Self-Oscillating Source for Vocal-Tract Synthesizers

JAMES L. FLANAGAN, SENIOR MEMBER, IEEE
LORINDA L. LANDGRAF

Abstract—A self-oscillating model of the human vocal cords is derived and simulated on a digital computer. The model is used as a source of excitation for a vocal-tract synthesizer, also programmed on the computer. Synthetic speech from the simulation is used to study the influence of glottal parameters upon signal features. The cord model produces glottal volume velocity functions which reflect the acoustic interaction between source and tract. Voice pitch and irregularities in excitation are generated intrinsically from specification of subglottal pressure, cord tension, and tract configuration. Pitch produced by the cord model is a monotone increasing function of subglottal pressure and tension. Mean air flow and glottal duty factor depend upon a combination of parameters, but primarily upon the properties of the contacting surfaces during cord closure.

Manuscript received October 10, 1967. This paper was presented at the 1967 Conference on Speech Communication and Processing, Cambridge, Mass.

The authors are with Bell Telephone Laboratories, Inc., Murray Hill, N. J.

ONE PROBLEM in generating synthetic speech of acceptable quality is the duplication of natural detail in vocal excitation information. This paper describes an excitation source which has certain similarities to the vocal-cord source in the human vocal tract. Its use with a vocal-tract synthesizer leads to automatic determination of glottal waveform, pitch, duty factor, and acoustic interaction of source and system. The parameters which are supplied correspond to the physiological factors of vocal-cord tension and subglottal pressure—two factors which the human manipulates overtly in producing speech.

For the production of voiced sounds, the vocal system can be schematized as in Fig. 1.[1] The acoustic compliance of the lung volume is represented by the capacitor C_L and its loss by the conductance G_L. The excess pressure in the lungs is P_L and, during speech at constant vocal effort, this pressure is maintained relatively constant by contraction of the ribcage. That is, as air is expelled and the charge on C_L diminished, the lung volume is diminished to maintain a constant ratio of charge to capacity.

The two parallel bronchi have a length of 3–5 cm and a total cross-section of about 400 mm^2, and are represented by a single equivalent T-section. Similarly, the single tube of the trachea, about 12 cm long and also about 400 mm^2 in section, is represented by a second T-element.

The opening between the vocal cords, the glottis, is a slit about 18 mm long in a man. Its area can vary from 0 to upwards of 20 mm^2, and its acoustic impedance is primarily resistive and inertive. It is represented in the circuit by the time-varying components R_g and L_g. The air pressure immediately under the glottis is P_s, and the volume velocity passing the glottal orifice is U_g. The acoustic load that the glottal system sees is the vocal-tract driving-point impedance Z_t. Sound radiation from the mouth is represented by the impedance Z_r. The acoustic volume velocity through the mouth and through the radiation impedance is U_m.

This model of vocal excitation has been studied previously, where the glottal area function was measured from high-speed movie film and supplied as a function of time to the circuit.[2] In the real larynx, however, the vocal cords operated as an aerodynamic oscillator, and their motion is a self-determined function of physica parameters, such as subglottal pressure, vocal-cor tension, and vocal-tract configuration. A more realisti model for speech synthesis and articulatory studie would, therefore, reflect these self-oscillating properties

The feedback controls for such a system can be illustrated as in Fig. 2. Here, because the lungs appear as a low-impedance constant-pressure source, and because the pressure drop across the large-area bronchi and trachea is relatively small, we approximate the subglottal pressure by the variable battery P_s. Using the experimental results of van den Berg,[3] we represent

the time varying glottal impedance by a viscous non-flow dependent resistance R_v; a kinetic flow-dependent resistance R_k; and an inertance owing to the mass of the glottal air plug L_g. In previous work[4] we have described these quantities as

$$R_v = 12\mu d l^2 A_g^{-3}$$
$$R_k = 0.44\rho \left| U_g \right| A_g^{-2}$$
$$L_g = \rho d A_g^{-1},$$

where

μ is the kinematic viscosity of air,
d is the vocal-cord thickness (depth),
l is the cord length,
A_g is the area of the glottal orifice,
ρ is the air density,

and

U_g is the acoustic volume velocity through the glottal orifice.

As indicated schematically in Fig. 2, $A_g(t)$ is determined from feedback relations involving subglottal pressure P_s, glottal flow U_g, and vocal cord tension Q. The vocal tract, for any given configuration, can be represented as accurately as desired by abutting T-sections of lumped-constant elements, and terminated in an impedance representing the radiation load at the mouth. The acoustic volume current at the mouth is U_m, and the pressure at a fixed point in front of the mouth is approximately the time derivative of this current. The cross-sectional area of the tract, as a function of distance along it, is described by the set of areas $A_1, \cdots A_m$. In connected speech, these areas vary with time.

The feedback relations necessary to generate $A_g(t)$ from P_s, U_g, and Q derive from coupled equations linking these parameters to those of a simple mechanical model of the vocal cords. As has been proposed in previous work,[4],[5] the vocal cords are considered a simple second-order system; that is, a simple mechanical oscillator of mass M, spring constant K, and viscous damping B. The system is driven by a forcing function $F(t)$. This situation is illustrated in Fig. 3. Motion of the mechanical oscillator is described by the equation $M\ddot{x}+B\dot{x}+Kx=F$, where $x(t)$ is the displacement of the mass. Although only a single movable mass is shown in Fig. 3, both vocal cords in the real larynx actually move. For our purposes here, we will assume symmetry and deal with only one mass. The mass is fixed at a value for the human cords, and the spring constant can be left a parameter corresponding to cord tension.[1] Damping can be determined experimentally.

[1] Actually, it is more convenient to deal with variations in undamped natural frequency, produced by changing the values of both K and M.

As shown in Fig. 3, the acoustic pressures at the inlet and outlet of the glottal orifice are P_1 and P_2, respectively. Experimental measurements[3] show that these pressures can be approximated as

$$P_1 = (P_s - 1.37P_B)$$
$$P_2 = -0.50P_B$$

where P_B is the Bernoulli pressure given by $P_B = \frac{1}{2}\rho \left| U_g \right|^2 A g^{-2}$. In the present study, the forcing function is taken as the mean inlet and outlet pressures, i.e.,

$$F(t) = \tfrac{1}{2}(P_1 + P_2)(ld)$$

acting on the vocal cord face (i.e., the intraglottal surface area).[2] At rest, the displacement of the mass is zero ($x=0$) and the glottal area is the "phonation neutral" value A_{g0}. When the displacement attains a critical value $x_c = -A_{g0}/l$, complete closure occurs, and A_g and U_g become zero until x becomes greater than x_c.

The boundary which the vibrating mass M strikes is analogous to the flesh of the opposing vocal cord. It may reflect mechanical properties of mass, stiffness, and viscosity, M', K', and B', respectively. For our present purposes, two conditions of the boundary at closure are interesting. One is the infinitely hard (massive) boundary which, upon striking, the mass gives up all its momentum instantaneously. For this condition, the minimum displacement is the critical value for closure x_c. During closure, the forcing function becomes $\frac{1}{2}(P_s)(ld)$, and the forces acting on the mass are immediately in a direction to open the port. Consequently, this boundary condition generally results in relatively brief and constant closure times.

A second condition of the boundary at closure is a purely viscous contact. Imagine that the opposing surface which the mass strikes is relatively massless and mainly viscous, with coefficient B'. (For example, imagine that the opposing wall is an oil bath into which the mass dips.)[3] Upon contact, the U_g flow is interrupted, but the mass, through its inertia, continues on to a displacement which exceeds x_c. All the while, the forcing function is constant at $\frac{1}{2}(P_s)(ld)$, and the computations of new value of $x(t)$ proceed under this condition and with total damping equal to the sum $(B+B')$ until $x > x_c$. At this time, the port opens, the forcing function becomes different from the constant value $\frac{1}{2}P_s(ld)$, and the system damping returns to B.

If the nonuniform vocal tract is represented by a transmission line of abutting T-sections, the first two loop equations for the circuit of Fig. 2 are simply

[2] It is clear that other forcing functions can be hypothesized and examined. One that leads to interesting and relevant results is $F(t) = -P_B(ld)$.

[3] While this, of course, is an extreme condition, one imagines the surfaces of the real cords to be somewhat viscous and to form into one another during contact.

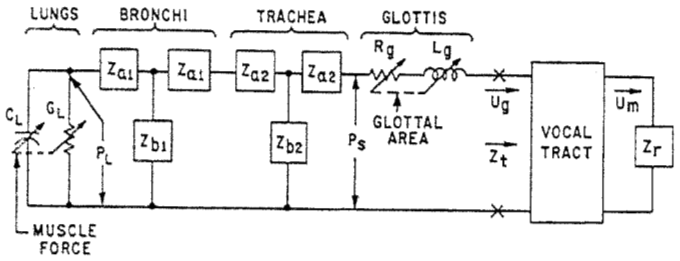

Fig. 1. Acoustic circuit representation for the production of voiced sounds.

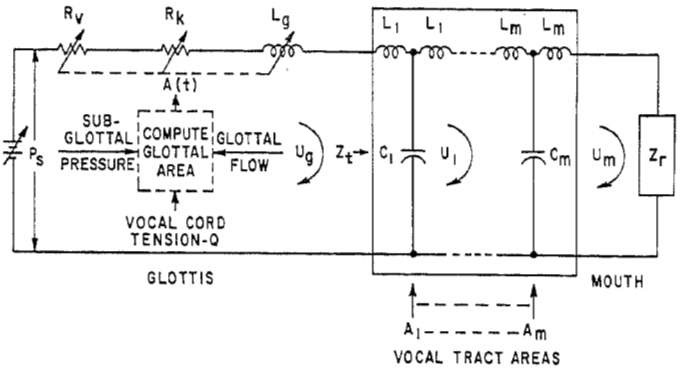

Fig. 2. Simplified equivalent circuit for voiced sounds.

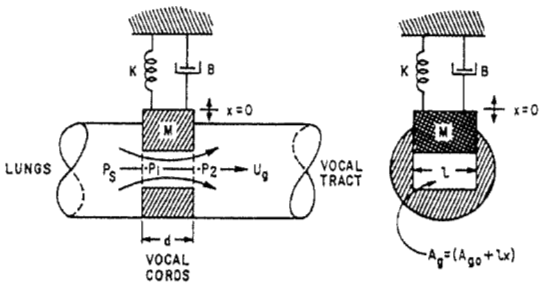

Fig. 3. Second-order mechanical model of the vocal cords.

$$(R_v + R_k)U_g + \frac{d}{dt}(L_g U_g) + \frac{d}{dt}(L_1 U_g) + \frac{1}{C_1}\int (U_g - U_1)dt = P_s \quad (1)$$

$$\frac{d}{dt}[(L_1 + L_2)U_1] + \frac{1}{C_2}\int (U_1 - U_2)dt + \frac{1}{C_1}\int (U_1 - U_g)dt = 0 \quad (2)$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$(m) \qquad \vdots \qquad (m)$$

The first equation of this set is coupled to the equation for the mechanical oscillator through the glottal flow U_g and its relation to the Bernoulli pressure. All circuit elements are, in general, functions of time. In addition, R_k is dependent upon $|U_g|$.

These continuous equations can be approximated by difference equations, and the latter can be solved by discrete numerical operations in a digital computer. We have programmed this set of equations on a GE 645 computer for the case of a 10-section vocal-tract synthesizer.[4] The program permits systematic examination of glottal parameters. In addition, we have calculated the sound pressure corresponding to the mouth current U_m, and have digital-to-analog converted this synthetic speech signal for listening evaluations.

The computation of A_g and U_g (and all of the subsequent loop volume veclocities $U_1 \cdots U_m$) proceeds as follows. For initial values of P_s and K, one starts with the neutral area A_{g0}. The first sample of U_g is then calculated for this area. From the first sample of U_g, the first sample of $F(t)$ is calculated, and then the first sample of x, taking past samples as zero. From the new sample of x, the first sample of A_g is computed, and the process iterated.[5]

For appropriate human constants, the mass generally executes nonsinusoidal oscillations (as do the human vocal cords), and the continuous function U_g is generated. The form and period of U_g will depend not only upon the glottal and subglottal constants, but also upon the vocal-tract configuration. In describing typical results from the computer simulation, it is convenient to specify a range of parameters. For present purposes, the following conditions hold: A normal chest-register mass of the moving vocal cords is taken as $M = \frac{1}{4}\sigma l^2 d$, where σ is the density of the cord flesh (assumed approximately 1.0), hence $M = 0.24$ gram. Corresponding

[4] This program, if desired, allows specification of a length parameter with each section area. If section lengths are considered fixed and equal, they correspond to approximately 1.7 cm of a man's vocal tract. A second program, not discussed here, represents the tract to 18 sections.

[5] From the first sample of U_g, the first samples of U_1 and subsequent loop currents are also computed by working successively through the set of m loop equations.

to this mass, a spring constant is chosen to produce an undamped natural frequency of the vibrating system, well below the fundamental frequency when forced; an appropriate value is $f_0 = 25$ Hz; then, $K = 4\pi^2 M (f_0)^2$. Damping is conveniently expressed as a percentage of the critical value, i.e., $B = k2\sqrt{MK}$. For present examples, we will use $B = 0$, and $B' = 2\sqrt{MK}$. As previously noted, it appears more realistic to introduce the cord-tension parameter as a combined stiffening and lightening of the moving system. That is, the cord-tension parameter Q may modify the undamped natural frequency of the system by the product $Qf_0 \equiv (1/2\pi)\sqrt{QK/(M/Q)}$.

The two types of boundary collisions presently of interest are the hard boundary (i.e., infinitely massive) and the viscous boundary (i.e., where the total damping becomes $(B+B')$ for the time that closure exists and is B for other times).

The values of U_g and A_g, computed from the program, are plotted under computer control on an SC-4020 microfilm plotter. Typical results of these computations are shown for several vocal tract shapes in Fig. 4.[6] For all these cases, the glottal conditions are $P_s = 8$ cmH_2O; $A_{g0} = 0.05$ cm^2; $Q = 2$, and the boundary during closure is viscous with $B = 0$ and $B' = 2\sqrt{MK}$.

From the results shown in Fig. 4, one notices that the vocal-tract shape can influence the waveform and fundamental frequency of the glottal flow.[7] The acoustic interaction between source and system is particularly evident in the U_g waveforms for the vowels |a| and |i|. In both these cases, one identifies temporal detail corresponding to the first normal mode of the tract (i.e., the first formant).

Also from Fig. 4, one sees that steady-state oscillation of the cord is reached by about the fourth cycle. The starting motion is, in general, irregular, and is accompanied by appreciable changes in waveshape and period. Similar irregularities apparently occur in the real larynx. The glottal flow wave U_g generally shows steeper slope on its falling phase than on its rising phase, consistent with inverse-filter measurements of human glottal flows. The results show, too, that the area wave and its corresponding glottal flow wave can differ substantially in form. The tendency is for the area wave to be the more symmetrical. In the data shown for |i|, the initial motion of the model is largely to diminish the glottal area, while in the data for |a|, the initial motion is more nearly to increase the area. The initial displacement depends upon parameter values, and both opening and closing starting phases are commonly observed.

An important aspect of the model is the pitch it produces for different subglottal pressure and cord tension. Measurements of this behavior are shown for both the hard and viscous boundaries in Fig. 5. In these plots, the vocal-tract shape is the neutral schwa |ə| $[(A_1 \cdots A_m) = 5$ cm$^2]$ and $A_{g0} = 0.05$ cm^2. The tension factor Q is shown as the parameter. The curves exhibit a monotone increase in pitch with subglottal pressure and, over parts of the range, the relation is reasonably linear.

Measurements on human sound production give pitch variations in the range 2.5 to 20 Hz/cm H_2O[6],[7] and the computer model falls into this range. Also, the variation of pitch with cord tension (or, more precisely, with the natural frequency of the system) is monotonically increasing and roughly linear.

Another aspect of interest is mean glottal flow $\overline{U}_g$ as a function of subglottal pressure. These calculations are plotted in Fig. 6 for the two boundary conditions. For comparable glottal conditions, the viscous boundary leads to greater displacements and peak flow values, and, hence, to greater air consumption. However, the ratio of closed time to whole period for the viscous boundary is greater than for the hard boundary, as shown by Fig. 7. Consequently, one might expect that the viscous condition could lead to a U_g spectrum rich in higher frequency components.

The present vocal cord model appears to be a useful tool for studying sound articulation in terms of parameters that have direct physiological correlates. In fact, with the complete circuit shown in Fig. 1, relations between respiration and articulation can be considered. Present efforts are being directed toward multiple-element representations of the cords to study factors such as cord asymmetry and phase differences in the motion of the lower and upper edges of the glottal orifice.

As already noted, the vocal cord model provides realistic voiced excitation for the programmed vocal-tract synthesizer. An additional objective is the study of synthetic speech produced from the self-oscillating source. A typical synthetic speech result, using the glottal model and the programmed vocal tract, is illustrated by the sound spectrogram in Fig. 8. For this case, glottal conditions are kept constant ($P_s = 8$ cm H_2O; $A_{g0} = 0.05$ cm^2, $Q = 2$) and the vocal-tract shape is changed linearly from the vowel configuration for |i| to that for |a|.[8] Each vowel has a duration of 100 ms, and the transition occurs in 100 ms. The pitch change evident in the spectrogram is produced solely by the influence of the changing tract shape; that is, by the changing acoustic load on the source.

[6] Using the computations of A_g, a computer movie, similar in format to the high-speed film of the vocal cords taken by D. W. Farnsworth, has been made. The two films, made at equivalent frame rates of 4000 sec^{-1}, permit comparisons of the computer model and the real larynx.

[7] Note that the final periods correspond to pitches of 114 Hz for |ə|, 93 Hz for |a|, and 104 Hz for |i|.

[8] These are precisely the conditions used for the U_g and A_g data shown in Fig. 4.

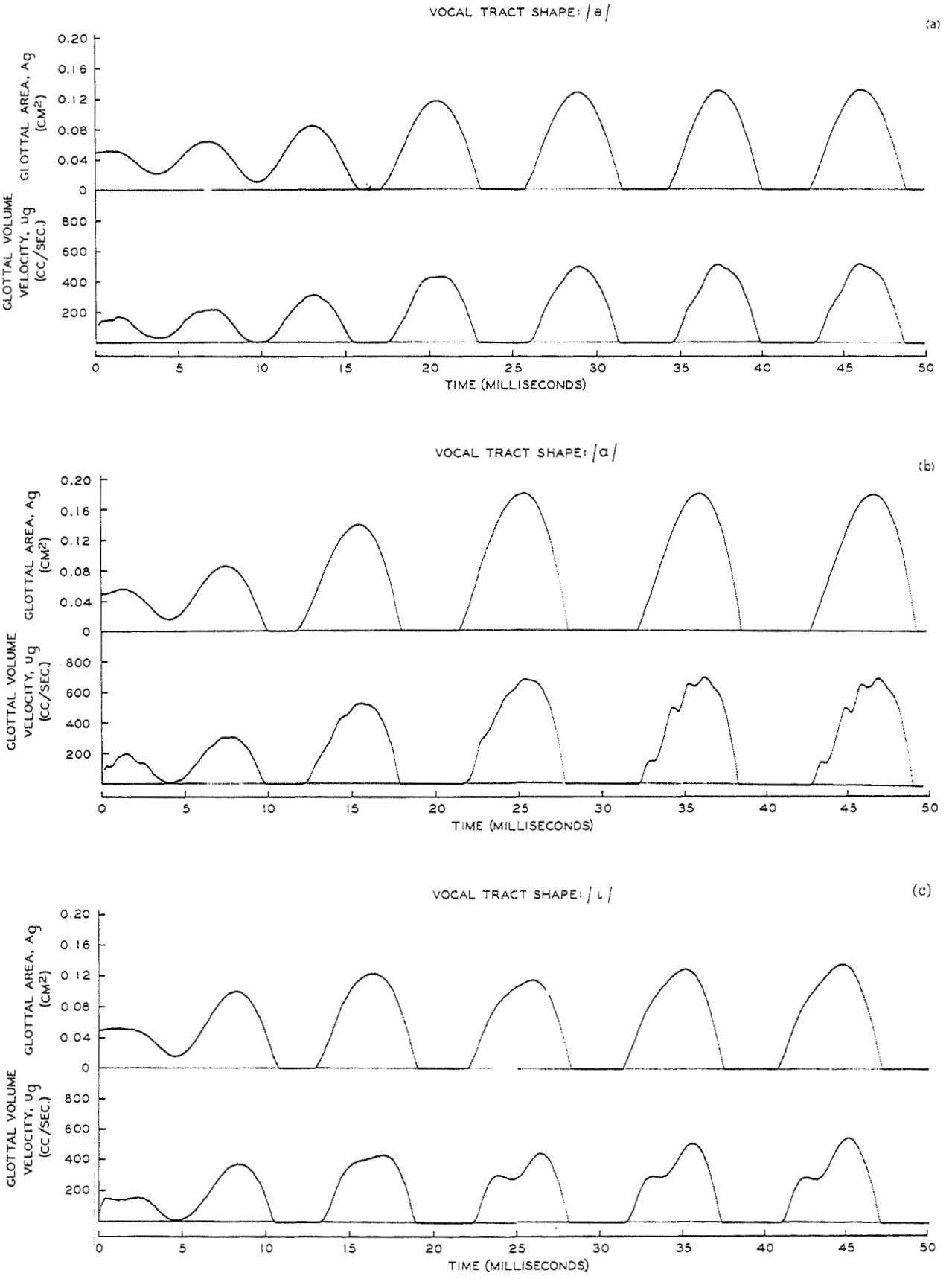

Fig. 4. Waveforms of glottal area and acoustic volume velocity produced by the computer model.

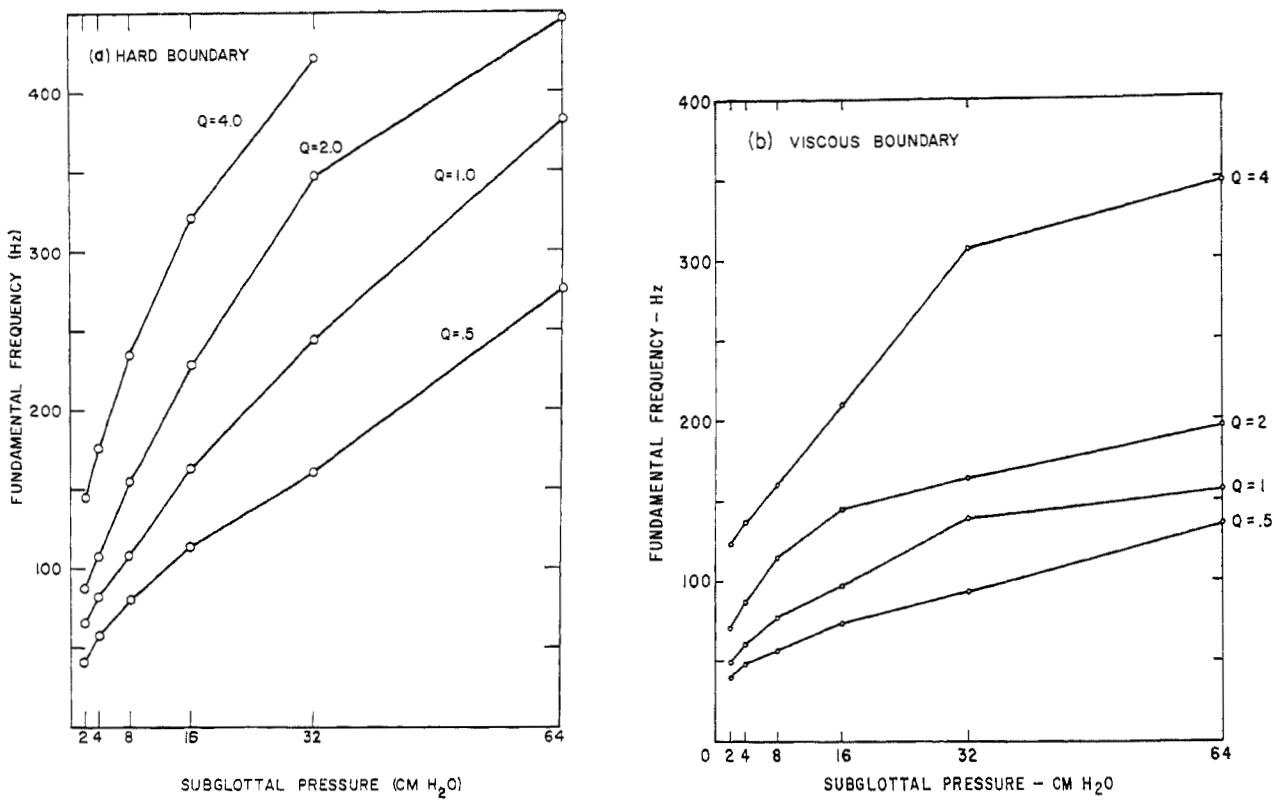

Fig. 5. Variation of fundamental frequency with subglottal pressure produced by the vocal cord model. Cord tension factor Q is the parameter. Vocal tract shape corresponds to the vowel |ə|.

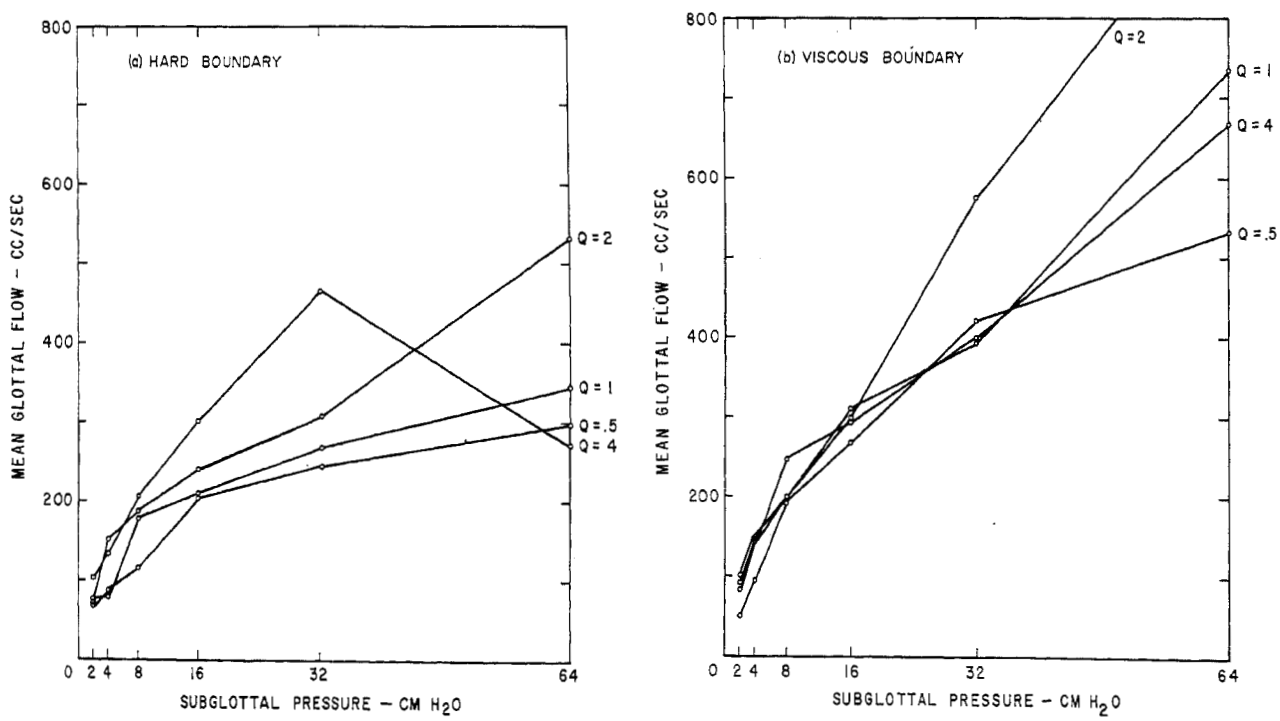

Fig. 6. Mean glottal flow as a function of subglottal pressure for the computer model.

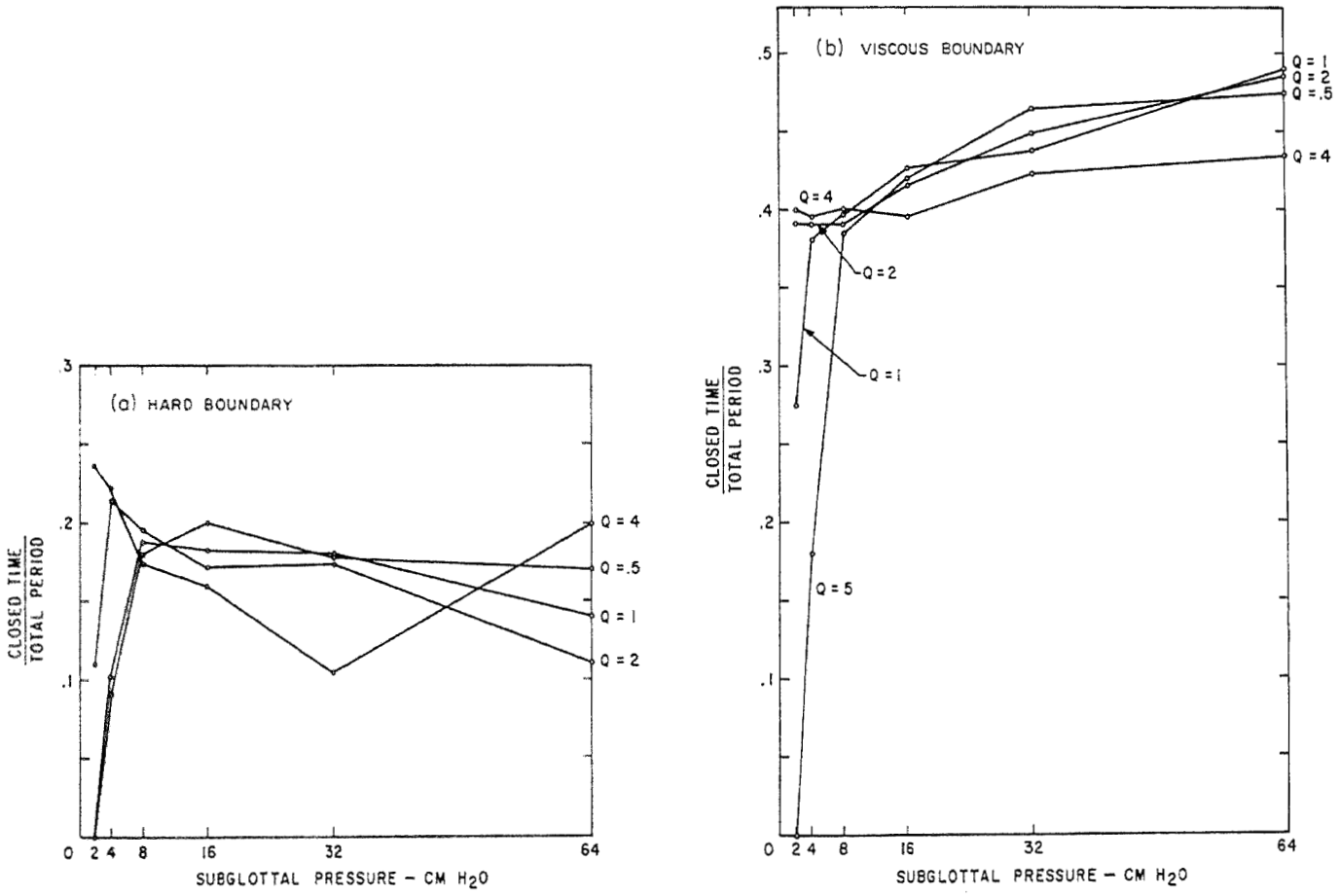

Fig. 7. Ratio of closed time to whole period for the vocal cord model.

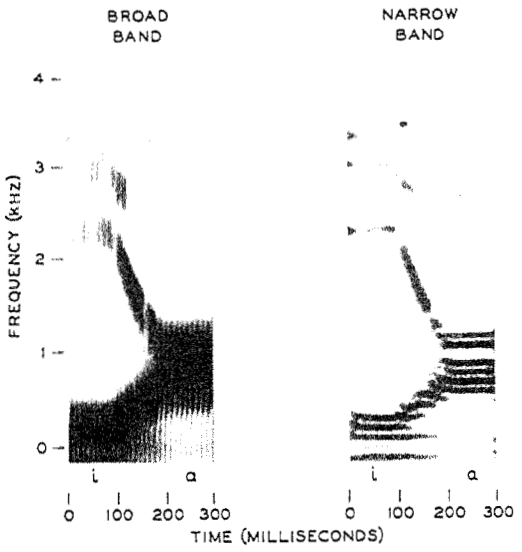

Fig. 8. Sound spectrograms of synthetic speech produced by the programmed vocal tract. The tract is excited by the self-oscillating model. The spectrograms show a linear transition from the vowel [i] to the vowel [a]. Glottal conditions are maintained constant.

Acknowledgment

In interpreting results from the single-element model of the vocal cords, and in extensions of the principles to multielement forms, we have enjoyed the collaboration and consultation of S. Öhman of the Royal Institute of Technology, Stockholm, Sweden. It is a pleasure to acknowledge his substantial contributions to these results.

References

[1] Additional details can be found in J. L. Flanagan, *Speech Analysis, Synthesis and Perception*. New York: Academic Press, 1965.

[2] J. L. Flanagan and D. I. S. Meinhart, "Source-system interaction in the vocal tract," *J. Acoust. Soc. Am.*, vol. 36, p. 2001(A), 1964.

[3] J. W. van den Berg, J. T. Zantema, and P. Doornenbal, Jr., "On the air resistance and the Bernoulli effect of the human larynx," *J. Acoust. Soc. Am.*, vol. 29, pp. 626–631, 1957.

[4] J. L. Flanagan, "Estimates of intraglottal pressure during phonation," *J. Speech and Hearing Research*, vol. 2, pp. 168–172, 1959.

[5] T. H. Crystal, "Model of larynx activity during phonation," M.I.T. Research Lab. of Electronics, Cambridge, Quart. Prog. Rept. 78, July 15, 1965.

[6] S. Öhman and J. Lindqvist, "Analysis and synthesis of prosodic pitch contours," Speech Transmission Lab., Royal Institute of Technology, Stockholm, Quart. Prog. and Status Rept., April 1965.

[7] J. W. van den Berg, "Direct and indirect determination of the mean subglottal pressure," *Folia Phoniatrica*, vol. 8, pp. 1–24, 1956.

James L. Flanagan (A'51–M'57–SM'67) was born in Greenwood, Miss., on August 26, 1925. He received the B.S. degree in electrical engineering from Mississippi State University, State College, in 1948, and the S.M. and Sc.D. degrees in electrical engineering from the Massachusetts Institute of Technology, Cambridge, in 1950 and 1955, respectively. He then completed two years of postdoctoral work at M.I.T. and at the A.F. Cambridge Research Center, Bedford, Mass.

He joined Bell Telephone Laboratories, Inc., Murray Hill, N. J., in 1957, and in 1961 became Head of the Speech and Auditory Research Department. In February, 1967, he was made Head of the Acoustics Research Department at BTL. His interests have centered on voice communication and its related areas and he has been concerned with signal processing methods for bandwidth conservation, and with fundamental acoustical studies of speech and human hearing. He holds a number of patents on speech-coding techniques, and is the author of a number of technical papers and a recent book, *Speech Analysis, Synthesis and Perception*.

Dr. Flanagan is a Fellow of the Acoustical Society of America, a member of its Technical Committee on Engineering Acoustics, and Chairman of its Publication Policy Committee. He is a member of the Committee on Hearing and Bioacoustics of the National Academy of Sciences, Tau Beta Pi, and Sigma Xi.

Lorinda L. Landgraf was born in Suffern, N. Y., on November 18, 1944. She received the B.A. degree in mathematics from the University of Delaware, Newark, in 1966 and is currently doing graduate study toward a masters in computer science at Steven's Institute of Technology, Hoboken, N. J.

During the summer of 1965 and since 1966, she has been a member of the Acoustics Research Department of Bell Telephone Laboratories, Inc., Murray Hill, N. J., where she is concerned with computer simulation of speech-coding techniques and computer methods for acoustic signal processing.