and can be replaced by $j\omega$ to obtain Fourier spectra, where ω is the angular frequency.

When the source is at the glottis and when there is no coupling to the nasal cavities, $T(s)$ is characterized by a number of poles, and can be written

$$T(s)=\frac{s_1s_1^*s_2s_2^*\cdots}{(s-s_1)(s-s_1^*)(s-s_2)(s-s_2^*)\cdots}, \tag{3}$$

where the asterisks designate complex conjugates and s_1, s_2, $\cdots$ are the poles corresponding to the various vocal-tract resonances or formants. The frequencies and bandwidths for the poles are, of course, dependent on the vocal-tract configuration. For an idealized source spectrum envelope with a decreasing slope of 12 db/octave, i.e., a spectrum envelope proportional to $1/\omega^2$, and for a radiation characteristic proportional to ω [cf. Eq. (2)], the magnitude of the spectrum envelope $|P_E(j\omega)|$ of the sound pressure for a nonnasal vowel is given by

$$|P_E(j\omega)|\propto(1/\omega)|T(j\omega)|. \tag{4}$$

Thus if the idealized source spectrum is assumed, the envelope $|P_E(j\omega)|$ is characterized by a pole in the vicinity of $\omega=0$ and by a set of conjugate-complex pairs of poles, corresponding to the poles of $T(s)$ in Eq. (3). Alternatively, if the sound pressure is transduced and passed through a circuit with a frequency characteristic that rises at 6 db/octave, the spectrum envelope of the resulting signal is characterized by the poles of $T(s)$ and only those poles, assuming the idealized shape for the source spectrum envelope.

When there is coupling between the vocal tract and the nasal tract, or when the vocal-tract excitation is at a point other than the glottis, the transfer function $T(s)$ is characterized by zeros as well as poles, and can, in general, be written

$$T(s)=K\frac{(s-s_a)(s-s_a^*)(s-s_b)(s-s_b^*)\cdots}{(s-s_1)(s-s_1^*)(s-s_2)(s-s_2^*)\cdots}, \tag{5}$$

where s_a, s_a^*, s_b, s_b^*, $\cdots$ are the zeros and K is a real quantity independent of frequency. The frequencies and bandwidths for the *zeros* depend both on the vocal-tract configuration and on the location of the source in the vocal tract, whereas for a given vocal-tract configuration the *poles* of $T(s)$ are independent of the location of the source.

Relations between the speech spectra and the articulatory processes that produce them are summarized by the sketches in Fig. 1 for three classes of speech sounds. For each class of sounds, the figure shows a typical articulatory configuration and source location, and approximate source spectrum, a representation of the poles and zeros of the transfer function, a plot of the magnitude of the transfer function vs frequency, and the output spectrum. The spectra for the vowel and for the nasal consonant are, of course, line spectra, whereas the fricative has a continuous spectrum. In each case, the output spectrum (in decibels) is obtained by adding the spectra of the source and the transfer function, and then applying a 6 db/octave correction for the radiation characteristic, as explained above.

The acoustical theory may be summarized, therefore, as follows. The spectrum of the vocal-tract output (in decibels) is the sum of a source spectrum, a transfer function and a radiation characteristic. For a given class of speech sounds, the source spectrum and the radiation characteristic are relatively invariant from one talker to another, and are largely independent of the articulatory configuration. The transfer function is determined by the articulatory configuration and the source location, and is completely described in terms of a set of poles in the case of nonnasal vowel and vowel-like sounds, and by a set of poles and zeros for other classes of sounds.

ANALYSIS-BY-SYNTHESIS MODEL

The term *analysis by synthesis* is used to refer to an active analysis process that can be applied to signals that are produced by a generator whose properties are known.[5,6] The heart of an analysis-by-synthesis system is a signal generator capable of synthesizing all and only the signals to be analyzed. The signals synthesized by the generator are compared with the signals to be analyzed, and a measure of error is computed. Different signals are generated until one is found that causes the error to reach some smallest value, at which time the analyzer indicates the properties of the internally generated signal. It has been suggested[6,7] that a scheme of this type has applications in the analysis of linguistic phenomena at various levels of representation: acoustic, graphic, phonological, morphological, and syntactic. Of concern in the present discussion is the analysis of linguistic events at the acoustic level.

The procedure used to accomplish analysis by synthesis at the acoustic level[7] is shown schematically in Fig. 2. The speech is passed first through a peripheral element in this case a *filter set*, the outputs of which are recitfied, smoothed, sampled at prescribed time intervals, and then stored. (The techniques used to process the speech by the filter system and to store the spectra in the computer memory are described in the Appendix.) The component labeled *spectrum generator*, when given appropriate instructions, can generate outputs that are compatible with the original stored speech data. In the present case, this component generates speechlike spectra when provided with information on the poles and zeros of the vocal-tract transfer function and on the type of vocal-tract excitation. The *comparator* computes a measure of the difference between the input speech spectra and spectra generated by the model. The order in which different trial spectra are synthesized by the model is prescribed by a control or *strategy* component that makes decisions on the basis of (1) previous error scores for the spectral sample under analysis, (2) the results of analyses of adjacent spectral samples, and (3) possibly the results of preliminary direct measure-

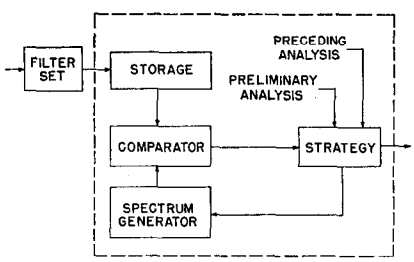

FIG. 2. Paradigm of an analysis-by-synthesis process for the reduction of speech spectra.

[5] D. M. Mackay, Brit. J. Philo. Sci. 2, 105 (1951).
[6] M. Halle and K. N. Stevens, "Analysis by synthesis," Proc. Sem. Speech Compression and Processing, edited by W. Wathen-Dunn and L. E. Woods, AFCRC-TR-59-198, December 1959, Vol. II, Paper D7.
[7] K. N. Stevens, J. Acoust. Soc. Am. 32, 47 (1960).

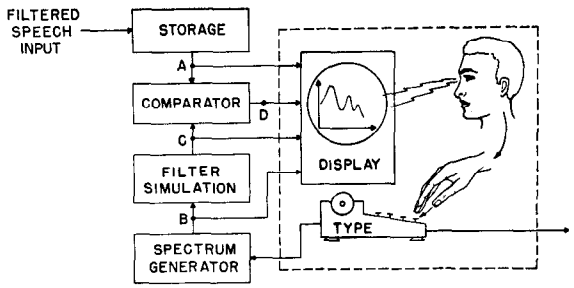

FIG. 3. Paradigm of the mode of the analysis-by-synthesis reduction scheme in which the experimenter controls the selection of parameters of the comparison spectra. The strategy or control element of Fig. 2 is realized by the contents of the dashed box at the right. The nodes A-D show points where functions can be selected and displayed on the cathode-ray tube output. The process is described fully in the text.

ments on the spectral sample. When a synthesized spectrum that provides minimum error is obtained, the analyzer indicates (or stores) the pole-zero locations and source characteristics of that spectrum.

Five operations, therefore, are performed in the analyzer: (a) storage of the speech data processed by the input filter set, (b) generation of speech spectra, (c) instruction of the spectrum generator by a control system, (d) calculation of measures of the difference between the input speech spectra and the spectra computed internally, and (e) display, in some form, of the parameters of the generator that yield minimum error.

The success and utility of the analysis-by-synthesis technique in comparison with other analysis methods depends largely upon the speed and accuracy with which speech spectra can be analyzed, and it is important, therefore, that the number of trial spectra that need to be synthesized in order to obtain a minimum error be kept as small as possible. Thus one of the central problems in the design of an analysis-by-synthesis scheme is that of devising a strategy to be used by the control component to assure rapid convergence to the desired result.

In the analysis procedures described here, two different methods have been used to implement the operations in the strategy component of Fig. 2. In one case, the control function is performed by the experimenter, and hence the problem of specifying a strategy for automatic analysis is circumvented. In the other case a rudimentary strategy that permits automatic analysis of speech spectra is employed. The former method is slower than the automatic procedure, but leads to greater accuracy of analysis, and can be used in the development of strategies that might ultimately be incorporated into a more sophisticated automatic procedure. The two analysis methods will be discussed in detail in the following sections.

EXPERIMENTAL MATCHING OF VOWEL AND CONSONANT SPECTRA

When control of the internal spectrum generator is placed in the hands of the experimenter, the analysis-by-

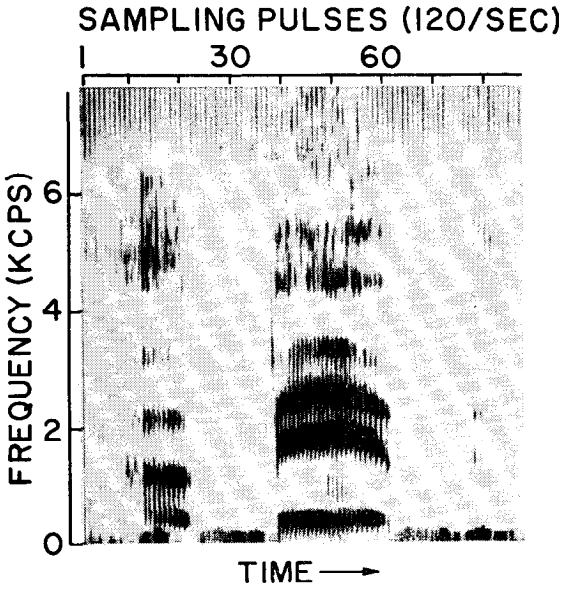

FIG. 4. Sound spectrogram of a nonsense utterance, [hɔ'bɪb]. The time sampling pulses (high-pass filtered) appear at the top of the spectrogram.

synthesis procedure takes the form shown in Fig. 3. In this figure all operations indicated in the blocks, except the decisions and actions of the operator, are performed within the digital computer or its peripheral equipment. Instructions are communicated to the spectrum generator through a typewriter that is operated by the experimenter. The locations (in the complex frequency plane) of a trial set of poles and zeros are typed, and the spectrum generator and filter simulation portions of the system compute a spectrum to be compared with the input speech spectrum that is under analysis. Measured or computed functions existing at various points in the analysis process can be displayed on a cathode-ray tube, as shown in the figure. The task of the operator is to adjust the positions of the poles and zeros until a "best fit" is obtained between the spectrum under analysis and the internally generated spectrum. The experimenter can use both a set of numerical error scores and visual examination of the displayed functions to determine how to adjust the set of poles and zeros in order to improve the match and to decide when a best fit has been obtained.

As indicated in the Appendix, the speech spectra in the present analysis scheme are obtained by passing the speech signal through a pre-emphasis circuit with a rising characteristic of 6 db/octave and then through a bank of 36 filters. Each filter output is rectified and smoothed by a low-pass filter with a time constant of about 10 msec. The recitfied and smoothed outputs are sampled periodically at intervals of 8.3 msec, processed by an analog-to-digital converter, converted to logarithmic values (to the nearest decibel), and stored in the computer memory. The spectral data are also stored on punched tape and are thus available for future analysis.

In order to facilitate analysis of specific components of the utterances, conventional sound spectrograms of the speech materials are made. A spectrogram of a typical dissyllabic utterance used in one experimental study is shown in Fig. 4. Time pulses indicating the instants at which the filter outputs are sampled by the computer are high-pass filtered and mixed with the speech signal before the spectrograms are made, and appear as closely spaced vertical lines across the top of the spectrogram. The numerical identification of the pulses shown on the spectrogram corresponds to the way successive spectra are labeled in the computer memory. Thus the experimenter can, if he wishes, use the spectrogram as a guide in the selection of a particular spectrum or group of spectra from the computer memory. The selection is achieved by a simple instruction to the computer identifying the spectrum to be displayed and analyzed.

The manner in which the speech spectra are displayed is shown by the example in Fig. 5, which was taken from the utterance whose spectrogram is given in Fig. 4. The number at the upper right of the display indicates that the spectrum number is 48, and reference to the spectrogram shows that the spectrum occurred during the stressed vowel [ɪ]. The points along the abscissa represent successive filters in the analyzing bank, and the ordinate is the amplitude in decibels. In terms of the processes portrayed by Fig. 3, this spectrum is found at node A.

As described above, the experimenter specifies the locations of a set of poles and zeros and the internal spectrum generator computes the corresponding spectrum that is to be compared with a speech spectrum such as that shown in Fig. 5. Computation of the synthesized spectrum is carried out in two steps. The first step is to calculate the logarithm of the magnitude of the transfer function as a function of frequency and the second step is to compute the effective spectrum that would be measured if a signal with a spectrum corresponding to this transfer function were processed by the analyzing filter bank. The second step is necessary since the original speech signal itself is, of course, processed by a bank of relatively broad filters, and thus a

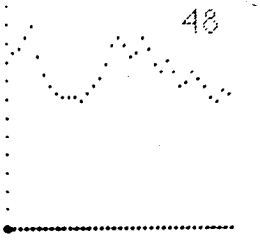

FIG. 5. Photograph of a spectral representation of an input speech sample (at node A in Fig. 3) displayed on the face of the output cathode-ray tube. Each point on the horizontal axis represents one of the 36 filter outputs; the points on the vertical axis represent 5-db steps in amplitude. The input speech has received a 6 db/octave pre-emphasis. The number 48 identifies a sample in the vowel [ɪ] in the word shown in Fig. 4.

valid comparison can be made only if both the inputsignal and the internally generated signal undergo the same sequence of operations.

For a single conjugate pair of poles at frequency F_n with bandwidth ΔF_n the function to be computed in the internal spectrum generator during the first step is[8]

$$20 \log |T_n| = 20 \log \left\{ \frac{F_n^2 + (\frac{1}{2}\Delta F_n)^2}{[(f-F_n)^2 + (\frac{1}{2}\Delta F_n)^2]^{\frac{1}{2}}[(f+F_n)^2 + (\frac{1}{2}\Delta F_n)^2]^{\frac{1}{2}}} \right\}, \quad (10)$$

where f is the frequency variable and T_n is the portion of the system function associated with the nth pair of poles. The logarithm of the magnitude of the system function for a zero is, of course, the negative of that for a pole with the same frequency and bandwidth. For a real-axis pole at a frequency minus F_a, the function to be computed is

$$20 \log\{F_a/(f^2+F_a^2)^{\frac{1}{2}}\}. \quad (11)$$

These functions are computed to the nearest $\frac{1}{8}$ db at 100-cps intervals of f. When more than one pole or zero is specified, the logarithm of the system function is obtained by adding the logarithms of the system functions corresponding to the individual poles and zeros. A result of this calculation for a typical set of poles appropriate for a vowel is shown in Fig. 6, which is a photograph of the cathode-ray tube display obtained from node B in Fig. 3. Although values up to 10 kc are displayed here, usually only the spectrum up to about 3000 cps is of interest for vowels. In this type of display a logarithmic scale is used for the abscissa.

The second step in the computation of comparison spectra is the evaluation of the effect of the filter bank on the computed transfer function. If the magnitude of the transfer function of a filter is designated as $|A_i(f)|$, where i represents the filter number from 1 to 36, then the magnitude of the square-law rectified and smoothed output of the filter when the input spectrum has a magnitude $|H(f)|$ is proportional to

$$\left[\int_0^\infty |H(f)|^2 |A_i(f)|^2 df\right]^{\frac{1}{2}}. \quad (12)$$

The difference between the result of this computation and the result of processing the data by full-wave recitfication can be expected to be less than 1 db. Thirty-six such integrals are evaluated in the computer to obtain the hypothetical rectified filter outputs corresponding to a given input spectrum. These numbers are expressed in decibels to permit direct comparison with the spectra that are under analysis.

The result of the filter calculation for the spectrum given in Fig. 6 is shown as one of the curves in Fig. 7. Figure 7 is an example of the display of the events at nodes A, C, and D in the system schematized previously in Fig. 3. The original speech spectrum (the one previously given in Fig. 5) and the difference curve (node D) are shown in addition to the internally synthesized spectrum (node C). The numbers at the left are three numerical measures of the error.

In all of the spectrum matching procedures described in this report, three different error scores were computed and were available as measures of the goodness of fit between the speech spectra and the internally synthesized spectra. The error curve is represented by 36 values (corresponding to the 36 filters) that will be designated as a set of numbers e_i. The error curve is always adjusted automatically such that the weighted mean of e_i is zero over the entire range of values of i. The three error measures are the following:

$$\text{Absolute error} = \sum_{i=1}^{36} |w_i e_i| ;$$

$$\text{Variation} = \sum_{i=1}^{35} |w_{i+1} e_{i+1} - w_i e_i| ;$$

$$\text{Square of error} = \sum_{i=1}^{36} w_i^2 e_i^2 ;$$

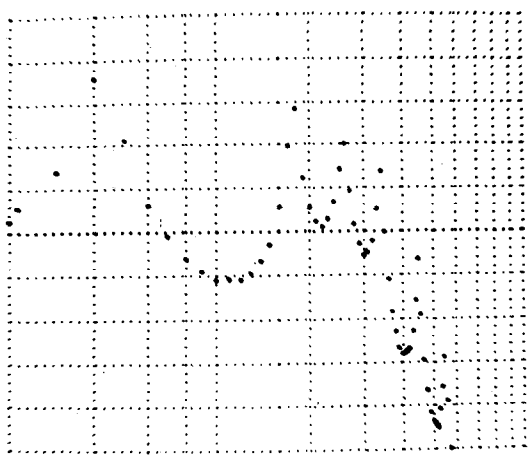

FIG. 6. Photograph of a computed spectrum (at node B in Fig. 3) displayed on the face of the output cathode-ray tube. The horizontal axis is a logarithmic frequency scale from 200 cps to 10 000 cps. The vertical axis represents amplitude in decibels, the small points indicating 1-db steps. The spectrum shown is characterized by resonant frequencies (bandwidths), in cps, of 430 (30), 1770 (80), 2580 (150), plus resonances every 1000 cps from 3500 cps to 9500 cps with bandwidths gradually increasing to 300 cps. Datum points are plotted every 100 cps. For convenience, the datum points at 100 and 200 cps are both plotted close to the 200-cps line.

[8] More precisely, ΔF_n is $1/\pi$ times the real part of the complex frequency of the pole (i.e., $1/\pi$ times the distance of the pole from the $j\omega$ axis in the s plane). When ΔF_n is small compared with F_n, then ΔF_n is very nearly equal to the bandwidth.

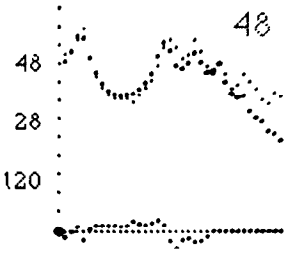

FIG. 7. Photograph of output display showing superposition of input spectrum (upper curve of light points) and comparison spectrum (upper curve of heavy points) obtained at node C of Fig. 3. The lower curve is the difference curve, obtained at node D of Fig. 3. The three numbers at the left show, from top to bottom, the magnitude of the absolute, variation, and squared error scores (see text) over 24 filter points, that is, up to about 3050 cps. The input spectrum is sample 48 shown in Fig. 5; the comparison spectrum is that shown in Fig. 6 after processing by the simulated filters. The error curve is typical of a situation in which a resonance in the comparison spectrum is improperly located.

where w_i represents arbitrary weighting factors which may be assigned depending on the frequency range considered to be important for matching a particular class of spectra. For example, it has been found convenient to match vowel spectra over the values of i from 1 to 24, corresponding to a frequency range of 100 to 3050 cps. A simple way of weighting the error in this case is to put $w_i=1$ over $i=1$ to 24, and $w_i=0$ over $i=25$ to 36. More sophisticated schemes for assigning weighting factors can, of course, be adopted. The variation, being a sum of first differences of the weighted error curve, provides an indication of the amount of fluctuation in the error curve, while the other two numbers provide measures of the amount of deviation of the curve from the zero axis. The significance of the different measures of error will be of particular interest in the discussion of automatic matching procedures in the next section. For experimental matching of various types of speech spectra the square of the error has been used more frequently than the other measures.

For the spectral match shown in Fig. 7, the square of the error, summed over 24 points in frequency, is 120 db^2. The error curve in this case has an irregularity at frequencies in the vicinity of the second resonance. This irregularity is due, apparently, to an incorrect selection of the frequency position of the second pole. When this frequency is adjusted upwards, the match between the two spectra would be expected to become better.

Some indication of the sensitivity of the error scores to small changes in the resonant frequency in this example is given by the upper curve in Fig. 8. This curve has a reasonably sharp minimum of 30 db^2 for a resonant frequency of 1870 cps. Evidently this error score is quite sensitive to small changes in the resonant frequency, and an accuracy better than 30 cps is to be expected in this case. The minimum for the variation error score generally is not as sharp as that for the squared error. Figure 8 also shows the squared error as a function of the frequency of $F2$ for a second formant in a lower frequency range, the second formant of the vowel [ɑ]. Curves of form similar to those of Fig. 8 have been obtained for variations in both bandwidth and resonant frequency in the process of matching a large number of vowel spectra.

When the second resonant frequency of the internally generated spectrum of Fig. 7 is given the value that yields minimum squared error in Fig. 8, the spectral match shown in Fig. 9 is obtained. It is to be noted that a good match to the vowel spectrum shown in Fig. 9 was obtained in the frequency range 100–3050 cps by synthesizing a spectrum characterized by three conjugate pole pairs in this frequency range together with a group of poles at higher frequencies. The high-frequency poles must be included simply to provide the proper levels for the lower resonances.[9] Since the original speech spectrum was pre-emphasized with a slope of 6 db/octave, then, as noted previously, the spectrum envelope of the resulting signal is characterized simply by the set of conjugate pairs of poles of $T(s)$, if an idealized source spectrum envelope with a falling characteristic of 12 db/octave is assumed. As a matter of fact, any significant deviation of the synthesized spectrum from the speech spectrum of Fig. 9 would indicate that the shape of the actual source spectrum differed from this ideal shape. The fact that a good match is obtained in this case indicates that a -12 db/octave slope is a reasonable approximation for the spectrum envelope of the glottal source.[10]

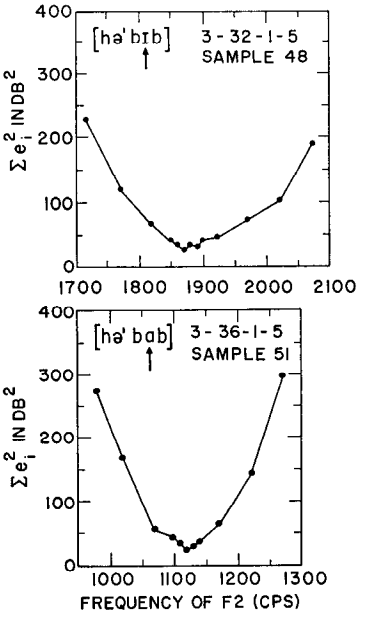

FIG. 8. Examples of the effect on the squared error score of varying the frequency of a single resonance of the comparison spectrum through a range of values in the vicinity of the actual vowel formant frequency. The upper graph refers to a sample taken centrally in a vowel characterized by a high-frequency second formant (sample 48, as in Figs. 4, 5, 7). The lower graph illustrates the same result for a sample from a vowel with a second formant at a lower frequency.

[9] G. Fant, "On the predictability of formant levels and spectrum envelopes from formant frequencies," *For Roman Jakobson*, edited by M. Halle *et al.* (Mouton and Company, 's-Gravenhage, 1956), pp. 109–120.

[10] The matching of a large number of spectral samples in voiced speech sounds has shown the same general results. These analysis procedures, however, are not highly sensitive to local variations in the shape of the glottal spectrum since the analog filters used in processing the speech materials are not very selective. Thus the form of the glottal spectrum derived by this method cannot be compared in detail with that derived from time-domain analyses.

By the procedure discussed above, matches have been obtained for a number of spectra associated with vowel and consonant portions of utterances by several male talkers. Systematic studies of the pole-zero patterns for various time locations through these utterances have been made[11] and detailed reports are in preparation. Examples of the matches obtained for three classes of speech sounds other than nonnasal vowels are shown in Fig. 10. In all cases it was possible to select a set of poles and zeros such that good fits were obtained with the data. In the matching of a spectrum such as one of these, the initial step was to determine the approximate locations of the poles and zeros from theoretical considerations and by examination of the general shape of the spectrum. Convergence to pole and zero locations yielding an optimum fit was achieved through a trial-and-error process, always with the constraint that the locations be consistent with known theoretical relations between vocal-tract configurations and the acoustic signal.

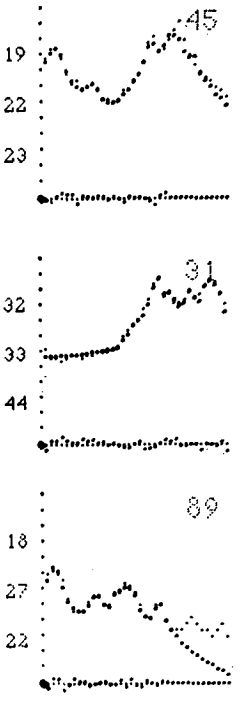

FIG. 10. Photographs of matches similar to that of Fig. 7 for spectra of speech sounds other than nonnasal vowels. The upper photograph shows a match to a (nasalized) vowel [*i*] occurring in the phonetic environment [*m-m*]. The comparison curve was constructed from three poles (and their conjugates) up to 3200 cps, a correction for higher poles, and a pole-zero pair in the vicinity of 1200 cps; the error scores were computed over 24 filters.[16] The middle photograph is a match to the fricative [ʃ] in initial position in a stressed syllable. In this case the error score was computed over 36 filters up to 7000 cps and the comparison curve was constructed from seven poles, three zeros, and three real-axis zeros close to zero frequency.[11] In the lower photograph the speech spectrum was sampled from an [*n*] in word final position and the error score was computed over 24 filters. The comparison curve was constructed from five poles and a zero up to 3000 cps, plus a higher-pole correction.[11]

AUTOMATIC MATCHING OF VOWEL SPECTRA; THE STRATEGY PROBLEM

Although application of the spectrum-matching technique described above usually resulted in good agreement between the speech spectra and the synthesized spectra, and, presumably, in reasonably accurate values for spectral poles and zeros, the method has the disadvantage that it is tedious and is not completely automatic. Attempts have been made, therefore, to reduce the time required for the generation of comparison spectra and to program the computer to perform the function of the experimenter in the analysis scheme of Fig. 3. The task of developing an optimum strategy whereby rapid convergence to a best-fitting vowel spectrum is achieved is by no means trivial, and is an example of the hill-climbing problem that has received considerable attention in the field of pattern recognition.[12] The strategy that is used in the present automatic matching scheme is a rather elementary one; more complex strategies are being developed to obtain more accurate and rapid analyses.[13,14] The automatic method to be described is applicable only to the analysis of the spectra of nonnasal vowels.

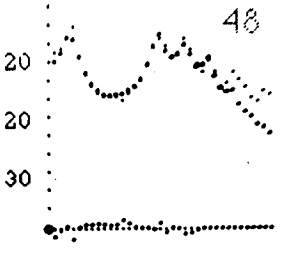

FIG. 9. Photograph of a display such as that described in Fig. 7, showing a good match between the input and comparison spectra. To obtain this match to the vowel [ɪ] the values of the lowest four resonant frequencies (bandwidths), in cps, of the comparison spectrum were 430 (30), 1870 (80), 2580 (150), and 3400 (120); additional resonances were spaced at 1000-cps intervals from 4500 to 9500 cps with gradually increasing bandwidths up to 300 cps.

Generation of Comparison Spectra

In the experimental analysis procedure described above, each internally generated spectrum is computed as needed, and the integrations to simulate the effect of the filter bank are performed for each trial spectrum. The calculation of each pole factor and the simulation of the filtering of the synthesized spectra are the most time-consuming portions of the above method, however, and automation of the matching procedure would not be practical unless the time required for these operations was reduced. A more rapid (but less precise) procedure for the generation of comparison spectra was therefore adopted. In this procedure the comparison spectra are assembled from a limited set of elemental spectra that are stored within the computer memory.

Five elemental spectra are added to obtain the spectrum envelope of a vowel (in decibels) in the frequency range that includes the first three vocal-tract resonances (up to about 3000 cps for adult male voices). Four of these spectra are simple resonance curves each of which corresponds to a conjugate pair of poles of the vocal-tract transfer function. An inventory of 78 simple resonance curves is stored in the computer memory with resonant frequencies spaced every 20 cps from 160 to

[11] A. S. House, K. N. Stevens, and H. Fujisaki, J. Acoust. Soc. Am. **32**, 1517 (1960); J. M. Heinz, J. Acoust. Soc. Am. **32**, 1517 (1960); O. Fujimura, J. Acoust. Soc. Am. **32**, 1517 (1960).

[12] M. Minsky, Proc. Inst. Radio Engrs. **49**, 8 (1961).

[13] A. Paul, S.M. thesis (unpublished), M.I.T. (1961).

[14] M. V. Mathews (personal communication, 1961).

500 cps, every 50 cps from 500 to 3000 cps, and every 100 cps from 3000 to 4000 cps. The bandwidth of the resonance associated with each of these curves is fixed at a value suggested by measured data on formant bandwidth; it is 60 cps for the low-frequency resonances and increases to 180 cps for the high-frequency resonances. Vowel spectra with various combinations of resonances can be assembled by selection of appropriate groups of four such curves. The fifth elemental spectrum is a curve that, in the frequency range of the first three vocal-tract resonances, accounts for the source spectrum, the radiation characteristic, and poles of the vocal-tract transfer function higher than the fourth. This "correction" spectrum is a relatively smooth curve, and its shape is not expected to change markedly from one adult male speaker to another or from one vowel to another, although some variation in the slope of the curve may occur. An inventory of six such correction spectra is stored in the computer memory, and one of these is always added to the group of four resonance curves to synthesize a complete vowel spectrum.

The 84 elemental spectra that are stored in the computer memory are actually the curves that would be obtained if each of the simple resonance spectra and correction spectra were processed by the filter bank in the manner discussed above [Eq. (12)]. The elimination of the necessity of spectrum calculation and filter simulation greatly reduces the time required for the generation of comparison spectra (by a factor of about 50 in the present case), but it inevitably leads to some error in the synthesized spectra, especially at high frequencies where the filter bandwidths are not constant. Correction for the primary effect of the filter bandwidths can be made, and is actually included in the above procedure. It can be shown, however, that compensation for this error cannot be made exactly, especially for cases in which two resonances are closely spaced. Consequently this procedure for assembling vowel spectra has some inherent error, although this error is usually quite small.

Description of Strategy

From the 84 stored elemental curves, it is possible to assemble about 5×10^5 vowel-like spectra, if reasonable assumptions are made concerning the frequency range for each of the first four formants. Since in the analysis of a given speech spectrum it is impractical to make a comparison with each of these synthesized spectra, it is essential to devise a strategy whereby only a small subset of comparison spectra needs to be assembled and tested before convergence to the best-fitting spectrum is achieved. It is possible to distinguish two situations that require different strategies. One situation arises when no prior information is available concerning the formant frequencies for the vowel spectrum under analysis, and/or when no previous data have been obtained for the talker who generated the utterance in which this spectrum occurs. Such a case would occur when, in the analysis of the formant frequencies during the vowel portion of a syllable, one spectrum is selected to be examined first. Here the basic task of the analyzer is to establish a good first approximation to the input spectrum. In the second situation, which occurs much more frequently than the first, approximate data concerning the formant frequencies and the appropriate correction spectrum are already available in the analyzer. These data may have been obtained either from analysis of a spectrum sample located adjacent to the spectrum to be analyzed or from a preliminary approximate analysis of the spectrum. In this case, the task of the analyzer is to optimize the match between the input and the synthesized spectra.

When there is no prior knowledge of the locations of the formants, one method that has been used to establish the approximate values of the formant frequencies consists of the following steps: (1) Elemental spectra corresponding to formant frequencies in the expected range of F_1 (plus a standard F_4 curve and a standard correction curve) are each compared with the speech spectrum to be analyzed, and the curve yielding the minimum variation error is selected tentatively as identifying F_1. (2) Elemental spectra corresponding to formant frequencies in the expected range of F_2 are each added in turn to the curve found in (1) and the composite curve yielding the minimum variation error is selected tentatively as identifying F_2. (3) Step (2) is repeated to find tentative values for F_3. (4) After approximate values for the first three formant frequencies are found in this way, elemental spectra corresponding to formant frequencies in the expected range of F_4 are each added in place of the standard F_4 curve adopted in previous steps, and the one yielding the minimum error score is found. (5) Step (4) is repeated to find the correction curve yielding the minimum error score. (6) The set of first four formant frequencies and the correction curve found by the above procedures are then used as starting points for the more exact analysis procedure that is employed when approximate data of this type are available.

It is to be noted that in the first step above, no elemental spectra corresponding to F_2 and F_3 are included in the synthesized spectrum. It can be shown, nevertheless, that the variation error score can serve to locate the approximate position of the lowest resonance in the input spectrum. The squared error score can give reliable results only after a reasonable approximation to the input spectrum is established, and is not a good criterion at this stage of the preliminary analysis. Throughout the automatic procedures pertaining to the analysis of vowels produced by adult male talkers error scores were computed with equal weighting for the 24 filters in the frequency range 100–3050 cps.

In the process of developing the preliminary analysis procedure outlined above, various alternative schemes for obtaining a first approximation to the formant fre-

quencies were tried. In one such scheme the spectrum of a neutral vowel, i.e., $F_1=500$ cps with subsequent formants occurring at intervals of 1000 cps, was used as the zero-order approximation, and the frequency positions of the formants were revised successively within appropriate ranges. Another scheme involved the matching of the input spectrum against members of a small stored set of standard vocalic spectra and the selection of the best approximation. A third procedure obtained estimates of approximate values of formant frequencies from direct measurements of certain gross features of the input spectrum.[15] Further studies with a large number of talkers and utterances will be required before the over-all performance of these various preliminary analysis procedures (or possibly combinations of them) can be compared quantitatively.

When approximate values for the formant frequencies and correction spectrum are available, an iterative procedure is employed, and the sequence of operations is the following: (1) With F_2, F_3, F_4, and the correction spectrum fixed at the given values, curves with resonant frequencies in the vicinity of the given F_1 are used to form a series of spectra that are compared with the speech spectrum to be analyzed. The value of F_1 yielding the minimum squared error is selected and used in subsequent steps. (2) Step (1) is repeated but with F_1, F_3, F_4, and the correction spectrum at the given values and F_2 as the variable. (3) Step (1) is repeated to find, in turn, revised values for F_3, F_4 and the correction spectrum. (4) Steps (1)–(3) are repeated. If the results are the same as those obtained after the first set of trials, the analysis of the given spectrum is terminated; otherwise the process is repeated until no improvement in the fit is obtained.

The automatic method for the analysis of vowels has been used to obtain data on the variation with time of the formant frequencies of stressed vowels in a number of dissyllabic utterances. The computer has been programmed to perform the analysis on each spectral sample in turn within a designated region of the utterance. The initial step in the procedure is to prescribe the range of spectral samples over which the analysis is to be performed and to select a sample located centrally within this range. The analysis is first carried out on the centrally located sample, following one of the procedures that require no *a priori* knowledge of the approximate formant frequencies. The more precise iterative procedure is then applied to this sample to locate the formant frequencies and correction spectrum more exactly. These values of formant frequencies and correction spectrum are used as first approximations in the analysis of the following spectral sample. In this manner the analysis is performed on each spectral sample in turn until the end of the designated interval is reached. Then the program returns to the centrally located sample, and uses the results previously obtained for this sample as first approximations to the next preceding sample. The analysis is carried out on each preceding sample moving toward the beginning of the designated time interval. The results for each spectral sample are stored in the computer memory. After the analysis of all samples is completed, an instruction can be given to the computer to print or punch out the results of the analysis of each sample in order, or to display the results on the oscilloscope in various ways.

[15] F. Poza, S.M. thesis (unpublished), M.I.T. (1959).

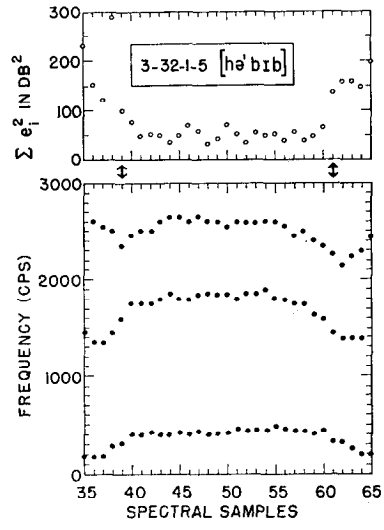

FIG. 11. Display of data on vowel formant frequencies derived by the automatic analysis procedure. The utterance is the same as that shown in Fig. 4. Time is on the horizontal axis and is indicated in terms of samples which occur at 8.3-msec intervals. The solid points represent the lowest three vowel formants in each sample as determined by the automatic procedure. The open points at the top of the figure give a measure of the error of fit between the input and comparison spectra. The arrows indicate points in time where study of the spectrogram of the utterance (see Fig. 4) suggests the locations of vocalic boundaries.

Typical Results

Figure 11 displays typical results of the automatic vowel analysis program for a portion of the utterance whose spectrogram is shown in Fig. 4. The first three formant frequencies found by the program are plotted for each spectral sample in the stressed vowel. The squared error score for each sample is also shown in the upper part of the figure. The arrows indicate the "vowel" boundaries suggested by study of the spectrogram. It is noted that the error score increases sharply at these boundaries, since it is not possible, of course, to obtain good matches with consonant spectra by assembling a set of simple resonance curves by a procedure based on a theory of vowel production.

Several limitations of the automatic procedure have already been pointed out, and further studies will be necessary to overcome these limitations. The automatic analysis procedure in its present form requires that many trials be made before convergence to a set of resonant frequencies is achieved, and consequently the

analysis takes a considerable amount of time (order of 1000 times real time for the computer and the programs used in these studies). Furthermore, small but systematic errors in formant locations occur as a result of (a) the incomplete correction for the effect of the filters in the construction of the comparison spectrum and (b) the inability to vary the bandwidths of the formants. Both of these types of errors can be eliminated if a more complex and time-consuming procedure is used to assemble the spectra, similar to the procedure used in the experimental method described in connection with Fig. 3. If, however, formant bandwidth were a variable in the matching process, then a more detailed strategy would be necessary to converge to both the frequencies and the bandwidths appropriate to a given spectral sample.[13]

Remarks on Extension of Automatic Analysis Procedure to Other Classes of Speech Spectra

The automatic speech reduction procedure just described is applicable only to spectra of nonnasal vowels or vowel-like sounds for which the vocal-tract transfer function is characterized by a set of conjugate pairs of poles. Thus for the matching of these types of spectra the internal spectrum generator in Fig. 2 need be instructed simply to synthesize spectra corresponding to a product of terms each of which represents a conjugate pair of poles. On the other hand, completely automatic procedures for reduction of spectra other than those of vowels or vowel-like sounds have not yet been developed. This lack of progress stems largely from the fact that the generation of these other classes of sounds is not yet understood in detail. While it is known, for example, that spectra occurring during the production of nasal, stop, and fricative consonants are characterized by zeros as well as poles, the numbers of zeros and poles required and the frequency ranges to be expected for each cannot be specified easily and systematically on the basis of present knowledge.

The spectrum of a nasalized vowel, for example, is characterized by about four poles and one zero in the frequency range up to 3000 cps,[16,17] but the problem of devising a strategy that would lead to automatic matching of such a spectrum is a formidable one. If the positions of the four poles and zero were varied independently, a large number of combinations would have to be tried, but in order to avoid erroneous results these should include only those combinations that could in fact represent outputs of a vocal tract. To meet this requirement constant reference to articulation would have to be made during the process of searching for suitable pole-zero combinations.

[16] O. Fujimura, "Analysis of nasalized vowels," Quart. Progr. Rept. **62**, Research Laboratory of Electronics, M.I.T. (1961), pp. 191–192.

[17] Reference 4, pp. 148 ff.

In view of these complications it is suggested that the strategy in an automatic analysis-by-synthesis procedure that is applicable to all types of spectra should consist of a search for parameters that are more directly related to articulation than are the pole-zero locations. In effect, the proposed strategy would require a search through a set of articulatory configurations. For each trial configuration the pole-zero locations, and hence the over-all spectrum, would be computed and compared with the spectrum under analysis. Different articulatory configurations would be tried until a spectrum yielding a best fit with the input spectrum was obtained. Thus in the case of matching the spectrum of a nasalized vowel, the strategy would try different vowel configurations and different amounts of coupling to the nasal cavities until an optimum spectral match was obtained.

The realization of this type of analysis scheme requires that a model be developed for specifying articulatory configurations in a simple yet meaningful way. Although various simple models have already been proposed[18–20] it is clear that much must be learned concerning articulatory constraints and the relations between articulation and the acoustic output before a suitable strategy for the automatic reduction of all kinds of speech spectra is developed.

DISCUSSION

Analysis-by-synthesis procedures for the reduction of speech spectra have been used in one form or another by several investigators. Early attempts to use a spectrum matching technique were reported by Steinberg[21] and by Lewis,[22] who matched simple resonance curves to vowel spectra in the vicinity of the spectral peaks. The method was carried much further by Fant,[4,23] who demonstrated how the spectra associated with simple linear circuits can be matched against vowel and consonant spectra. In Fant's studies, the experimenter can be said to have been situated within the feedback loop (as in Fig. 3 above) and the comparison spectra were either computed or measured from simple analog circuits. The goodness of fit was assessed by visual examination of the curves. Similar procedures were used by Heinz and Stevens[24] for the matching of the spectra of fricative consonants. Matching of the spectra of several vowels was achieved by Mathews, Miller, and David,[25] who used digital computer techniques for the analysis

[18] Reference 4, pp. 71 ff.

[19] K. N. Stevens and A. S. House, J. Acoust. Soc. Am. **27**, 484 (1955).

[20] O. Fujimura (personal communication, 1960).

[21] J. C. Steinberg, J. Acoust. Soc. Am. **6**, 16 (1934).

[22] D. Lewis, J. Acoust. Soc. Am. **8**, 91 (1936).

[23] C. G. M. Fant, "Transmission properties of the vocal tract, II." Quart. Progr. Rept. Acoustics Laboratory, M.I.T. (Oct.–Dec. 1950), pp. 14–19.

[24] J. M. Heinz and K. N. Stevens, J. Acoust. Soc. Am. **33**, 589 (1961).

[25] M. V. Mathews, J. E. Miller, and E. E. David, Jr., J. Acoust. Soc. Am. **33**, 179 (1961).

of spectra computed from individual periods of the glottal output. They devised procedures for finding a set of poles corresponding to the vocal-tract transfer function and zeros to approximate the detailed form of the glottal spectrum such that best fits were obtained with the spectra under analysis. By performing a "pitch synchronous" analysis, they were able to obtain a rather detailed picture of the characteristics of the glottal excitation as well as the vocal-tract resonances, although the procedure was complicated by the necessity for adjusting a large number of parameters in order to converge to a best fit. The principles of the active speech analysis procedure have also been enunciated by Inomata,[26] who, in connection with a program concerned with automatic speech recognition, has used computer techniques to search for a set of poles that yield a spectrum that matches a given vowel spectrum.

Whereas the methods just summarized, as well as those described in this paper, involve the matching of speech spectra and thus are carried out in the frequency domain, analysis procedures based on the same principle can also be applied in the time domain. The "inverse filtering" techniques described by Miller[27] involve the processing of the vowel sounds by a cascaded set of filters that are characterized by a set of conjugate pairs of zeros. When the frequencies of the zeros are adjusted to coincide with those of the poles that describe the vocal-tract transfer function for the vowel, then the output of the filters represents the waveform of the glottal source. Since the general shape of the glottal pulse is known, then a procedure can be devised for adjusting the zeros until the expected shape is obtained. The processing of the signal by a cascaded sequence of filters in the time domain is analogous to subtracting elemental resonance spectra (in decibels) from the speech spectrum. It would appear difficult, however, to devise an automatic analysis procedure based on time-domain methods, since criteria for optimum cancellation of a pole by a zero might be difficult to devise.

The various versions of analysis-by-synthesis or feedback methods of speech spectrum analysis such as those that have been described here and by others are considered to have important advantages over other analysis schemes. For the feedback analysis method, once a set of parameters is found such that a good replica of the input signal is generated when these parameters are applied as instructions to the internal generative model, then there is little question that this set constitutes an adequate representation of the input. In contrast to this method are the passive or open-loop analysis procedures in which simple attributes of the spectra, such as the major spectral peaks, are measured directly and are used to provide a simple representation of the speech signal. There is no assurance in these cases that important data have not been discarded or that an error has not been made in the extraction of a particular parameter.

Other potential advantages of the feedback analysis procedure stem from the fact that it permits certain quasi-invariant features of the speech signal to be accounted for in a relatively straightforward manner. Thus, in principle, once certain properties of a given talker, such as the spectrum of the glottal output or the approximate range of variation of his formant frequencies, have been evaluated, then these properties can be assumed to remain relatively unchanged over a period of time, and the strategy during this period is simplified. In a sense, the method is geared to the extraction of features of the signal that are changing, and spends little time on the extraction of features that do not change or that change only slowly.

The similarity between an analysis-by-synthesis procedure and certain aspects of human perception have led several investigators to speculate that man manipulates sensory data such as speech by an active internal replication process.[5–7,26,28,29] If there is any basis for such speculation, then analysis techniques of the type described here would have the additional advantage that they bear at least some resemblance to the process of human speech reception.

ACKNOWLEDGMENTS

The authors have profited from stimulation, counsel and technical assistance contributed by their associates. Discussions with Osamu Fujimura have particularly influenced the course of the research. Ideas contributed by George Rosen and Fausto Poza are also acknowledged with gratitude, as is the criticism and encouragement of Morris Halle, the programming aid of Paul T. Brady and the technical assistance of Jane Arnold. Finally, the work would not have been possible without the availability of the TX-O computer, a facility of the Department of Electrical Engineering, M.I.T., and the cooperation and help of its technical staff.

This work was supported in part by the U. S. Army Signal Corps, the Air Force Office of Scientific Research, and the Office of Naval Research; and in part by the Air Force Cambridge Research Laboratories.

APPENDIX: SPEECH INPUT SYSTEM

Sampled speech data are introduced into the computer in spectral form using equipment the block diagram of which is shown in Fig. 12. Speech is recorded on one channel of a two-channel magnetic tape loop and sampling pulses are recorded on the other channel. The speech is played back through a pre-emphasis network into a

[26] S. Inomata, Bull. Electro-Tech. Lab. (Tokyo) **24**, 597 (1960).
[27] R. L. Miller, J. Acoust. Soc. Am. **31**, 667 (1959).
[28] G. A. Miller, E. Galanter, and K. H. Pribram, *Plans and the Structure of Behavior* (Henry Holt and Company, New York, 1960).
[29] L. A. Chistovich, Soviet Phys.—Acoustics **6**, 393 (1961); [Akust. Zhur. **6**, 392 (1960)].

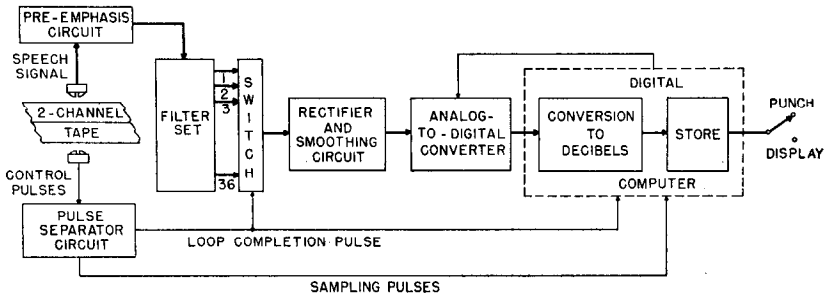

FIG. 12. Block diagram of the procedure used in preparing speech materials for computer analysis.

bank of 36 simple-tuned filters. The pre-emphasis network has a rising frequency characteristic of 6 db/octave. The center frequencies of the filters range from 150 to 7025 cps and are selected so that the half-power points of adjacent filters are coincident. The filter bandwidths are constant at 100 cps for center frequencies up to 1550 cps and then increase gradually until reaching a value of 475 cps for a center frequency of 7025 cps. During the read-in process, the outputs of the filters are selected in sequence by a stepping switch that steps after each cycle of the tape loop. Thus the loop is played 36 times to obtain a complete spectral analysis of the speech sample. The selected filter output is full-wave rectified and smoothed before being converted from analog to digital form. A commercial analog-to-digital encoder performs this conversion.

The second tape channel contains recorded control pulses. A pulse train of positive polarity in which the pulses occur every 8.3 msec is used to indicate times at which the data are to be sampled. A train of opposite polarity marks the end of the tape loop and initiates the stepping switch. These control pulses enter two light-pen flip-flop registers of the computer, so that the sampling can then be controlled by the computer.

The computer is programmed to search the light-pen flip-flop registers for "sample" pulses and to transfer data from the encoder when such a pulse appears. The filter outputs are encoded into 10 bits and are read into the computer, where the data are then converted into decibels, encoded into six bits, and rearranged so that three samples are stored in each 18-bit memory register. Thus each group of 12 registers contains outputs of the 36 filters at one sample time. Successive groups of 12 registers contain speech spectra at successive 8.3-msec intervals. With the present 8192-word memory, 3648 registers are used for data storage, and thus approximately 2.5 seconds of speech can be processed. The program provides routines that allow the data to be displayed on an oscilloscope or punched out on paper tape for later use. In addition, several error-checking routines are built into the program to maintan the accuracy of the read-in process.