XII. SPEECH COMMUNICATION*

Prof. K. N. Stevens	Dr. T. T. Sandel	S. Inomata†
Prof. M. Halle	Jane B. Arnold	A. P. Paul
Prof. J. B. Dennis	J. M. Heinz	J. R. Sussex
Dr. A. S. House	W. L. Henke	E. C. Whitman

A. SPEECH SYNTHESIS

The primary objective of our research in synthetic speech is to develop a thorough understanding of the process of speech production - an understanding that may be used as a basis for study of the perception of speech sounds and the development of new techniques of speech communication. We are concerned with the synthesis approach to the study of speech production in which a model of the vocal mechanism is used as a basis for speech synthesis. The resulting speech can be evaluated by physical comparison with natural speech or through psychological judgement by a listening panel. Parameters of the model may be adjusted to reduce the error in a physical comparison, or to obtain the most satisfactory listening performance. In particular, our goal is the improvement of the M.I.T. dynamic vocal-tract analog (DAVO) so that it will be a precise and flexible tool for future speech research.

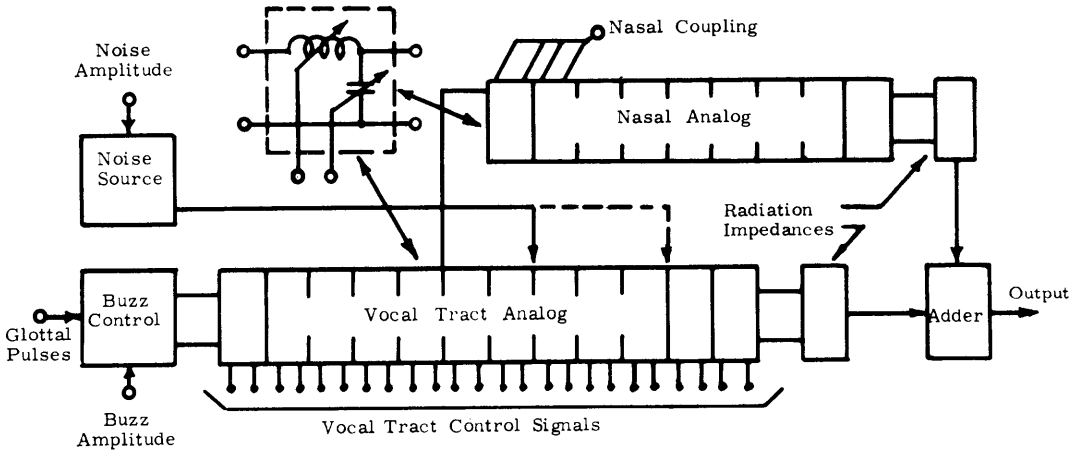

Fig. XII-1. Functional schematic diagram of the dynamic vocal-tract analog showing the required control signals for performing a synthesis.

*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.

The dynamic vocal-tract analog was constructed by George Rosen,[1] and is described by Fig. XII-1. In the past, the control signals for DAVO have been obtained from a number of trapezoidal waveform generators, operated by a timing generator, which supplied the buzz amplitude, noise amplitude, and nasal coupling signals to the vocal-tract model.[2] One trapezoidal waveform was used to switch the vocal-tract control voltages smoothly between two vocal-tract area configurations selected from six configurations set up by means of potentiometers. An impulse oscillator with frequency controlled by a trapezoidal waveform provided glottal pulses to the vocal-tract model. Relays turned on and off at preset times were used to perform selection functions.

1. Control by Digital Computer

The use of a general-purpose digital computer to generate the control signals for the vocal-tract analog offers two advantages over the specialized control system previously used: The computer can generate a long, continuous sequence of control signals in which the motions of the vocal-tract parameters are under very precise control, while the special control system is limited to one or two segments of trapezoidal shape for any control signal. The computer also has the advantage that it can perform sophisticated numerical and symbolic manipulations. This allows a language to be developed for the description of utterances to be synthesized which will simplify the process of experimental speech synthesis. As many as one hundred separate adjustments are required to set up the synthesis of a monosyllabic utterance with the special control system.

For these reasons equipment has been obtained to permit control of the vocal-tract model by the TX-0 computer. This apparatus is shown on the right side of Fig. XII-4. This figure also describes the control program that will be discussed later. A set of digital-to-analog converters provides the control signals for the synthesizer. A pulse generator issues a single glottal pulse whenever it is commanded by the computer program. A relay register performs such selection functions as determining the vocal-tract insertion point for the noise source. All of these devices are actuated through a special computer instruction provided for operating users' equipment. The glottal pulses are produced in the manner indicated above, rather than by a controlled oscillator, so that a synthesized utterance will be exactly reproducible, and to allow the greatest precision in mimicking natural speech.

2. Program Facility for Speech Research

Our aim is to establish a facility that will enable experimenters to study the speech phenomenon by using the dynamic vocal-tract analog with the greatest convenience and flexibility that can be achieved through the powers of the computer. For this purpose the arrangement illustrated in Fig. XII-2 is envisioned. The vocal-tract model is

operated by the control program outlined below. Its purpose is to simplify the specification of the speech sample to be synthesized, while placing as little constraint as possible on the range of sounds that can be produced. The input to the control program is a time-sequenced list of events to take place in the vocal-tract model. The event compiler will allow the experimenter to conveniently prepare the list of events required for a synthesis. It will also permit him to modify the sequence of events and vary parameters of a synthesis as desired. This facility will allow the investigator to prepare the specification of an utterance, perform the synthesis, evaluate it through informal listening or by physical comparison, readjust the specification, and make further trials within one experimental session.

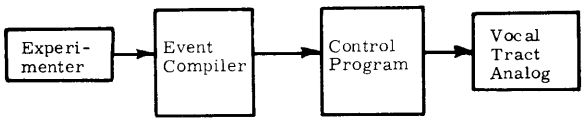

Fig. XII-2. Program facility to give the experimenter flexible control of the dynamic analog.

3. Control Program

To simplify the task of specifying the signals that must be presented to the analog vocal tract, it is necessary to provide a means of organizing this information. This is the purpose of the control program.[3] However, to retain the greatest amount of flexibility at the same time, it must be possible to operate the analog in any meaningful pattern of which it is capable by means of an appropriate input to the control program. Ideally, the control program should place no restriction on the gamut of sounds which the vocal-tract model may produce.

There are two dimensions in which organization can be supplied by the control program – time and space. In the time dimension, the motion of the buzz and noise amplitude signals and the nasal coupling signal can be described conveniently by a succession of parabolic segments as indicated in Fig. XII-3. Each member of the sequence is

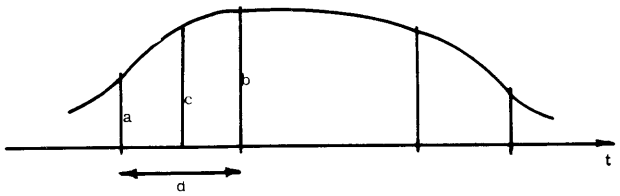

Fig. XII-3. Representation of a control signal by a sequence of parabolic segments. Each segment is specified by four parameters as indicated by a, b, c, and d.

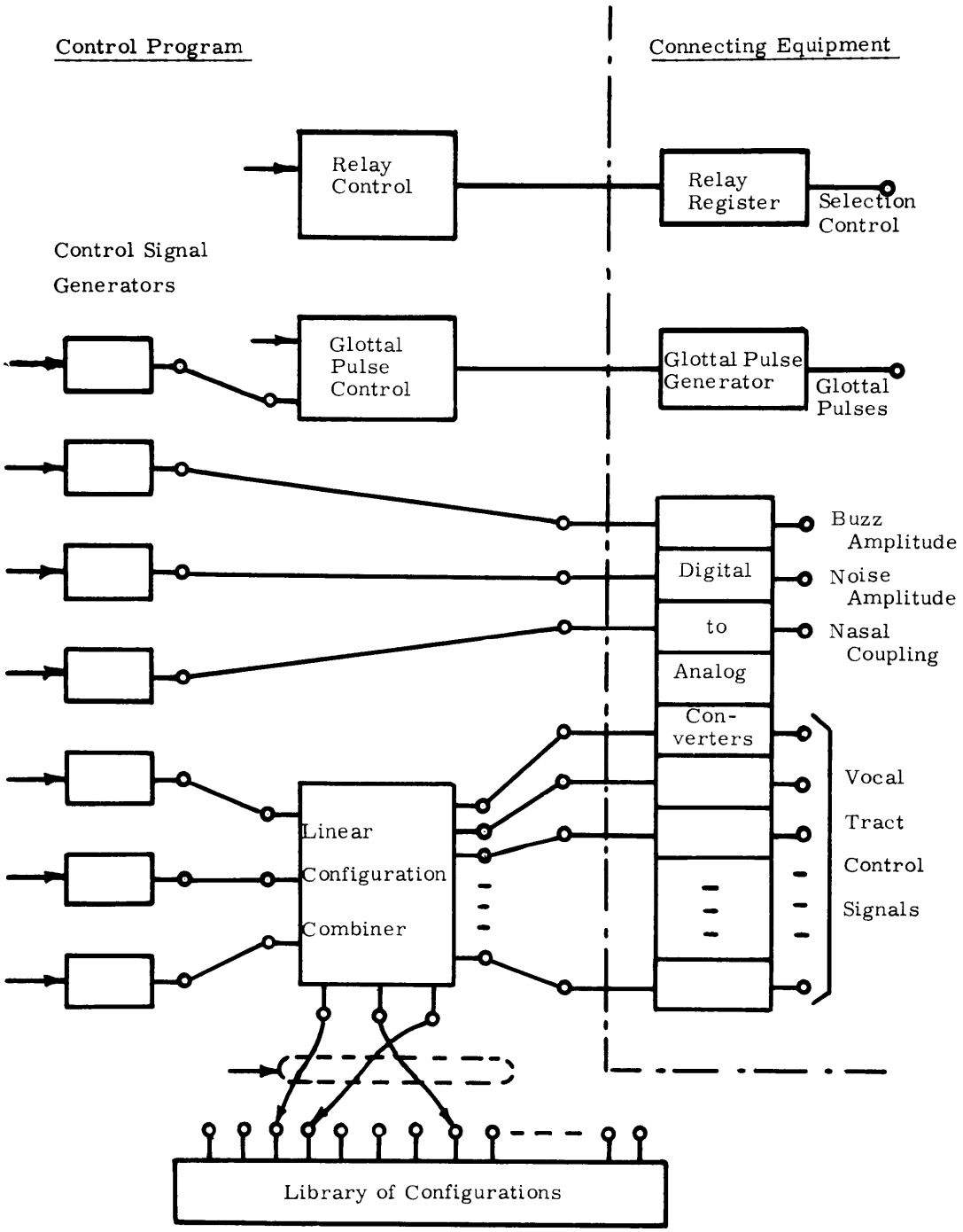

Fig. XII-4. Descriptive diagram of the operation of the control program. Equipment for connecting the TX-0 computer to the dynamic analog is shown on the right.

specified by initial, intermediate and final values, and its duration. No significant loss of flexibility is involved, since the length of any segment may be made as small as desired.

A general method of organizing the transmission-line parameters is to form a library of section area functions and section length functions, a pair of which will be called a vocal-tract configuration. The specification of the time pattern of vocal-tract parameters could be done by selecting one configuration from the library for each time unit. A more general scheme is to represent the set of vocal-tract parameters in any time unit by a linear combination of several stored configurations selected from a library. The time behavior of the coefficients of the linear combination is given as a sequence of quadratic segments in the manner described above. With this arrangement, it is possible to specify vocal-tract action in the greatest detail by stating the configuration separately for each time unit. However, the representation as a linear combination of stored configurations should permit a great reduction, both in the amount of detail needed to construct a synthesis and in the necessary number of stored configurations.

The control program is designed to accept the specification of an utterance in the form outlined above. The operation of this digital computer program is best presented by giving a block diagram (Fig. XII-4) expressed in terms of the corresponding operations on analog signals. Each of the control signal generators produces a succession of quadratic segments in time. Three of these boxes are used to control vocal-tract excitation and nasal coupling. Three or four more may be used to supply inputs to the linear combiner which computes the vocal-tract parameters as a linear combination of stored configurations.

The list of events comprising the input to the control program may contain the following kinds of entries:

(a) Control signal generator initiation - The control program supplies the specified control signal generator with a set of four parameters for the next parabolic segment that it will produce.

(b) Glottal pulse entry - This entry tells the control program whether to produce pulses with frequency given by a control signal generator, or to produce a pulse at a time stated in the entry.

(c) Relay control entry - The contents of the relay register are changed by the control program to perform a selection function in the vocal-tract analog.

(d) Configuration entry - A new configuration from the library replaces another as input to the linear configuration combiner.

The control program performs the computation indicated in Fig. XII-3 once for each time unit. When the accumulated time units sum to the time of the next entry in the event list, the next event is interpreted.

J. B. Dennis

References

1. G. Rosen, Dynamic Analog Speech Synthesizer, J. Acoust. Soc. Am. 30, 201-209 (1958).

2. M. H. L. Hecker, Studies of nasal consonants with articulatory speech synthesizer, J. Acoust. Soc. Am. 34, 179-188 (1962).

3. J. R. Sussex, Computer Control of a Dynamic Analog Speech Synthesizer, S.M. Thesis, Department of Electrical Engineering, M.I.T., August 1962.

B. REDUCTION OF SPEECH SPECTRA TO DESCRIPTIONS IN TERMS OF VOCAL-TRACT AREA FUNCTIONS

A general procedure by which speech spectra may be analyzed in terms of parameters that are simply related to the configuration of the vocal tract was given in a previous report.[1] Also reported was a preliminary computer program written for the TX-0 computer for calculating the poles of the vocal-tract transfer function from a specification of the cross-section area at a number of points along the tract. This program has now been extended to allow both the poles and zeros of the transfer function to be calculated for a "pressure-difference" source located at some point along the tract.

In order to reduce the number of input parameters that must be varied to obtain spectral matches, a modification of the Stevens-House three-parameter model of articulation is used.[2,3] For this modification, four parameters must be specified: (a) r_o, the effective radius of the vocal tract at the tongue constriction; (b) d_o, the location of the tongue constriction; (c) a_o, the cross-section area of the mouth opening, and (d) ℓ_o, the length of the mouth section. A computer program has been written which accepts these four parameters, r_o, d_o, a_o, and ℓ_o, and calculates the corresponding cross-section area parameters according to the rules of the model. The output area parameters of this program are used as input parameters for the previous program in order to calculate the pole and zero locations. Finally, a spectrum is calculated from the pole and zero locations for comparison with input speech spectra by using the procedure and computer program outlined by Bell, and others.[4] The combination of these programs which is used to find spectral matches with input speech data in terms of the four articulatory parameters is called TRACT III.

The procedure just described has been used to obtain spectral matches for several vowels and consonants. The experimenter types in the values of the four parameters and a spectrum is then calculated for comparison with the input speech spectrum. Initially, all bandwidths are set to 100 cps, although they may be adjusted by the experimenter from the typewriter.

Figure XII-5 shows a spectral match obtained for the vowel /ɑ/ occurring in the syllable /hə'ʃɑʃ/. The two curves at the top of the figure show the input data (small dots)

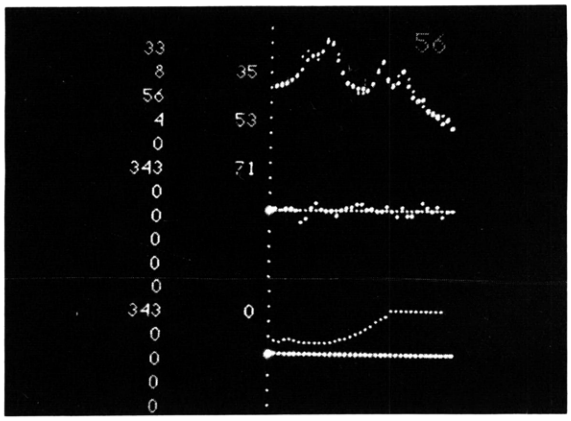

Fig. XII-5. Photograph of cathode-ray tube displays for the vowel /ɑ/ from the syllable /ʃɑʃ/ depicting: top curve, comparisons of speech data (small dots) with corresponding calculated points (large dots), middle curve, differences between the speech data and the calculated points, and bottom curve, the corresponding vocal-tract configuration, the glottis being at the left and the lips at the right.

and the internally generated spectrum (large dots). The difference curve is shown at the middle of the figure. The bottom curve depicts the effective radius as a function of distance along the vocal tract, the glottis being at the left and the lips at the right. The configuration is characterized by a constriction well back in the vocal tract toward the glottis and a fairly open mouth cavity. The four upper numbers in the column at the left represent the four parameters; in this case, r_o=33(0.44 cm), d_o=8(4 cm), a_o=56(11 cm^2), and ℓ_o=4(2 cm). Thus the total vocal-tract length is 16.5 cm. The three error scores are given by the numbers in the right-hand column and indicate a sum of absolute errors of 35 db, a variation error of 53 db, and a sum of square errors equal to 71 db^2. The number at the upper right-hand corner indicates the location of this sample in the utterance being considered. The formants of the internally generated spectrum occur at the frequencies 840, 1210, 2530, 3420, 4520, 5560, and 6650 cps. Spectral matches have also been obtained for the vowels /u/, /i/, and /ɛ/ occurring in fricative consonant environments.

The procedure used for matching some initial voiceless fricative consonants in consonant-vowel-consonant syllables consists of, first, matching spectra in the steady-state vowel portion of the syllable and, second, tracking the parameters back to the consonant-vowel boundary. Since the articulatory configuration must change continuously and relatively slowly, the configuration that is appropriate to the last sample in the consonant should be very nearly the same as the configuration that is appropriate to the first sample in the vowel. The major difference between the two spectra results from a change in the location and spectral envelope of the source. Once the source

location and spectral envelope have been determined for the last consonantal sample, the parameters may be tracked back to the steady-state consonant portion of the syllable.

The program has been arranged so that initially the source location is set to be 2 cm anterior to the point of constriction. If necessary, changes can be made in the source location from the flexowriter.

To speed up the tracking process, matches were obtained over a frequency range that included only the first 3 formants for the vowel portion of the syllable. The range was extended, however, to include the first 7 formants for the vowel sample adjacent to the consonant-vowel boundary and for all of the consonant spectra. The results of this matching process over a portion of the syllable /ʃɛʃ/ are shown in Fig. XII-6.

The four curves at the right in Fig. XII-6 trace the values of the four parameters that are required in order to obtain matches, beginning with the middle of the initial consonant and ending in the middle of the vowel portion of the syllable. The parameters do appear to change in a relatively continuous manner as the matching process proceeds from the consonant to the vowel. In going from the fricative /ʃ/ to the vowel /ɛ/, the mouth section becomes somewhat shorter and much more open. The position of the constriction generally moves back toward the glottis and the amount of constriction decreases. A

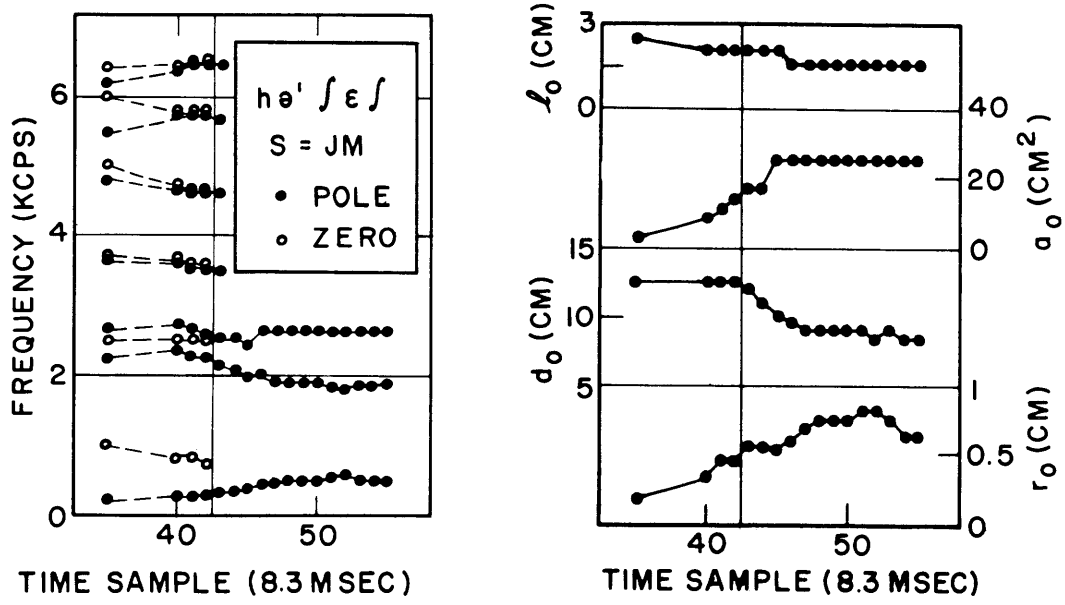

Fig. XII-6. Parameter patterns obtained from spectral matches over a portion of the initial /ʃ/ and /ɛ/ from the syllable /ʃɛʃ/ as produced by the speaker JM for: right, the four articulatory parameters r_o, d_o, a_o, and ℓ_o, and left, the resulting pole-zero parameters. The vertical line in the middle of each graph marks the location of the consonant-vowel boundary. The abscissas identify sample numbers indexed from the beginning of the utterance.

source location 2.5 cm anterior to the point of constriction was used throughout the fricative portion of the syllable.

The curves at the left in Fig. XII-6 trace the pole and zero locations for the vocal-tract transfer function corresponding to the values of the four parameters shown at the right. Beginning with the sample in the middle of the /ʃ/ portion of the syllable, each pole is generally cancelled by a neighboring zero, except for the poles that correspond to the third and sixth formants. There is also a "free" zero at 1000 cps. Near the vowel boundary, the fifth zero moves in such a way as to cancel the effect of the sixth formant also.

A spectral match for a sample in the steady-state portion of the consonant is displayed in Fig. XII-7. The two prominent peaks corresponding to the third and sixth

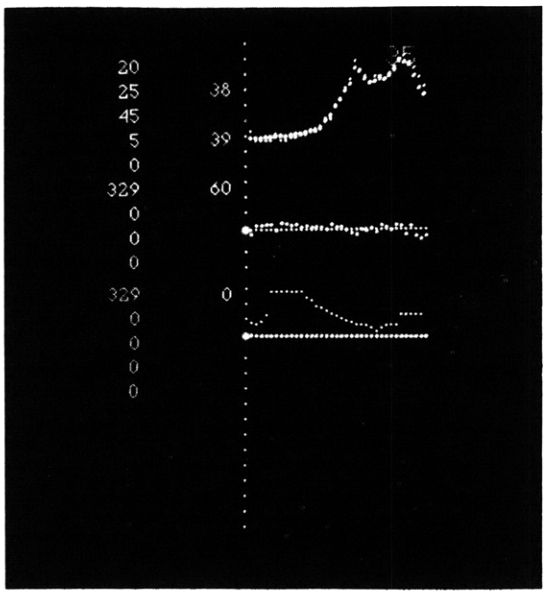

Fig. XII-7. Photograph of cathode-ray tube displays for the initial fricative consonant /ʃ/ from the syllable /ʃɛʃ/ depicting: top curves, comparisons of speech data (small dots) with corresponding calculated points (large dots), middle curve, differences between the speech data and the calculated points, and bottom curve, the corresponding vocal-tract configuration, the glottis being at the left and the lips at the right.

formants, as well as the effect of the free zero at 1000 cps, can be clearly seen in Fig. XII-7. The vocal tract has a rather narrow constriction well forward and a fairly long and narrow mouth opening. Spectral matches have also been obtained in a similar manner for the voiceless fricative consonant /f/ occurring in the syllable /fɛf/.

The results of this initial study are generally encouraging. It has been possible to

match a number of speech spectra in terms of articulatory parameters. Particularly interesting is the fact that spectra from consonant and consonant-vowel-transition portions of syllables were matched with little more difficulty than that required for matching vowel spectra. In each case the number of input parameters was the same and their values varied in a relatively continuous manner across consonant-vowel boundaries. Only a very limited amount of data was considered in this initial investigation, and at this time it seems improbable that reasonable values for the four parameters corresponding to spectra from all nonnasal classes of speech sounds can be obtained because of limitations in the four-parameter model. The results, however, indicate that the method has promise for an analysis procedure that makes it possible to study the dynamics of the articulatory system through analyses of the acoustic signal.

The material presented here is discussed in greater detail in the author's thesis submitted to the Department of Electrical Engineering, Massachusetts Institute of Technology, in partial fulfillment of the requirements for the degree of Doctor of Science, August 1962.

J. M. Heinz

References

1. J. M. Heinz, Reduction of speech spectra to descriptions in terms of vocal-tract area functions, Quarterly Progress Report No. 64, Research Laboratory of Electronics, M.I.T., January 15, 1962, pp. 198-203.

2. K. N. Stevens and A. S. House, Development of a quantitative description of vowel articulation, J. Acoust. Soc. Am. 27, 484 (1955).

3. K. N. Stevens and A. S. House, Studies of formant transitions using a vocal-tract analog, J. Acoust. Soc. Am. 28, 578 (1956).

4. 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 (1961).