XXII. SPEECH COMMUNICATION*

Prof. K. N. Stevens
Prof. M. Halle
Dr. A. S. House
Jane B. Arnold
M. H. L. Hecker
J. M. Heinz
S. Inomata†
A. R. Møller‡
A. P. Paul
T. Sakai**
T. T. Sandel
J. R. Sussex
E. C. Whitman

RESEARCH OBJECTIVES

The objectives of our work are to further our understanding of: (a) the process whereby human listeners decode an acoustic speech signal into a sequence of discrete linguistic symbols such as phonemes; and (b) the process whereby human talkers encode a sequence of discrete linguistic symbols into an acoustic signal.

Current research activities related to these objectives include experiments on the generation of speech by electrical analog speech synthesizers, development of means for controlling analog speech synthesizers by a digital computer, measurements of movements of the speech-generating structures during speech production, studies of methods of speech analysis, accumulation of data on the acoustic characteristics of utterances corresponding to phonemes in various linguistic contexts, and studies of the perception of speechlike sounds.

K. N. Stevens, M. Halle

A. PERFORMANCE OF THE ARTICULATORY ANALOG OF THE SPEECH MECHANISM: A REPORT ON THE STATUS OF RESEARCH

On other occasions Rosen[1,2] has described the various components of the dynamic transmission-line vocal-tract analog (familiarly, DAVO) in detail, and has also described experiments on vowel production and fricative-consonant production; Fujimura[3] has discussed stop-consonant production. More recently, Hecker[4] has reported on the development of the nasal components that have been added to the analog and has described experiments dealing with the production of nasal consonants. During the past several months the sound vocabulary of the synthesizer has been extended still further, and the results of the earlier studies have been incorporated into a general scheme for programming the analog.

The analog of the vocal tract and nasal cavities which is under discussion is realized by two electrical transmission lines that represent the acoustic pathways of the speech mechanism above the level of the vocal folds. Sources are available for appropriately

*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, Grant G-7364, and Grant G-16526); and in part by the National Institutes of Health (Grant MP-4737).

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

‡On leave from the Speech Transmission Laboratory, Department of Telegraphy-Telephony, Royal Institute of Technology, Stockholm, Sweden.

**Visitor from the University of Kyoto, Japan.

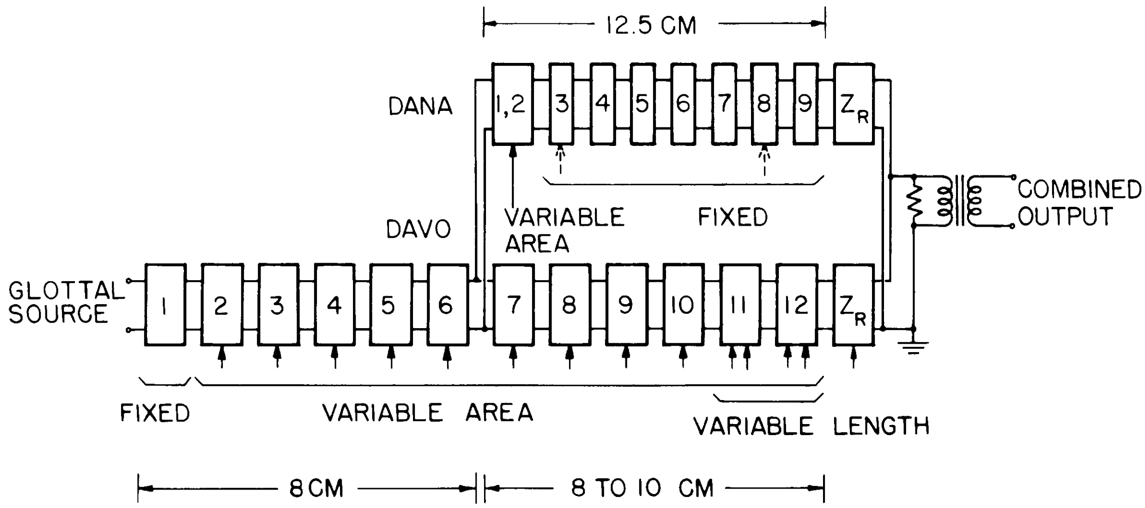


Fig. XXII-1. Sectional representation of the dynamic analog of the vocal tract (DAVO) and the dynamic analog of the nasal cavities (DANA), showing electrical interconnections. Solid arrows indicate manual control; Z_R designates radiation impedance.

exciting this electrical system; buzz excitation is used to simulate the glottal output, and noise excitation simulates the noise of turbulence. The transmission lines are approximated by lumped sections, each section having a series inductance and a shunt capacitance, as shown in block-diagram form in Fig. XXII-1.

The lower portion of the figure represents the analog developed by Rosen, and consists of 12 electrical sections, each simulating a portion, approximately 1.5 cm long, of the vocal tract. Section 12 terminates in a radiation impedance corresponding to the lip opening. The small arrows at the bottom of the sections indicate that the cross-section areas and lengths of certain sections can be varied by means of control voltages. An array of 14 voltages is needed to describe the geometric configuration of the vocal tract. Such arrays of control voltages are stored on potentiometer matrices.

The upper portion of Fig. XXII-1 represents the acoustic pathway of the nasal passages. Section 9 is terminated in a radiation impedance corresponding to the nostrils, and sections 1 and 2 provide for electronic control of the acoustic coupling between the nasal passages and the vocal tract at the level of the soft palate.

Through studies of the synthesis of speech by means of the articulatory analog of the speech system we hope to further our understanding of the processes by which human beings encode a sequence of discrete symbols such as phonetic symbols, into a continuous acoustic signal, and perform the converse. Such experiments can provide useful information on this human speech-generating process, since they bring us into contact with several levels of description of speech events. The synthesizer provides a fairly accurate representation of the acoustic system used in human speech production, and there is a relatively direct relation between this representation and the spectrographic representation or the speech waveform – a relation that is now fairly well understood. On the other hand, control of the configurations and excitations of the transmission-line analog is achieved, in effect, through simulation of what might be called phonetico-anatomical events and physiological speech events.

In this report we will try to indicate the nature of the rules that can be used to control the activities of DAVO when a sequence of input phonetic symbols is given. The rules themselves have not been specified for all possible sequences of symbols in a completely quantitative form, but a table of the primary features of the rules is largely complete for most consonants and vowels as they occur in monosyllabic CV and VC sequences. Little work has yet been done on the generation of syllables that contain consonant clusters. The rules are also subject to the restrictions inherent in the design of DAVO and its present control system. These rules require that the signals controlling the synthesizer be piecewise-linear in form, and that all sections of the vocal-tract analog change synchronously. The primary features of the rules can apparently be established in spite of these restrictions, but a detailed specification of secondary features must await the design of a more flexible control system for DAVO.

The rules must specify two types of things – first, they must describe the control of configurations of the vocal-tract analog as a function of time, and second, they must specify the temporal variations of the excitation and nasal coupling. We shall indicate the nature of these two aspects of the rules by describing the production of a simple consonant-vowel syllable by the analog synthesizer. In Fig. XXII-2 are shown tracings of radiographs which were taken in the midsagittal plane of the head during the production of two speech sounds. The upper tracing represents the position of the structures during the production of a postdental stop consonant; the lower one, a high front vowel.

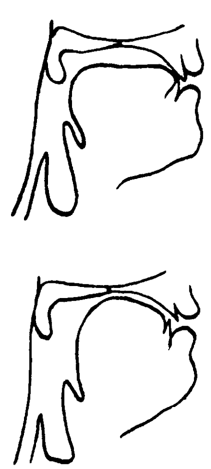


Fig. XXII-2. Tracings of midsagittal x-ray views of the vocal organs taken during the production of a postdental stop consonant (top) and a high front vowel (bottom).

It is necessary to derive from such tracings and other related anatomical data estimates of the area of the vocal tract as a function of position in the tract. Figure XXII-3 shows two area functions corresponding to postdental consonant and high front

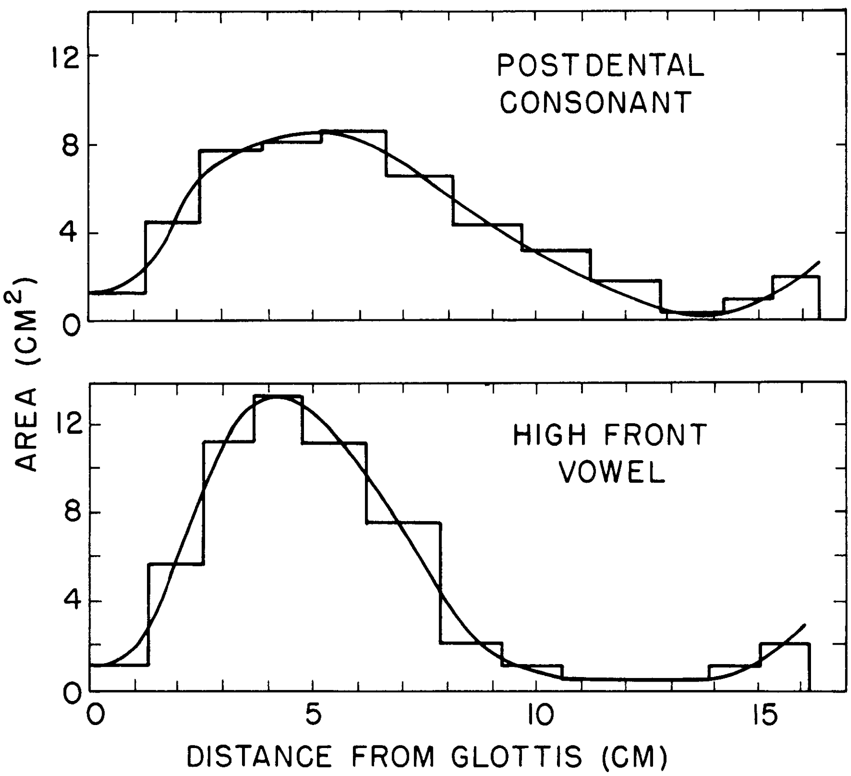


Fig. XXII-3. Area functions describing the target configuration of the analog synthesizer for the production of speech sounds as indicated.

vowel configurations. When DAVO is arranged to represent these configurations, measurements show that the resonant frequencies of the tract are appropriate to these two classes of sounds.

The production of a syllable consisting of a postdental consonant followed by a high front vowel presumably involves maneuvering the vocal tract from one of these configurations to the other. Furthermore, it can be postulated that such a maneuver is appropriate to the production of more than one syllable, since there is more than one consonant produced by a configuration such as the one labeled postdental in the figure.

The analog equipment allows the study of temporal variables through experiments in which timing instructions are generated by a central programming device. In essence, synchronized pulse trains control a set of function generators that provide transitions between two static states or conditions. In Fig. XXII-4, for example, the upper line presents a description in time of the maneuvering of the vocal tract from the postdental to the vowel configuration. The change begins at a point labeled zero time, and its duration is approximately 30 msec.

The synthesis of particular initial postdental consonants in the syllable is accomplished through appropriate manipulations of the timing of the excitations and of the

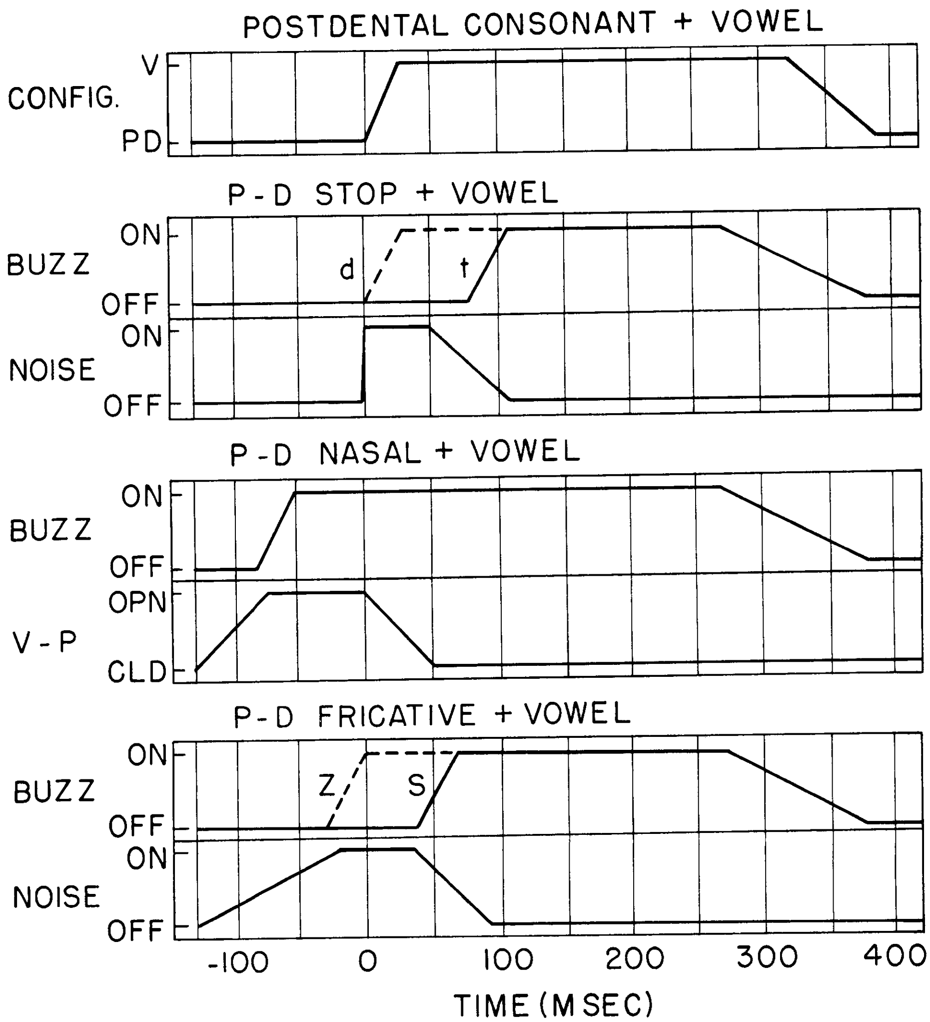


Fig. XXII-4. Curves describing the control parameters required for the production of syllables consisting of various postdental consonants followed by a vowel. The upper curve indicates the time course of the maneuver from the consonant to the vowel configuration. Subsequent curves demonstrate the conditions necessary for generating different postdental consonants /t d n s z/.

velopharyngeal coupling. A voiceless stop, for example, can be produced when there is no velopharyngeal coupling, the buzz excitation starts at approximately 80 msec, and noise excitation corresponding to frication and aspiration starts as the configuration begins to change, as shown in the second and third lines of the figure. The noise is inserted at a point in the vicinity of the vocal-tract constriction. If, in addition, the time course of the fundamental frequency of the buzz source is specified, the syllable will sound like /ti/.

The syllable can be changed to sound like a voiced stop consonant plus vowel, that

m			n	ŋ
p			t	k
b			d	g
			r l	
	f	θ	s ʃ	h
	v	ð	z ʒ	
ʍ				
w			j	

Fig. XXII-5. Tabular arrangements of phonetic symbols according to their manner of production (rows) and place of articulation (columns). The column to the left contains consonants produced with a constriction at the lips, and successive columns contain consonants whose articulation is characterized by progressively posterior locations of the point of maximum constriction.

is, /di/, if the buzz excitation is started 80 msec earlier, as shown in the second line of the figure, and all of the other control signals are kept the same. When the onset of the buzz is made still earlier, as in line 4 of the figure, the noise excitation is removed, the coupling to the nose follows the time course shown in line 5, and the syllable sounds like a nasal consonant plus vowel, /ni/. By the same kind of manipulation of timing functions of buzz and noise excitation exemplified at the bottom of the figure, voiceless and voiced fricative consonants can be produced.

The point of the discussion is that for the same configurations of the vocal-tract analog as a function of time, a variety of syllables can be generated by making simple changes in the temporal variations of the excitations and velopharyngeal coupling. Figure XXII-4 can be considered as a statement summarizing primary rules for the production of a set of related consonants in syllable initial position. These rules have already been stated in a qualitative manner in classical phonetic descriptions – they are, however, being quantified in our present research.

Figure XXII-5 is a summarizing statement of consonant production, from a phonetic point of view, and is arranged in a somewhat systematic way. The phonetic symbols are arranged in columns that specify places of major constriction in the vocal tract, with bilabial configurations to the left and velar to the right. The rows identify general manners of production, such as nasal, stop, and fricative. When voiceless and voiced cognates exist, they are arranged in adjacent rows. Such an arrangement implies that a small number of so-called target configurations corresponding to the various columns in the chart will suffice for consonant production. Figure XXII-4 and the listening experiences by which it is supported indicate, for example, that a single configuration is adequate for postdental consonants – the column headed by /n/ in this figure – and that the duration of the transition from the consonant configuration to that of the adjacent vowel is roughly independent of whether the consonant is nasal, stop, or fricative. Along any row of the chart, the rules describing temporal relations among the control signals for the synthesizer, particularly those involving buzz and noise excitation and velopharyngeal coupling, will be similar, as Hecker[4] has demonstrated for the nasal consonants on the top row of this figure. The durations of the transitions to and from the consonant configuration depend, however, upon the place of production of the consonant, being longer for velar than for bilabial and postdental configurations. From one row to another the timing functions differ in relatively simple ways. Stops and fricatives, for example, differ primarily in the onset rates of excitation; cognate consonants differ primarily in duration of buzz excitation and in velopharyngeal coupling.

We have been working on the specification of the rules that describe the configurations and timing functions required for the generation of all of the sounds indicated in Fig. XXII-5. Target configurations that are reasonable approximations to those used in producing speech have been established on the basis of radiographic data, acoustic analyses, and listening tests. To derive the timing functions we have performed both formal experiments and carried out many informal investigations, limited, of course, by the characteristics of the programming device at our disposal. Within this general frame of reference we have prepared a demonstration of the present capabilities of DAVO and its operators for producing syllables involving a variety of consonants and vowels.

The demonstration consists of a short song and two short sentences. The song is the well-known one by which young children learn the alphabet – it consists of the names of the 26 alphabet letters followed by the refrain, "Oh how happy we shall be, when we know our ABC." These materials include all of the consonant phonemes of American English except /θ ð ŋ ɡ ʒ/. The first two missing phonemes are included in the sentence materials; the third and fourth have been demonstrated by Hecker[4] and Fujimura,[3] respectively; the last has been produced informally.

The use of the song sidesteps the problem of dealing with prosodic features, since the demonstration tape was produced by superimposing on the rules we have been proposing the information provided by a simple musical score – that is, the stress, duration, pitch, etc., follow the score. The production was simplified still further by using an infantile singing style, allowing for separation of syllables, discontinuous vocalization, and some "pitch wobble" within the syllable.

The second part of the demonstration deals with connected speech. With the present programming equipment the production of such materials is rather laborious because it is necessary to synthesize syllables with appropriate transitions, store them on tape, and splice such syllables together. The specification of prosodic features – intonational patterns, stress, vowel durations, and so on – was guided by studying spectrograms of utterances generated by a human talker. The materials prepared are as follows: "This is the voice of DAVO at M.I.T. Tech is Hell!"

M. H. L. Hecker, A. S. House, K. N. Stevens

References

1. G. Rosen, Dynamic analog speech synthesizer, J. Acoust. Soc. Am. 30, 201-209 (1958).

2. G. Rosen, Dynamic Analog Speech Synthesizer, Technical Report 353, Research Laboratory of Electronics, M.I.T., February 10, 1960.

3. O. Fujimura, Some synthesis experiments on stop consonants in the initial position, Quarterly Progress Report No. 61, Research Laboratory of Electronics, M.I.T., April 15, 1961, pp. 153-162.

4. M. H. L. Hecker, Construction and Evaluation of a Dynamic Analog of the Nasal Cavities, S.M. Thesis, Department of Electrical Engineering, M.I.T., May 20, 1961.

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

The analysis-by-synthesis technique as applied to the analysis of speech spectra requires a model that is capable of generating speech spectra when the values of certain parameters that are descriptive of the speech production are given.[1] The analysis procedure consists of varying the input parameters of the model until an output spectrum is obtained which optimally matches the spectrum that is to be analyzed. The values of the parameters which are needed to give this best match constitute the results of the analysis.

Models that have been utilized previously have as their input parameters the locations in the complex-frequency plane of the poles and zeros of the vocal-tract transfer function. Also, some information is included to describe the source characteristics. For a particular type of source, however, these characteristics are generally found to vary only slightly from person to person, and from time to time for a particular person. The result of an analysis with this model, then, is a description of a speech spectrum in terms of the poles and zeros of the vocal tract which produced that spectrum. The poles of the transfer function are independent of the source location and depend only on the vocal-tract configuration.[2] Thus, the poles move in a continuous manner during the production of connected speech, even though the source may change in type and location. One difficulty with models of this type arises from the fact that the zeros do depend on source location, as well as on the vocal-tract shape, and so may appear and disappear and, in general, change in a discontinuous fashion as different types of sources are introduced at different locations within the tract. This discontinuous behavior of the zeros and the resulting drastic changes in the general characteristics of the speech spectrum make the tracking through connected speech of the parameters involved in this type of model very difficult.

In order to investigate the possibility of an analysis in which the parameters employed are more closely related to the physical processes involved in speech production, a new model is being studied. In this model the input parameters are the cross-section areas of the vocal tract at a specified number of positions along the tract. Changes in these area parameters are closely related to changes in the physical dimensions of the vocal tract and therefore vary in a continuous way. Thus, it is hoped that these area parameters, or other physical parameters simply related to them, may prove useful in an analysis-by-synthesis scheme for speech analysis. The rest of this discussion is directed toward the problem of describing a model for generating speech spectra when vocal-tract area functions and the source locations are known.

The calculation of a spectrum takes place in two steps, although ultimately these steps will be combined into one. First, the locations of the poles and zeros of the transfer function are determined from the area parameters, and, second, the spectrum itself is determined on the basis of the previous model.[1] Since the distribution of acoustic losses within the vocal tract is not known in detail (and is of secondary importance here), an initial version of this model provides only the imaginary-axis coordinates (center frequencies) of the poles and zeros. Thus the real-axis coordinates (half-bandwidths) also constitute input parameters for this version of the model.

Webster's horn equation expressing the approximate behavior of acoustic volume velocity $U(x,t)$ in a nonuniform tube in terms of the cross-section area of the tube $A(x)$ is taken as an acoustic description of the vocal tract,[3-6] and is given by

$$A \frac{\partial}{\partial x}\left(\frac{1}{A}\frac{\partial U}{\partial x}\right) - \left(\frac{1}{c}\right)^2 \frac{\partial^2 U}{\partial t^2} = 0. \tag{1}$$

This expression is correct when only plane waves propagate within the tube, but it gives a good approximation in many cases in which the waves are not plane but may be considered one-dimensional.[7, 8] For exponential time dependence of complex frequency $s = \sigma + j\omega$, Eq. 1 reduces to an equation involving the complex volume velocity $U(x,s)$.

$$A \frac{d}{dx}\left(\frac{1}{A}\frac{dU}{dx}\right) - \left(\frac{s}{c}\right)^2 U = 0. \tag{2}$$

The appropriate homogeneous (source-free) boundary conditions represent: (a) an assumption of infinite glottal impedance, and (b) a constraint on the ratio of complex sound pressure $p(x, s)$ to complex volume velocity $U(x, s)$, imposed by the radiation mass loading at the mouth. Thus

(a) at $x = 0$ (glottis), $U = 0$

(b) at $x = \ell$ (mouth), $p/U = (M_{sr}/A) \cdot s$

where M_{sr} is the specific acoustic radiation mass at the mouth. The radiation mass is considered to be very nearly the radiation mass appropriate to a piston in an infinite baffle. The second boundary condition may be expressed alternatively in terms of the volume velocity and its derivative by introducing an expression for the acoustic compressibility of a gas.[7]

$$p = -\frac{\rho_o c^2}{sA}\frac{dU}{dx} \tag{3}$$

ρ_o = the ambient density of the gas.

The second boundary condition then becomes

(b') $\frac{dU}{dx} + \left(\frac{s}{c}\right)^2 \left(\frac{M_{sr}}{\rho_o}\right) U = 0.$

Since no sources are present and the boundary conditions are homogeneous, Eq. 2 can have nontrivial solutions for only certain values of s, the natural frequencies of the tube. The natural frequencies of the vocal tract are, of course, the locations in the complex frequency plane of the desired poles of the vocal-tract transfer function. Since losses are not included in these equations, all such characteristic values of s must be imaginary, so that the poles obtained all lie on the $j\omega$-axis.

For numerical solution, the vocal tract of length ℓ is divided into N equal intervals, each of length h. If $U(x=nh)$ is denoted by U_n and $A(x=nh)$ by A_n, the differential equation (2) can be represented approximately by the difference equation

$$U_{n-1} = \frac{A_n}{A_{n+1}}\{U_n - U_{n+1}\} + \left\{1 + \left(\frac{sh}{c}\right)^2\right\} U_n \tag{4}$$

and the boundary conditions become

$$U_o = 0$$

and

$$U_{N-1} = U_N \left\{ 1 + \left(\frac{sh}{c}\right)^2 \frac{M_{sr}}{h\rho_o} \right\}.$$

A program for the TX-0 digital computer has been written to solve Eq. 4. A trial-and-error procedure is used in which, for a trial value of s, U_N is chosen arbitrarily, U_{N-1} is calculated from the second boundary condition, and U_o is obtained by an iteration of Eq. 4. In general, U_o will not be zero unless the trial value of s is a characteristic value for the particular vocal-tract configuration that is being considered. The program is continued by trying different values of s in a systematic manner until all of the values of s which are smaller than some preset value are found for which the resulting U_o is zero. These characteristic values for s are then the desired lossless pole locations.

The present version of the program can handle up to 40 intervals of length as small as 0.5 cm. Figure XXII-6 shows a cathode-ray tube display of results of solutions provided by the preliminary program for three different values of S. The vocal-tract configuration is one appropriate to the vowel /i/. A curve depicting the diameter of the vocal tract at 33 positions along its 16-cm length is shown in Fig. XXII-6. Figure XXII-7 shows distributions of volume velocity along the vocal tract for three different trial values of complex frequency s. These three values correspond to: (a) the first-formant frequency; (b) a frequency between the first and second formant; and (c) the second-formant frequency.

With the present program we locate only the poles of the vocal-tract transfer function, and so it is useful primarily for sources located at the glottis. With a slight modification, however, the program can also be used to locate the zeros of a transfer function for a pressure-difference source (a pressure source "in series" with the vocal tract) located at a distance ℓ_1 from the glottis. Zeros in the output spectrum for such a source occur at frequencies that are "open-circuit" natural frequencies of that portion of the vocal tract lying between the glottis and the source. These are the frequencies at which the impedance, looking back from the source toward the glottis, has poles.[9, 10] Since the source is in series with this impedance, the output at these frequencies is zero. If $\ell_1 = N_s h$, the boundary conditions corresponding to open-circuit conditions are: $U_o = 0$, and $U_{N_s} = 0$. Thus, the present program can be used with these boundary conditions to find the open-circuit natural frequencies of the back portion of the vocal tract and hence the zeros of the transfer function.

A final version of the vocal-tract spectrum-generating program, which is now being prepared, includes provisions for sources located within the vocal tract, as well as at the glottis. With this program it will be possible to investigate the problem of converging as fast as possible to solutions for the natural frequencies of the vocal tract.

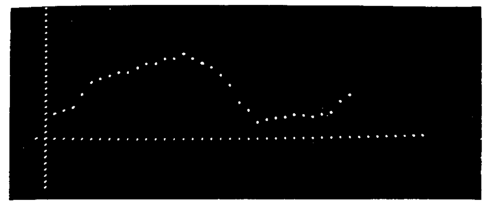
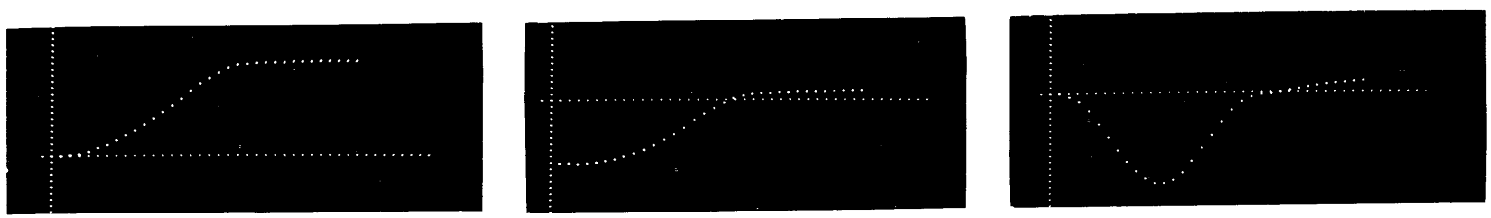


Fig. XXII-6. Photograph of a TX-0 cathode-ray tube display depicting effective vocal-tract diameter (ordinate) versus distance along a 16-cm tract (abscissa) with points plotted every 0.5 cm. The glottis is represented at the left and the mouth at the right of the curve. The configuration was derived from unpublished data for the vowel /i/ (R. W. Wendahl, Ph.D. Thesis, University of Iowa, 1957).



(a) (b) (c)

Fig. XXII-7. Photographs of TX-0 cathode-ray tube display depicting volume-velocity distribution along a vocal tract with the diameter function shown in Fig. XXII-6 for: (a) the first-formant frequency; (b) a frequency between the first- and second-formant frequencies; and (c) the second-formant frequency. The numbers at the right in each case, starting from the top, indicate the frequency, the interval size (here, 2^{-1} cm), and the total number of intervals. Note that the volume velocity in (a) and (c) (corresponding to formant frequencies) satisfies the boundary condition $U_o = 0$, but in (b) it does not.

An investigation of the usefulness of this type of spectrum generator in an analysis-by-synthesis scheme for speech analysis will then be initiated.

J. M. Heinz

References

1. 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. (in press).

2. C. G. M. Fant, Transmission Properties of the Vocal Tract with Application to the Acoustic Specification of Phonemes, Technical Report No. 12, Acoustics Laboratory, M.I.T., January 1952.

3. A. G. Webster, Acoustical impedance and the theory of horns and the phonograph, Proc. Natl. Acad. Sci. U.S. 5, 275 (1919).

4. T. Chiba and M. Kajiyama, The Vowel: Its Nature and Structure (Tokyo-Kaiseikan Publishing Company, Ltd., Tokyo, Japan, 1941).

5. E. S. Weibel, Vowel synthesis by means of resonant circuits, J. Acoust. Soc. Am. 27, 858 (1955).

6. W. Meyer-Eppler and G. Ungeheurer, Die Vokalartikulation als Eigenwertproblem, Z. Phonetik 10, 245 (1957).

7. P. M. Morse, Vibration and Sound (McGraw-Hill Book Company, Inc., New York, 1948).

8. E. S. Weibel, On Webster's horn equation, J. Acoust. Soc. Am. 27, 726 (1955).

9. G. Fant, Acoustic Theory of Speech Production (Mouton and Company, The Hague, 1960).

10. J. M. Heinz and K. N. Stevens, On the properties of voiceless fricative consonants, J. Acoust. Soc. Am. 33, 589-596 (1961).