An articulatory synthesizer for perceptual research[a)]

Philip Rubin and Thomas Baer

Haskins Laboratories, 270 Crown Street, New Haven, Connecticut 06510

Paul Mermelstein

Bell-Northern Research and INRS Telecommunications, University of Quebec, Verdun, Quebec, Canada H3E 1H7

(Received 15 March 1979; accepted for publication 12 May 1981)

A software articulatory synthesizer, based upon a model developed by P. Mermelstein [J. Acoust. Soc. Am. 53, 1070–1082 (1973)], has been implemented on a laboratory computer. The synthesizer is designed as a tool for studying the linguistically and perceptually significant aspects of articulatory events. A prominent feature of this system is that it easily permits modification of a limited set of key parameters that control the positions of the major articulators: the lips, jaw, tongue body, tongue tip, velum, and hyoid bone. Time-varying control over vocal-tract shape and nasal coupling is possible by a straightforward procedure that is similar to key-frame animation: critical vocal-tract configurations are specified, along with excitation and timing information. Articulation then proceeds along a directed path between these key frames within the time script specified by the user. Such a procedure permits a sufficiently fine degree of control over articulator positions and movements. The organization of this system and its present and future applications are discussed.

PACS numbers: 43.70.Bk, 43.70.Jt, 43.70.Qa, 43.60.Qv

INTRODUCTION

Articulatory synthesis has not usually been considered to be a research tool for studies of speech perception, although many perceptual studies using synthetic stimuli are based upon articulatory premises. In order to address the need for a research articulatory synthesizer, this paper provides a brief description of a software articulatory synthesizer, implemented on a laboratory computer at Haskins Laboratories, that is presently being used for a variety of experiments designed to explore the relationships between perception and production. A related paper, by Abramson *et al.* (1979), describes in greater detail some of these experiments, which focus on aspects of velar control and the distinction between oral stop consonants and their nasal counterparts. The intent of the present paper is to provide an overview of the actual design and operation of this synthesizer, with specific regard to its use as a research tool for the perceptual evaluation of articulatory gestures.

Briefly, the articulatory synthesizer embodies several submodels. At its heart are simple models of six key articulators. The positions of these articulators determine the outline of the vocal tract in the midsagittal plane. From this outline the width function and, subsequently, the area function of the vocal tract are determined. Source information is specified at the acoustic, rather than articulatory, level, and is independent of the articulatory model. Speech output during each frame is obtained after first calculating, for a particular vocal-tract shape, the acoustic transfer function for both the glottal and fricative sources, if they are used. For voiced sounds the transfer function accounts for both the oral and nasal branches of the vocal tract. Continuous speech is obtained by a technique similar to key-frame animation (see below).

Although the synthesizer is capable of producing short segments of quite natural speech with a parsimonious input specification, its primary application is not based on these characteristics. The most important aspect of its design is that the articulatory model, though simple, captures the essential ingredients of real articulation.[1] Thus, synthetic speech may be produced in which articulator positions or the relative timing of articulatory gestures are precisely and systematically controlled, and the resulting acoustic output may be subjected to both analytical and perceptual analyses. Real talkers cannot, in general, produce utterances with systematic variations of an isolated articulatory variable. Further, for at least some articulatory variables—for example, velar elevation and the corresponding degree of velar port opening—simple variations in the articulatory parameter produce complex acoustic effects. Thus, the synthesizer can be used to perform investigations that are not possible with real talkers, or investigations that are difficult, at best, using acoustic synthesis techniques.

I. THE MODEL

The model that we are using was originally developed by Mermelstein (1973) and is designed to permit simple control over a selected set of key articulatory parameters. The model is similar in many respects to that of Coker (Coker and Fujimura, 1966; Coker, 1976), but was developed independently. The two models emphasize different aspects of the speech production process; Coker's implementation stresses synthesis-by-rule, while the focus of the present model is on interactive and systematic control of supraglottal articulatory configurations and on the acoustic and perceptual consequences of this control. The particular set of parameters employed here provides for an adequate description of the vocal-tract shape, while also incorporating both individual control over articulators and physiologically constrained interaction between articulators. Figure 1 shows a midsagittal section of the

[a)]Portions of this paper were presented at the 95th Meeting of the Acoustical Society of America, 16–19 May 1978, Providence, R. I.

 0001-4966/81/080321-08$00.80

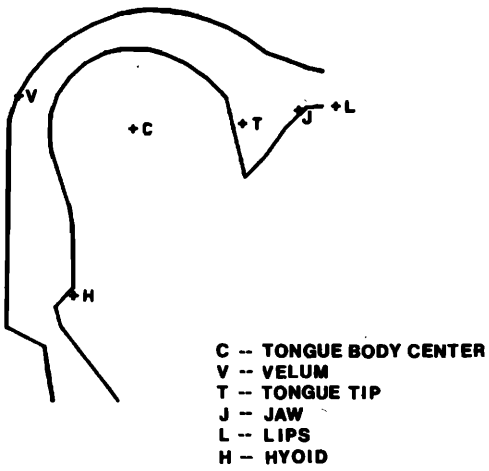


FIG. 1. Key vocal tract parameters.

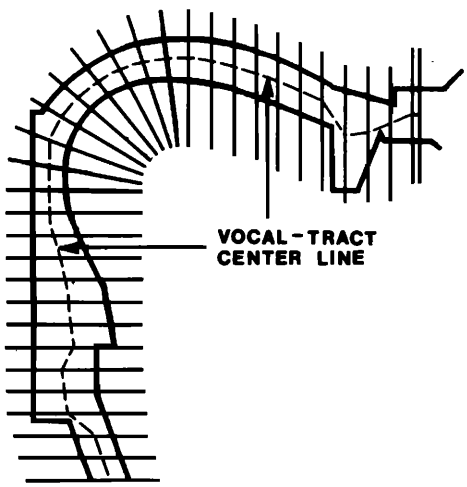


FIG. 2. Grid system for conversion of midsagittal shape to cross-sectional area values.

vocal tract, with the six key articulators labeled: tongue body, velum, tongue tip, jaw, lips, and hyoid bone position. (Hyoid bone position controls larynx height and pharynx width.) These articulators can be grouped into two major categories: those whose movement is independent of other articulators (the jaw, velum, and hyoid bone); and articulators whose positions are functions of the positions of other articulators (the tongue body, tongue tip, and lips). The articulators of this second group all move relative to the jaw. In addition, the tongue tip moves relative to the tongue body. In this manner, individual gestures can be separated into components arising from the movement of several articulators. For example, the lip-opening gesture in the production of a /ba/ is a function of the movement of two articulators: the opening of the lips themselves, and the dropping of the jaw for the vowel articulation. Movements of the jaw and velum have one degree of freedom, all other articulators move with two degrees of freedom. Movement of the velum has two effects: it alters the shape of the oral branch of the vocal tract, and in addition, it modulates the size of the coupling port to the fixed nasal tract. A more detailed description of the models for each of these articulators is provided in Mermelstein (1973). Specification of the six key articulator positions completely determines the outline of the vocal tract in the midsagittal plane.

Validation of the articulatory model involved systematic comparisons with data derived from x-ray observations of natural utterances. The data used were obtained from Perkell (1969). An interactive procedure was used in which the experimenter matched the model-generated outline to x-ray tracings for the utterance, "Why did Ken set the soggy net on top of his desk?" Acoustic matching procedures are detailed in Mermelstein (1973). Further aspects of the model (described below) were refined by comparisons of the final synthesizer output with the accompanying acoustic recordings for these utterances. This articulatory model was designed to provide a restricted, manageable set of control parameters, while retaining physiological credibility. The structure of the synthesizer additionally incorporates the ability to explore the possibility of modified or supplemental control parameters. For example, additional parameters suggested by factor analyses of tongue shapes (Harshman *et al.*, 1977; Kiritani and Sekimoto, 1977; Maeda, 1979), may be included.

Once articulator positions and, thus, the vocal-tract outline, have been specified, cross-sectional areas are calculated by superposing a grid structure fixed to the maxilla on the vocal-tract outline and computing the points of intersection of the outline and the grid lines (see Fig. 2). Grid resolution is variable within a limited range. In the default mode, grid lines are 0.25 cm apart where parallel and at 5° intervals where they are radial. Figure 2 shows the lowest resolution case: parallel grid lines are separated by 0.5 cm and the radial sections are at 10° intervals. The intersection of the vocal-tract center line with each grid line is determined by estimating the point at which the distances to the superior–posterior and inferior–anterior outlines are equal. The center line of the vocal tract is formed by connecting the points from adjacent grid lines. The length of this line represents the length of the vocal tract modeled as a sequence of acoustic tubes of piecewise uniform cross-sectional areas. Vocal-tract widths are measured using the same algorithm as that for finding the center line.

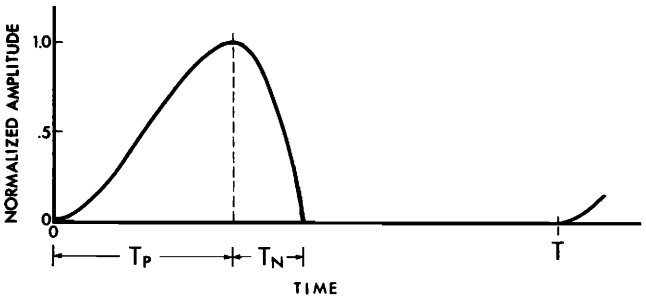
Sagittal cross-dimensions are converted to cross-sectional areas with the aid of previously published data regarding the shape of the tract along its length. Different formulas are used for the pharyngeal region (Heinz and Stevens, 1964), oral region (Ladefoged *et al.*, 1971), and labial region (Mermelstein *et al.*, 1971). More specifically, cross-sectional area in the pharyngeal region is approximated as an ellipse with the vocal-tract width (w) as one axis, while the other axis increases from 1.5 to 3 cm. Cross-sectional area in the soft-palate region is given by $2w^{1.5}$, in the hard-palate region by $1.6w^{1.5}$, and between the alveolar ridge and incisors by $1.5w$ for $w < 0.5$, $0.75 + 3(w - 0.5)$ for $0.5 < w < 2$, and $5.25 + 5(w - 2)$ for $w > 2$. The cross sections are assumed to be elliptical in the labial region, with horizontal width in centimeters given by $2 + 1.5(s_1 - p_1)$, where p_1 is lip protrusion and s_1 is the vertical lip separation. The area function is then

smoothed and approximated by a sequence of uniform tubes of fixed section length (0.875 cm). To determine the cross-sectional area of the last section (nearest the lips), the tract is extended as a parabolic horn continuous in area with the computed cross-sectional area at the lips. This continuation allows truncation of the tract at a length value that is an integral multiple of the section length, and its termination by the appropriate acoustic impedance. Movements of the articulators, especially hyoid height and lip protrusion, affect the overall length of the vocal tract, resulting in a variable number of vocal-tract sections.

After the area function has been determined and approximated by an equivalent series of uniform-tube sections, the acoustic transfer function is calculated. Briefly, the theory for this calculation is based on the Kelly–Lochbaum model (1962), which provides for frequency-independent propagation losses within sections and reflections at section boundaries. (See Appendix A for a description of the detailed mathematical theory for this calculation.) Nonideal terminations at the glottis, lips, and nostrils are accurately represented. However, the effects of other nonideal conditions, such as yielding vocal-tract walls, are accounted for only phenomenologically—by incorporating lumped-parameter elements within the nasal section and at the glottis. The values of these elements are determined empirically to optimize the model's fit to actual acoustic data. For example, adjustments to the glottal inductance are capable of simulating the soft-wall effects in raising *F*1 for most articulatory configurations. The accuracy of these effective parameters clearly does not rival that of models which account more explicitly for the propagation of sound along the tract (e.g., Flanagan *et al.*, 1975). However, these more realistic techniques are too computationally complex to be practical for use in an interactive research synthesizer, given existing technology. Appendix B provides a more detailed description of the nonideal termination and loss parameters.

Acoustic excitation of the vocal-tract transfer function is specified in the form of an acoustic waveform, rather than being simulated in terms of aerodynamic and physiological variables. Thus, acoustic output of the vocal tract is obtained by supplying the acoustic excitation (glottal or fricative) as input to the appropriate acoustic transfer function implemented as a digital filter. This procedure has been adopted largely in the interest of computational speed and efficiency, but also reflects our focus, for the purposes of this device, on vocal-tract anatomy and dynamic behavior, rather than laryngeal physiology and aerodynamic effects.

Control of the voice source involves specification of amplitude, fundamental period, and two additional parameters affecting the shape of the glottal volume velocity pulse. Normally, these two parameters specify open quotient, the ratio of pulse duration to pitch period, and speed quotient, the ratio of the rising to falling pulse durations (Timcke *et al.*, 1958). The specific form of this time-domain description uses different polynomial functions for the rising and falling phases,



POSITIVE SLOPE

$0 \le t \le T_P$

$3\left(\frac{t}{T_P}\right)^2 - 2\left(\frac{t}{T_P}\right)^3$

NEGATIVE SLOPE

$T_P \le t \le T_P + T_N$

$1-\left(\frac{t-T_P}{T_N}\right)^2$

$$\text{OPEN QUOTIENT} = \left(\frac{T_P + T_N}{T}\right)$$

$$\text{SPEED QUOTIENT} = \frac{T_P}{T_N}$$

FIG. 3. Glottal pulse shape.

and includes a single slope discontinuity at the end of the open phase (see Fig. 3). This form was previously found to be most natural in perceptual tests (Rosenberg, 1971). Optionally, the glottal pulse can be specified in the frequency domain as the impulse response of a two-pole filter. In this case, the two input parameters specify the pole frequencies. However, the time domain description is preferred because its input parameters seem more natural and its output produces perceptually better results.

For fricative excitation, the amplitude and place of insertion of pseudorandom noise must be specified. Optionally, the location of the noise source can be automatically set anterior to the tube section with minimum area. Direct control of frication amplitude is admittedly unnatural, and it has also proven difficult to use in practice. Automatic control of the onset and amplitude of frication in terms of aerodynamic parameters would clearly be preferable. Such control of frication has been reported by Flanagan and his coworkers (Flanagan *et al.*, 1975; Flanagan and Ishizaka, 1976), using techniques which are presently too computation bound for use in a research synthesizer. However, a simplification of these methods could be developed for automatic control of frication by generalizing the glottal source to account for dc flow, perhaps by specifying the glottal waveform as an area function and also introducing subglottal pressure as a control variable. Turbulence amplitude and internal source resistance at a constriction could then be calculated from flow velocity and cross dimension at the constriction.

II. PROGRAM ORGANIZATION

Figure 4 is a block diagram of the conceptual organization of the software for the articulatory synthesizer, as implemented on a DEC PDP-11/45 minicomputer. Input to the synthesizer is in the form of a list of posi-

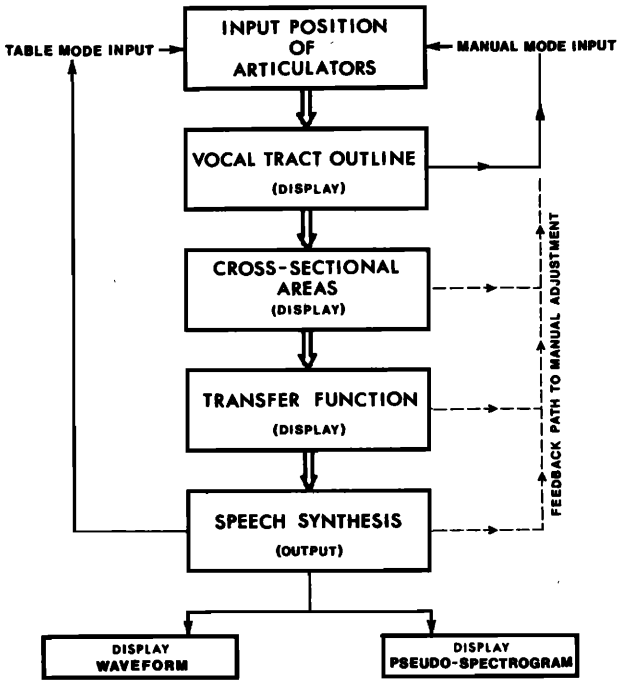


FIG. 4. Steps in articulatory synthesis.

tions of the key articulators that is arrived at in one of two ways. In what is called the *manual mode* of operation, input is derived from the static vocal-tract configuration that results from manual manipulation of the articulators described above—a form of synthesis-by-art. Alternatively, input can be read from a previously stored table of values—synthesis-by-script. This second procedure, called the *table mode* of operation, will be discussed in more detail in the next section. (A third input mode, not shown in Fig. 4, can also be used. In this mode an array of cross-sectional area values is specified as input, and earlier stages of the synthesis process are bypassed.)

A. Control by graphical modification

Figure 1 shows the graphical display provided by the synthesizer system that permits the user to simply modify the midsagittal vocal-tract shape. The user selects one of the indicated six key parameters, moves a set of cross-hair cursors to specify its new position, and the resultant vocal-tract outline is immediately calculated and displayed on the graphics terminal. After the positions of the key articulators have been provided as input, the program fleshes out this framework as a midsagittal section of the vocal tract, as was seen in Fig. 1, and displays this shape. Next, the width function and corresponding area function are calculated (as described above). When the determination of area values is completed, the vocal-tract transfer function is computed (Mermelstein, 1971, 1972) and displayed. Speech output is then generated, at a sampling rate of 20 kHz, by feeding the source waveform through the digital filter representation of the transfer function.

The resulting speech signal is obtained by concatenating the responses to individual pitch pulses of varying durations. Acoustic energy is usually propagated between pitch pulses, but may optionally be set to zero at the onset of every pulse. Output is generally produced within 20–60 times real time, which permits the kind of interactive use necessary for hypothesis-and-test research. Further, in the manual mode of control the user can obtain feedback at a number of different stages, in the form of displays of the vocal-tract outline, the cross-sectional area array and the acoustic transfer function. These varying forms of feedback information are extremely useful for providing the user with complementary descriptions of the particular articulatory shape being examined. In addition, they provide him with the opportunity to return to the manual adjustment stage if changes in the articulatory configuration are required.

B. Synthesis-by-script

The articulatory synthesizer, in its manual mode, provides an excellent means for examining vowel quality as a result of the excitation of the static vocal-tract shape. The use of the synthesizer in this mode, however, does not allow for the dynamic simulation necessary to model actual continuous speech. Therefore, another procedure has been implemented to provide for time-varying control over the movements of the articulators. The approach used is similar to *key-frame animation*: the framework for a desired dynamic articulation is represented by a series of configurations of the vocal tract. The actual path of articulation is obtained by interpolating between these key frames. In essence, the user provides the synthesizer with a *script*, in the form of a table of values, for the complete articulation. Each line of the script consists of a "snapshot" of the vocal tract at some point in time. The exact form of an articulation-over-time is then determined by linearly interpolating the articulatory parameters and computing the corresponding sequence of vocal-tract shapes. Movements of the vocal tract are simulated using a quasistatic approximation. The positions of the articulators are determined, or specified, at the onset of every pitch period and the corresponding acoustic transfer function is computed.

An example of this procedure can be seen in Fig. 5 for the synthesis of the utterance /da/. There are two "key" vocal-tract shapes specified as input. The first shape is an articulation appropriate for the onset of the production of the /da/, with the tongue tip occluding the vocal tract at the alveolar ridge. The vocal-tract shape appropriate for an /a/ is the second key configuration. The script, then, for this production begins with the first key shape specified at the onset of the utterance and at time 50 (ms), which permits a period of prerelease voicing. Release occurs at 50 ms, and all movement is completed by 120 ms, at which time the second key configuration is achieved. In this 70-ms period of rapid movement, a number of different intermediate shapes are calculated by linear interpolation, of which two (at 75 and 100 ms) are indicated in the figure. The production of this syllable continues for another 255 ms as specified by the time of the final /a/ configuration in the input. The additional

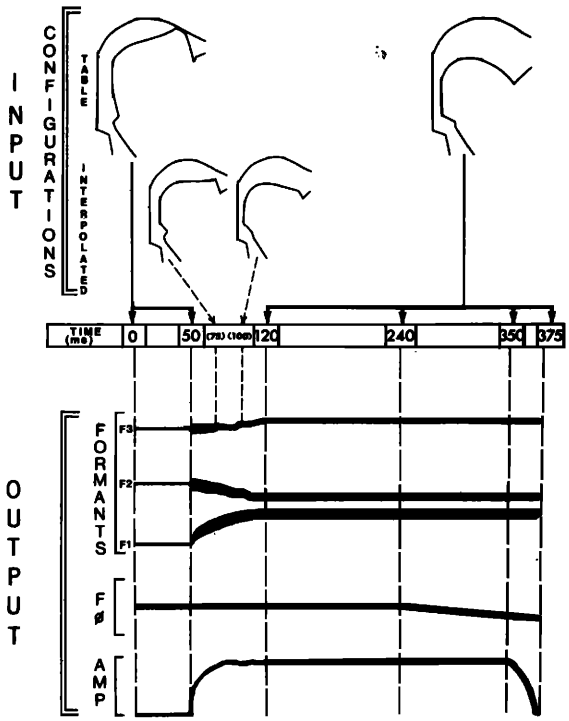


FIG. 5. Articulatory synthesis of /da/.

specifications of this shape are necessary to indicate the changes in the excitation parameters, such as a falloff in fundamental frequency towards the final third of the utterance (beginning at 240 ms), and a rapid decrease in the amplitude in the final 25 ms. The bottom half of Fig. 5 shows the output generated from this articulation script in the form of a stylized spectrogram representing the first three formants, and the time-synchronized plots of fundamental frequency and overall output amplitude.

This straightforward procedure affords the user a flexible means of approximating productions observed in actual speech. The movements of the articulators are controlled by directing their paths from shape to shape, with critical configurations serving as guides along the way. This allows for a simple specification of input information by the user. For example, the articulation for a /da/ could be represented, without refinements for naturalness, in the form of a script consisting merely of the two key vocal-tract shapes. Changes and comparisons between related articulations are easily accomplished. To produce a /na/ one can use the /da/ articulation previously described with the single modification of opening the velum to allow for the required amount of nasalization. A series along the continuum from /da/ to /na/ can be created, then, by using ordered steps of velar opening, from a completely closed velum to one open to the degree desired for an acceptable /na/. Further, changes in timing relationships are also simply accomplished by varying the time required to move between key configurations. Once an utterance has been completely generated, the speech signal produced can be played out, be stored on a disk, or be examined in more detail in a waveform editor program that is linked to the synthesizer. If necessary, the signal can be compared with previously produced utterances or edited in a variety of ways. Also available as final output is a stylized spectrographic display, which serves as a summary statement of information about formant frequencies and their bandwidths, amplitude, and fundamental frequency. In addition, an animated version of the synthesizer can be used for observing the dynamics of the overall articulatory sequence.

III. CONCLUSION

The design and implementation of the articulatory synthesizer is intended, as previously noted, to provide researchers with a flexible interactive tool for examining relationships between speech perception and production. Input parameters to the synthesizer are the positions of a limited set of major articulators and excitation and timing information. An important aspect of this model's design is that speech sounds can be generated using controlled variations in timing or position parameters, and used in formal perceptual tests. Another important aspect is that the synthesis procedure is fast enough to make on-line interactive research practical.

One present application of the synthesizer is an investigation of detailed relationships between velar control and the perceptual oral–nasal distinction. Here, an important attribute of the synthesizer is its ability to produce complex variations in the acoustic output from a simple, and natural, variation of a single articulatory parameter—as contrasted with the more complicated procedures necessary to generate oral–nasal series by acoustic synthesis methods. In another application (Raphael *et al.*, 1979), the articulatory synthesizer has been used to test hypotheses about articulation made on the basis of physiological (EMG) evidence on one hand, and acoustic evidence on the other hand. Additional future applications include a series of experiments intended to study the perceptual effects of variations in the relative timing of articulatory movements. Such investigations address the nature of the underlying organization of the speech act in terms of its dynamic "units." A planned technical improvement will be the development of a flexible display system that can function like a stop-frame projector.

Due to our current hardware and software limitations, a number of sacrifices have been made in modeling that, however, permit a level of computational ability that is sufficient for the use of this synthesizer as an adequate research tool. The nature of these sacrifices, it is felt, still leave us with a model that captures the essentials of articulation. As we gain further insight into the anatomy and physiology of speech production, we would like to incorporate this additional knowledge into the model. Future modifications will be considered if the constraint of adequate speed for a research tool is not violated and, additionally, if they provide a mode of control that remains both flexible and articulatorily natural. The articulatory synthesis system, as described in this paper, already serves as a powerful research tool for examining perception–production re-

lationships. We expect that the synthesizer's usefulness will grow as the system evolves and as we refine the issues to be investigated with its aid.

ACKNOWLEDGMENTS

We thank Charles W. Marshall, Leonard Szubowicz, and Steven B. Davis for their significant contributions to the development of the articulatory synthesis program. This research was supported under NSF Grant BNS-76-82023 and BRSG Grant RR-05596 to Haskins Laboratories.

APPENDIX A: AREA-FUNCTION TO ACOUSTIC-TRANSFER-FUNCTION CALCULATION

The acoustic model of speech production, given the vocal-tract area function, is indicated in Fig. 6. The various branches of the vocal tract are treated as linear two-port networks. For voiced or aspirated sounds [Fig. 6(a)], where the velar port may be partially open, the glottal source U_g, feeds the left part of the pharyngeal branch. (For convenience, the glottal source impedance has been brought inside the box.) The right side of the pharyngeal branch is connected in parallel with the nasal and oral branches. On the right side of these boxes appear the radiated nasal and oral pressure, each across an open circuit, since radiation characteristics have been brought inside the boxes. The output sound is the sum of the nasal and oral pressures. The junction point, at which the three subsystems are connected in parallel, corresponds anatomically to the top of the pharynx, at the level of the velopharyngeal port. However, if the nasal port is closed, the nasal branch drops out, and it no longer matters at what anatomical level the two remaining boxes split.

For fricative sounds [Fig. 6(b)], there is a noise source anterior to a constriction. The system splits into two parts; a front cavity, and a back cavity (which anatomically includes the constriction and also includes the source resistance associated with the noise). Across the other side of the front cavity is the radiated pressure from the mouth, where, again, the radiation characteristics have been brought inside the box. Across the other port of the back cavity appears the glottal source, if any. As before, glottal impedance has been brought inside the box.

In both parts of Fig. 6, the glottal source and the leftmost box can be replaced by their Norton equivalent. If the two-port network obeys reciprocity, the Norton equivalent source, $U_{g\,eff}$, is related to the actual glottal source, U_g, by the relation

$$U_{g\,eff} = U_g G_p ,$$

where G_p is the open-circuit pressure gain $p_g/p_p \,|\, U_g = 0$.

Using the Norton equivalent for the pharyngeal branch, as indicated in Fig. 7, it can be seen that the output for Fig. 6(a) is

$$p_m + p_n = [U_{g\,eff}/(1/Z_m + 1/Z_n + 1/Z_p)]\,(G_m + G_n) ,$$

where G_m and G_n are the open-circuit gains $p_m/p_p \,|\, U_g = 0$

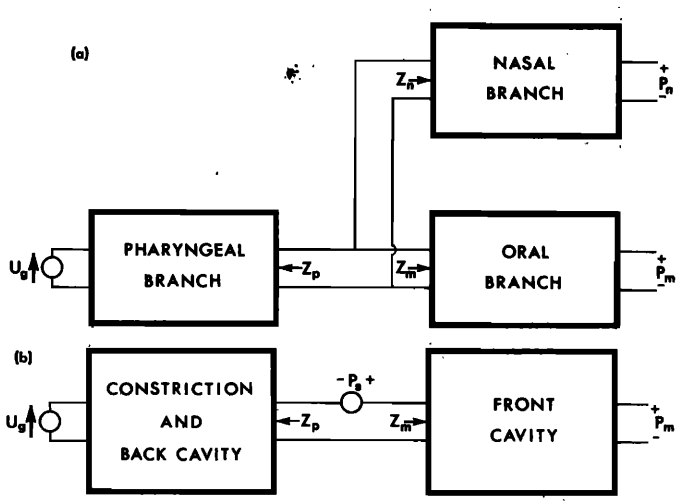


FIG. 6. (a) Block diagram for voiced and aspirated sounds, (b) block diagram for fricatives.

and $p_n/p_p \,|\, U_n = 0$, respectively. Therefore the transfer function is

$$(p_m + p_n)/U_g = [G_p(G_m + G_n)]/(1/Z_m + 1/Z_n + 1/Z_p) . \tag{A1}$$

If the nasal tract is not present (that is, if the velopharyngeal port is closed), then $Z_n = \infty$ and $G_n = 0$. Then, a corresponding equation accounts for the glottal component in Fig. 6(b).

The transfer function for the fricative component in Fig. 6(b) is

$$p_m/p_s = [Z_m/(Z_m + Z_p)]G_m . \tag{A2}$$

Thus all the relevant transfer functions can be calculated if the input impedances looking from the junction point and the open circuit pressure gain functions of all branches of the vocal tract are known. An iterative procedure for calculating these functions is described below.

For the purpose of calculations, the vocal tract is modeled as a series of uniform tubes with varying cross dimension but uniform length of 0.875 cm. A plane wave entering one end of such a section reaches the other end with a half time-unit (0.025 ms) delay and an attenuation $\alpha^{1/2}$, which depends on the cross-sectional area. We will consider one such section with cross-sectional area A, looking into an acoustic impedance Z_L from one end. When seen from inside the

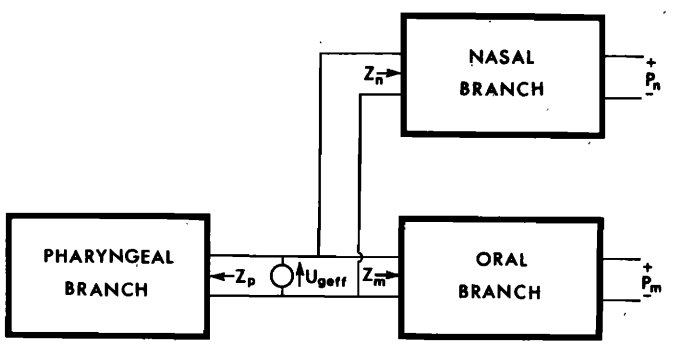


FIG. 7. Equivalent block diagram for voiced and aspirated sounds.

tube, this impedance produces a complex reflection coefficient

$$\Gamma = (Z_L - Z_0)/(Z_L + Z_0)\,, \tag{A3}$$

where $Z_0 = 40/A$ is the characteristic impedance of the tube. (All physical quantities are expressed in cgs units.) The impedance, Z, looking into the other end of the tube is then

$$Z = Z_0(1 + \alpha z^{-1}\Gamma)/(1 - \alpha z^{-1}\Gamma)\,,$$

where z is the Z-transform variable and Γ is expressed in Z-transform notation. The pressure gain across the tube is

$$G = (\alpha^{1/2} z^{-1/2})(1 + \Gamma)/(1 + \alpha z^{-1}\Gamma)\,.$$

Consider now tube section n, of area A_n, looking into a load impedance Z_{n-1}, which produces the reflection coefficient Γ_n. The next section, which has area A_{n+1}, sees an acoustic impedance

$$Z_n = (40/A_n)(1 + \alpha_n z^{-1}\Gamma_n)/(1 - \alpha_n z^{-1}\Gamma_n)\,,$$

which can be considered a reflection coefficient

$$\Gamma_{n+1} = (r_n + \alpha_n z^{-1}\Gamma_n)/(1 + r_n \alpha_n z^{-1}\Gamma_n)\,,$$

where

$$r_n = (A_{n+1} - A_n)/(A_{n+1} + A_n)\,.$$

This can, in turn, be used to find the impedance or reflection coefficient on the other side of section $n+1$, and the gain across it.

We now express the reflection coefficients as ratios of polynomials, so that

$$\Gamma_n = P_n/Q_n\,,$$

where P and Q are polynomials in z. Therefore,

$$P_{n+1} = r_n Q_n + \alpha_n z^{-1} P_n \tag{A4}$$

and

$$Q_{n+1} = Q_n + r_n \alpha_n z^{-1} P_n\,. \tag{A5}$$

The impedance into section n from the end of section $n+1$ is

$$Z_n = (40/A_{n+1})(Q_{n+1} + P_{n+1})/(Q_{n+1} - P_{n+1})\,, \tag{A6}$$

and the pressure gain across section n is

$$G_n = (\alpha_n^{1/2} z^{-1/2})(Q_n + P_n)/(Q_n + \alpha_n z^{-1} P_n)\,.$$

But

$$Q_n + \alpha_n z^{-1} P_n = (Q_{n+1} + P_{n+1})/(1 + r_n)\,,$$

so that

$$G_n = \alpha_n^{1/2} z^{-1/2}(1 + r_n)(Q_n + P_n)/(Q_{n+1} + P_{n+1})\,,$$

and the gain over sections 1 to N can be calculated:

$$G = z^{-n/2}\,\frac{Q_1 + P_1}{Q_{N+1} + P_{N+1}} \prod_{n=1}^{N} [\alpha_n^{1/2}(1 + r_n)]. \tag{A7}$$

To begin the transfer-function calculation, the source impedance or the radiation impedance (and gain) at the end of each branch of the vocal tract must be known and expressed as a ratio of polynomials in z. These are then used to determine the Q and P polynomials at the external end of the branch, using Eq. (A3), and the iterative equations (A4) and (A5) are applied one section at a time until P and Q at the other (internal) end of the branch are determined. Lumped losses can also be introduced during these iterations. Both the impedance and gain can then be calculated, using Eqs. (A6) and (A7), respectively. When this is done for all branches of the vocal tract, the glottal and fricative transfer functions can be calculated, using Eqs. (A1) and (A2). Standard techniques are then used to implement the transfer functions as digital filters to perform the synthesis.

APPENDIX B: VOCAL-TRACT TERMINATIONS AND LOSS PARAMETERS

Radiation impedance at the mouth or nose is modeled by a parallel resistor–inductor circuit, as described by Flanagan (1972). This impedance, Z, is given by

$$Z = (1 - z^{-1})/[2/R + 0.7(1 - z^{-1})]\,,$$

where R is the effective radius of the orifice. For the nose, an effective radius of 0.7 cm is used, while the value at the lips is calculated from area at the lips, A_n, by the formula:

$$R = (A_n/\pi)^{1/2}\,.$$

Relative contributions from the nose and mouth to the final acoustic output are in proportion to these radii.

Glottal impedance is modeled by a series resistor–inductor circuit. The values of these elements are adjusted to account for the effects of yielding vocal-tract walls. In the default state, a resistive component of 50 Ω and a reactive component of 1200 Ω are used. Fricative source resistance may also be changed, but is normally set to the characteristic impedance of the tube section behind the source.

Within the oral and pharyngeal parts of the vocal tract, the propagation loss $\alpha^{1/2}$, associated with a single transversal of each tube section is given by:

$$\alpha^{1/2} = 1 - 0.007/\sqrt{A}\,,$$

where A is the cross-sectional area of the tube. Within the nasal tract, a constant propagation loss factor of 0.99 is used, but additional losses are added by the use of lumped parameter elements. Specifically, a series resistor and inductor with a real axis pole at 150 Hz is inserted in the middle of the tract, and an additional resistive component is added at the coupling point.

[1] Although there are synthesizers which have more sophisticated and realistic models of the acoustic sources and of the area function to sound transformation (e.g., Flanagan *et al.*, 1975; Flanagan *et al.*, 1980), these systems are too computationally inefficient to serve as interactive research tools on equipment which is generally available to most laboratories.

Abramson, A. S., Nye, P. W., Henderson, J., and Marshall, C. (1979). "The perception of an oral–nasal continuum gen-

erated by articulatory synthesis," Haskin Labs. Status Rep. Speech. Res. SR-57, 17–38.
Coker, C. (1976). "A model of articulatory dynamics and control," Proc. IEEE 64, 452–460.
Coker, C., and Fujimura, O. (1966). "Model for specification of the vocal-tract area function," J. Acoust. Soc. Am. 40, 1271.
Flanagan, J. L. (1972). *Speech Analysis Synthesis and Perception*. (Springer–Verlag, New York).
Flanagan, J. L., and Ishizaka, K. (1976). "Automatic generation of voiceless excitation in a vocal-cord/vocal-tract speech synthesizer," IEEE Trans. Acoust. Speech Sig. Process. ASSP-24, 163–70.
Flanagan, J. L., Ishizaka, K., and Shipley, K. L. (1975). "Synthesis of speech from a dynamic model of the vocal cords and vocal tract," Bell Syst. Tech. J. 54, 485–506.
Flanagan, J. L., Ishizaka, K., and Shipley, K. L. (1980). "Signal models for low bit-rate coding of speech," J. Acoust. Soc. Am. 68, 780–791.
Harshman, R., Ladefoged, P., and Goldstein, L. (1977). "Factor analysis of tongue shapes," J. Acoust. Soc. Am. 40, 693–707.
Heinz, J. M., and Stevens, K. N. (1964). "On the derivation of area functions and acoustic spectra from cineradiographic films of speech," J. Acoust. Soc. Am. 36, 1037.
Kelly, J. L., Jr., and Lochbaum, C. (1962). "Speech synthesis," in *Proceedings of the Stockholm Speech Communications Seminar* (R. I. T., Stockholm, Sweden).
Kiritani, S., and Sekimoto, S. (1977). "Parameter description of the tongue movements in vowel production," in *Articulatory Modeling and Phonetics*, edited by R. Carre, R. Descout, and M. Wajskop (G. A. L. F. Groupe de la Communication Parlee, Brussels).
Ladefoged, P., Anthony J., and Riley, D. (1971). "Direct measurements of the vocal tract," J. Acoust. Soc. Am. 49, 104.
Maeda, S. (1979). "An articulatory model of the tongue based on a statistical analysis," in *Speech Communication Papers*, edited by J. J. Wolf and D. H. Klatt (Acoustical Society of America, New York), pp. 67–70.
Mermelstein, P. (1971). "Calculation of the vocal-tract transfer function for speech synthesis applications," *Proceedings of the Seventh International Congress on Acoustics, Vol. 3* (Akademiai Kiado, Budapest), pp. 173–176.
Mermelstein, P. (1972). "Speech synthesis with the aid of a recursive filter approximating the transfer function of the nasalized vocal tract," in *Proceedings of the 1972 International Conference on Speech Communication and Processing* (Boston, Mass.).
Mermelstein, P. (1973). "Articulatory model for the study of speech production," J. Acoust. Soc. Am. 53, 1070–1082.
Mermelstein, P., Maeda, S., and Fujimura, O. (1971). "Description of tongue lip movement in a jaw-based coordinate system," J. Acoust. Soc. Am. 49, 104.
Perkell, J. S. (1969). *Physiology of Speech Production: Results and Implications of a Quantitative Cineradiographic Study* (MIT, Cambridge, MA).
Raphael, L. J., Bell–Berti, F., Collier, R., and Baer, T. (1979). "Tongue position in rounded and unrounded front vowel pairs," Lang. Speech 22, 37–48.
Rosenberg, A. E. (1971). "Effect of glottal pulse shape on the quality of natural vowels," J. Acoust. Soc. Am. 49, 583–590.
Timcke, R., von Leden, H., and Moore, P. (1958). "Laryngeal vibrations: measurements of the glottic wave. Part I. The normal vibratory cycle," AMA Arch. Otolaryngol. 68, 1–19.