9.4, 9.8 Received 25 November 1968

Excitation of Vocal-Tract Synthesizers*

J. L. Flanagan and L. Cherry

Bell Telephone Laboratories, Incorporated, Murray Hill, New Jersey 07974

A method is described for representing voiced and voiceless excitation in the vocal tract. Three physiological factors suffice for synthesis of nonnasal speech: subglottal pressure, vocal-cord tension, and vocal-tract shape. Voiced excitation is obtained from an oscillator model of the cords. Voiceless excitation is produced by elements that approximate turbulent flow in the tract. These elements are controlled automatically by factors related to the acoustic volume velocity in the tract and the tract shape. Voiced, voiceless, and voiced-fricative sounds are synthesized from a vocal tract incorporating these sources. The whole system is computer simulated.

A DESCRIPTION of speech in physiological terms promises one route to improved synthesis. The physiological description should embrace the mechanism of sound generation in the vocal tract, as well as the tract geometry and its transmission properties. Toward this objective, we have made a computer simulation of the vocal tract, and of the mechanisms of voiced and voiceless excitation, at least as we presently understand them. Our aim is to synthesize all sounds in terms of three physiological factors: namely, subglottal pressure, vocal-cord tension, and vocal-tract shape. This report describes preliminary experiments to test the feasibility of this approach.

I. VOICED EXCITATION

For calculation of voiced excitation, we represent the vibrating vocal cords as a second-order mechanical system. The components are shown in Fig. 1. Air from the lungs on the left passes through the glottal constriction formed by the cords and into the vocal tract on the right. The acoustic volume velocity through the glottis is U_g, and the subglottal air pressure is P_s. The vibrating vocal cords are represented by the single mass M, the stiffness K, and the viscous loss B. The cords have thickness d and length l. Vertical displacement x, of the mass changes the glottal area A_g, and varies the flow U_g. At rest, the glottal opening has the phonation neutral area A_{g0}.

The mechanical oscillator is forced by a function of the subglottal pressure and the Bernoulli pressure in the orifice. The Bernoulli pressure is dependent upon U_g^2, which, in turn, is conditioned by the nonlinear, time-varying acoustic impedance of the glottal opening. In qualitative terms, the operation is as follows: the cords are set to the neutral or rest area, and the sub-

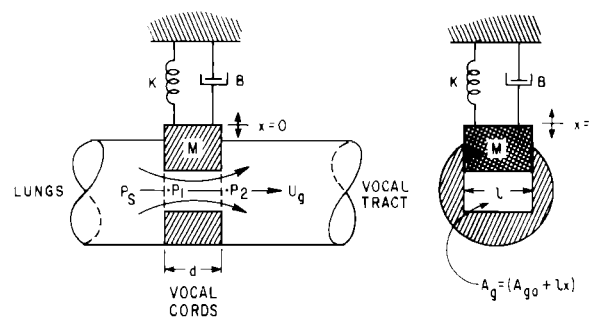

FIG. 1. Oscillator model of the vocal cords.

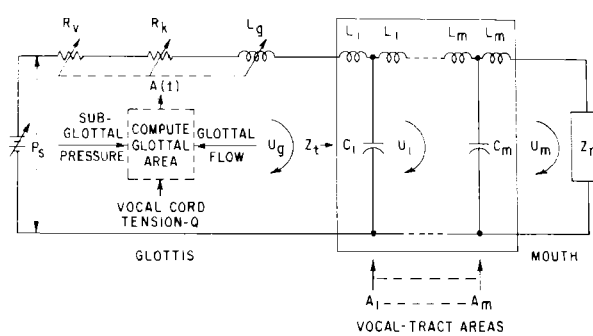

FIG. 2. Network representation of the vocal system for voiced sounds.

* Based upon material presented at the 6th International Congress on Acoustics, Tokyo, Japan, August 1968.

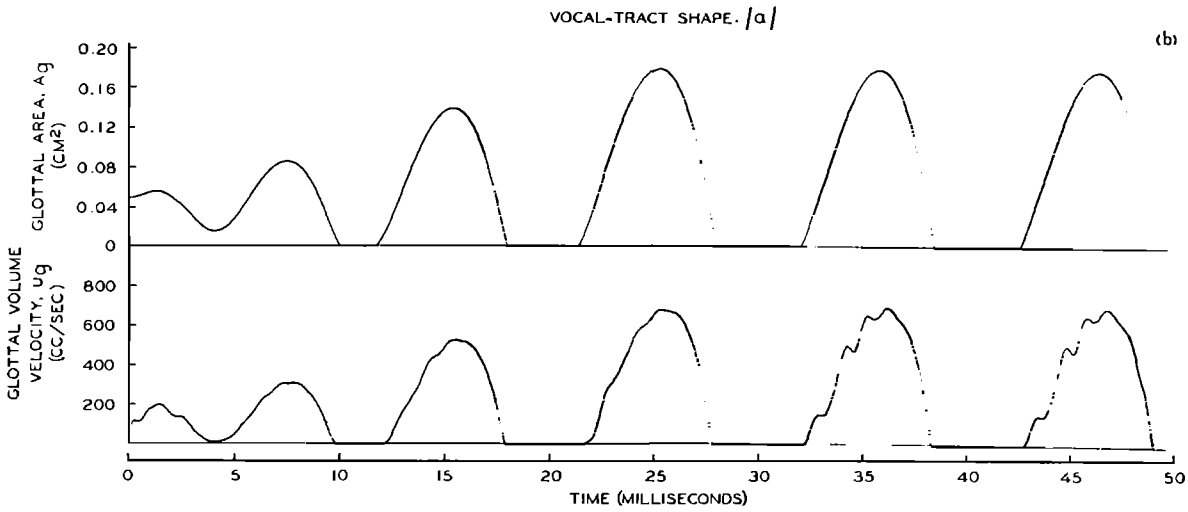

FIG. 3. Glottal area and acoustic volume velocity calculated from the vocal-cord model.

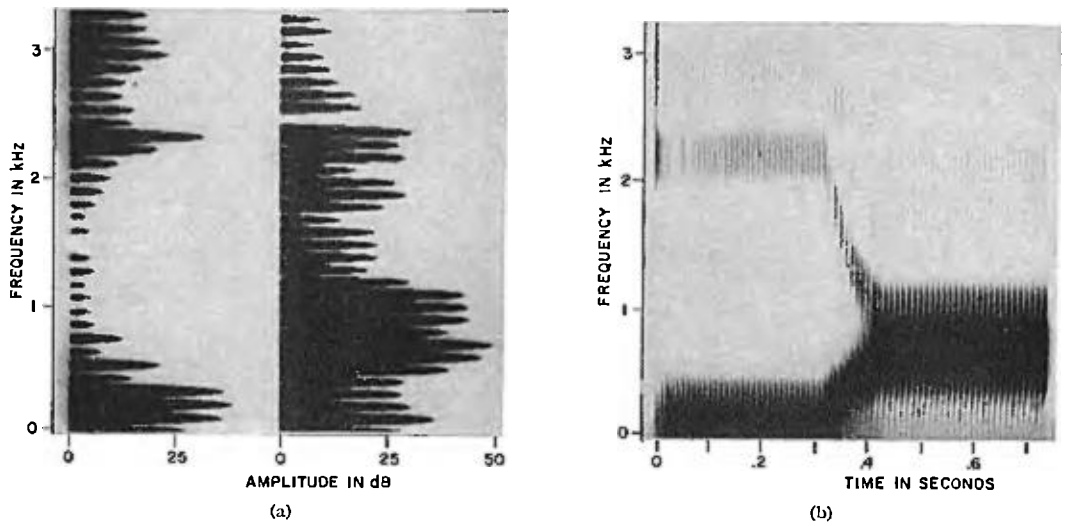

FIG. 4. Sound spectrogram of the synthesizer output. The sound corresponds to a linear transition from vowel /i/ to the vowel /a/. Amplitude sections are shown for the central portion of each vowel.

glottal pressure applied. As the flow builds up, so does the negative Bernoulli pressure. The latter draws the mass down to interrupt the flow. As the flow diminishes, so does the Bernoulli pressure, and the spring acts to retrieve the mass. Under appropriate conditions, stable oscillation results.[1]

The undamped natural frequency of the oscillator is proportional to $(K/M)^{\frac{1}{2}}$. It is convenient for us to define a vocal-cord tension parameter Q, which scales the natural frequency by multiplying the stiffness and dividing the mass. This is analogous to the physiological tensing of the cords, which stiffens them and reduces their distributed mass.

[1] Computational details of this formulation are described in J. L. Flanagan and L. Landgraf, "Self Oscillating Source for Vocal-Tract Synthesizers," IEEE Trans. Audio Electroacoust. AU-16, 57–64 (1968).

Using this formulation for glottal behavior, the complete vocal system, for voiced nonnasal sounds, may be represented by the network in Fig. 2.

The tract is represented as a nonuniform transmission line whose sections correspond to plane wave propagation in abutting cylindrical elements. The tract configuration is described by the cross-sectional area values $(A_1 - - - A_m)$. In our present case, $m=10$. The radiation load at the mouth is Z_r and is taken as the radiation impedance of a piston in a plane baffle. The acoustic volume velocity through the mouth, and hence into the radiation impedance, is U_m. The acoustic driving point impedance seen by the glottis is Z_t.

The acoustic impedance of the glottal orifice is characterized by two loss elements, R_v and R_k, and an inertance, L_g.[1] The values of these impedances depend upon the time-varying glottal area $A_g(t)$. In addition,

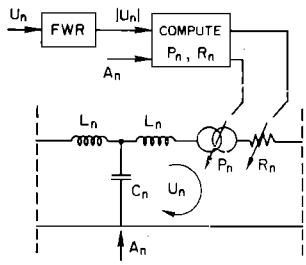

FIG. 5. Network modification for simulating the properties of turbulent flow in the vocal tract.

R_k is dependent upon $|U_g|$. The glottal area is linked to P_s and to U_g through the differential equation that describes the vocal-cord motion and its forcing function. The value of the tension parameter Q is also introduced into this equation. In other words, the dashed box of Fig. 2 represents iterative solutions to the differential equation for the system described in Fig. 1.

This continuous system can be represented by $(m+2)$ differential equations, which, in turn, can be approximated by difference equations. We have programmed these difference equations for simultaneous solution on a GE 645 digital computer. The program accepts as input data time-varying samples of the subglottal pressure P_s, the cord tension Q, and the vocal tract areas $(A_1 - - - A_m)$, and it computes sampled values of all volume velocities, including the glottal flow and mouth output. The results can be plotted for visual display and D/A converted for audible output. A typical glottal area and volume velocity, plotted by the computer for a vocal-tract shape corresponding to the vowel /ɑ/, is shown in Fig. 3. This Figure shows the initial 50 msec of voicing.

The top curve is the glottal area result, and the lower curve the glottal flow. The calculation is for a subglottal pressure of 8 cm H_2O and a tension value that places the cord oscillation in the pitch range of a man. One notices that by about the fourth period a steady state is achieved. One sees, in this case, irregularities in the glottal flow that are caused by acoustic interaction at the first formant frequency of the tract. One also notices that this temporal detail in the volume flow is not noticeably reflected in the mechanical behavior, that is in the area wave.

We have examined the behavior of the vocal-cord model for a wide range of glottal conditions and the results suggest that it duplicates many of the features of human speech (for example, the variation of fundamental frequency with P_s).[2–6] One of the best ways to assess the behavior is to listen to the sound quality resulting from the program. A spectrogram of the audible output for a linear transition from the vowel /i/ to the vowel /ɑ/ is shown in Fig. 4. The glottal conditions in this case are constant and are: $P_s = 8$ cm H_2O and $Q = 2.0$. The resulting fundamental frequency of these sounds is not only a function of the glottal parameters, but also of the tract shape; that is, a function of the acoustic loading that the tract presents to the vocal cords. The spectral sections indicate realistic formant and pitch values.

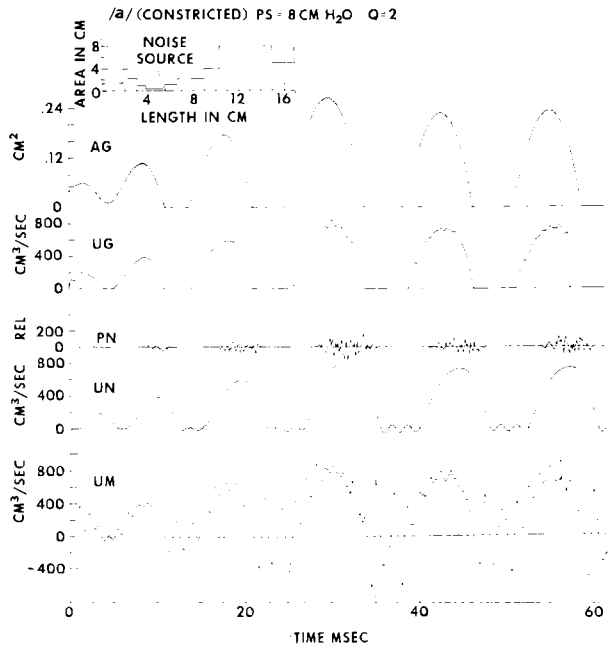

FIG. 6. Waveforms of functions calculated for a voiced fricative articulation corresponding to the constricted vowel /ɑ/.

II. VOICELESS EXCITATION

With slight modification, and with no additional control data, the synthesis can be arranged to include fricative and stop excitation. Fricative excitation is generated by turbulent air flow at a constriction, and stop excitation is produced by making a complete closure, building up pressure and abruptly releasing it. The stop release is frequently followed by a noise excitation owing to turbulence generated at the constriction after the release.

Experimental measurements indicate that the noise sound pressure generated by turbulence is proportional to the square of the Reynolds number for the flow.[7] To the extent that a one-dimensional wave treatment is valid, the noise sound pressure can be taken as

[2] Data on the behavior of the model as a function of its parameters are given in Ref. 1. Variation of fundamental frequency, mean glottal flow and glottal duty cycle are given as functions of subglottal pressure and cord tension. These calculated functions can be compared to corresponding data for natural speech in Ref. 3–6.

[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. Amer. **29**, 626–631 (1957).

[4] J. W. van den Berg, "Direct and Indirect Determination of the Mean Subglottal Pressure," Folia Phoniat. **8**, 1–24 (1956).

[5] S. Öhman and J. Lindqvist, "Analysis and Synthesis of Prosodic Pitch Contours," Speech Transmission Lab., Roy. Inst. Tech., Stockholm, Quart. Progr. Status Rep. (Apr. 1965).

[6] P. Lieberman, "Intonation, Perception and Language" (MIT Res. Monogr. 38, 1967).

[7] W. Meyer-Eppler, "Zum Erzeugungsmechanismus der Geräuschlaute," Z. Phonetik **7**, 196–212 (1953).

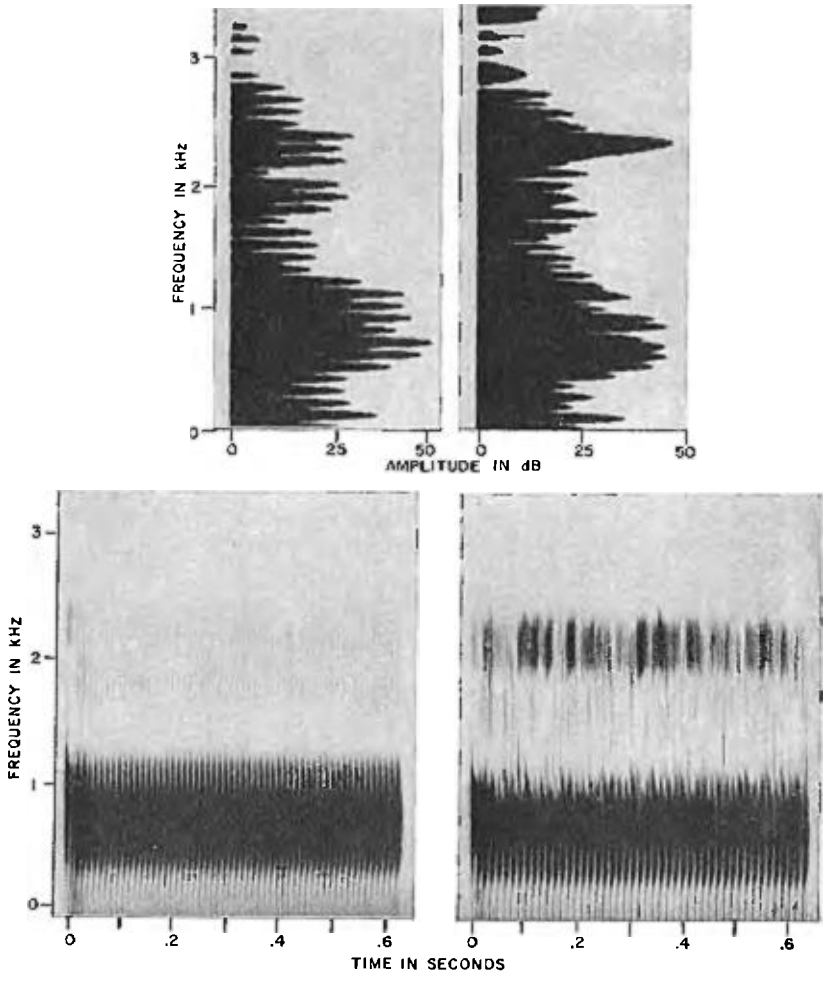

FIG. 7. Spectrograms of the synthesized output for normal vowel /a/ and the constricted /a/ shown in Fig. 6. Amplitude sections are shown for the central portion of each vowel.

proportional to the square of the volume velocity and inversely proportional to the constriction area. Measurements also suggest that the noise source is spatially distributed, but generally can be located at, or immediately downstream of the closure. Its internal impedance is primarily resistive, and it excites the vocal system as a series pressure source. Its spectrum is broadly peaked in the midaudio range and falls off at low and high frequencies.[8]

The transmission-line vocal tract can be modified to approximate the nonlinearities of turbulent flow. Figure 5 shows a single section of the transmission line so modified. A series noise source P_n, with internal resistance R_n is introduced into each section of the line. The area of the section is A_n and the volume current circulating in the right branch is U_n. The level of the noise source and the value of its internal resistance are functions of U_n and A_n. The noise source is modulated in amplitude by a function proportional to the squared Reynolds number, namely, U_n^2/A_n. The source resistance is a flow-dependent loss similar to the glotta resistance. To first order, it is proportional to $|U_n|$ and inversely proportional to A_n^2. The diagram indicates that these quantities are used to determine P_n and R_n on a sample-by-sample basis.

By continually noting the magnitudes of the volume currents in each section, and knowing the corresponding areas, the synthesizer detects conditions suitable to turbulent flow. Noise excitation and loss are therefore introduced automatically at any constriction. Small constrictions and low Reynolds numbers produce inaudible noise. The square-law dependence of P_n upon U_n has the perceptual effect of a noise threshold. (A real threshold switch can be used on the noise source, if desired.) The original control data, namely, vocal-tract shape, subglottal pressure, and cord tension, in effect, determine the place of the constriction and the loss and noise introduced there.

For the P_n source, we have used Gaussian noise, bandpassed between 500 and 4000 Hz. Also, to ensure stability, the volume flow U_n is low-pass filtered to 500 Hz before it modulates the noise source. In other

[8] J. M. Heinz, "Model Studies of the Production of Fricative Consonants," MIT Res. Lab. Electron. Quart. Progr. Rep. (15 July 1968).

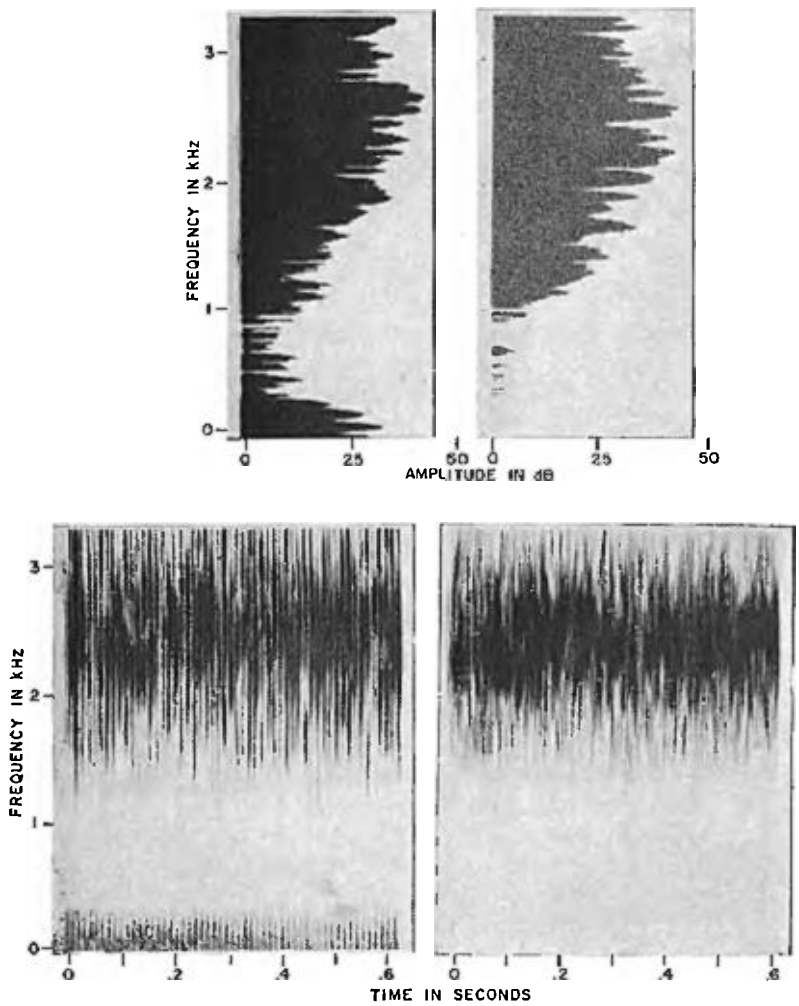

FIG. 8. Sound spectrograms for the voiced–voiceless cognates /ʒ/ and /ʃ/. Amplitude sections are shown for the central portion of each sound.

words, the noise is produced by the low-frequency components of U_n, including the dc flow.

This noise excitation works equally well for both voiced and unvoiced sounds. The operation for voiced fricatives includes all features of the formulation, and is a good vehicle for explanation. For example, consider what happens in a vowel when the constriction is made substantially smaller, giving rise to conditions favorable for turbulent flow. Since we have already shown results for the vowel /ɑ/, consider the same vowel with the constriction narrowed. (This configuration is not proposed as a realistic English sound, but merely to illustrate the effect of tightening the vowel constriction.) The situation is shown in Fig. 6. All glottal conditions are the same as before, but the constriction is narrowed to less than half the normal vowel constriction (namely, to 0.3 cm²).

The top trace shows the glottal area, and one notices that it settles to a periodic oscillation in about four periods—the final pitch here is somewhat less than that in Fig. 3 because the acoustic load is different. The second trace from the top shows the glottal flow. The glottal flow is about the same in peak value as before and is conditioned primarily by the glottal impedance and not by the tract constriction. At about the third period, noise that has been produced at the constriction by the flow buildup has propagated back to the glottis and influences the U_g flow. Note, too, that noise influence on the mechanical oscillator (i.e., the area function) is negligible.

The third trace shows the output of the noise source at the constriction. This output is proportional to the constriction current squared, divided by the constriction area. The fourth trace shows the low-passed constriction current that produces the noise. One sees that the tendency is for the noise to be generated in pitch-synchronous bursts, corresponding to the pulses of glottal volume flow. The result is a combined excitation in which the voicing and noise signals are multiplicatively related, as they are in the human.

The final trace is the volume flow at the mouth, and one can notice noise perturbations in the waveform. Note, too, that the epoch of greatest formant excitation corresponds to the falling phase of the glottal flow. A

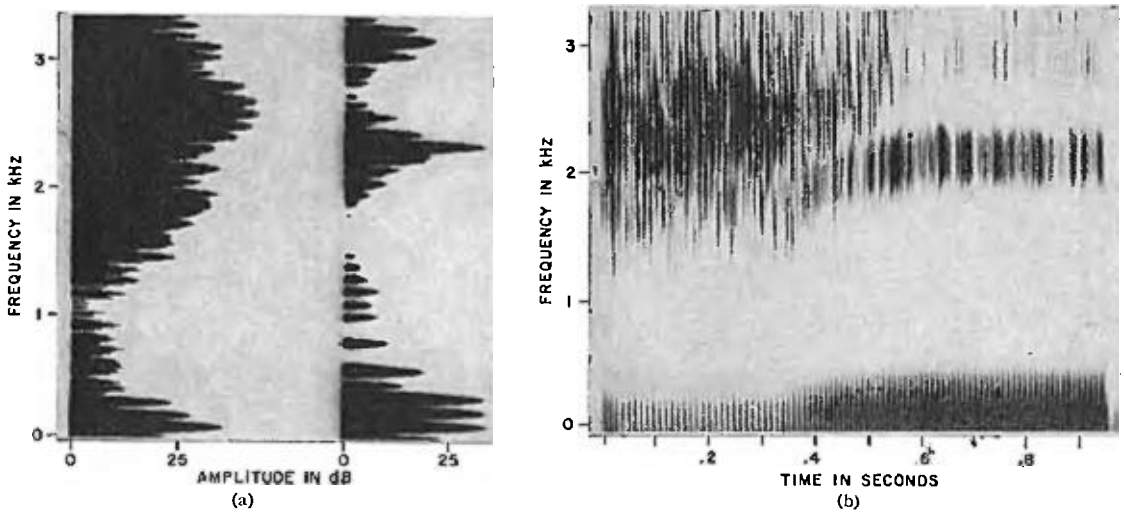

FIG. 9. Spectrogram for the synthesized syllable /ʒi/. Amplitude sections are shown for the central portion of each sound.

spectrogram of this audible output is compared with that for a normal /ɑ/ in Fig. 7. Note in the constricted, noisy /ɑ/ that: (1) the first formant has been lowered in frequency, (2) the fundamental frequency is slightly lower, and (3) pitch-synchronous noise excitation is clearly evident, particularly at the higher frequencies.

The voiceless sounds are produced simply by setting the neutral area of the vocal cords to a relatively large value, for example 1 cm². This change can be signaled by the cord tension parameter. As this is done, the Bernoulli pressure in the glottal orifice diminishes, the oscillations of the vocal cords decay, and the cord displacement assumes the steady large value. Measurements on real speech suggest this kind of effect in passing from voiced to voiceless sounds.[9,10] Corresponding exactly to this change, spectrograms of the audible output for the voiced voiceless cognates /ʒ/ and /ʃ/ are compared in Fig. 8. The vocal-tract shape is the same for both sounds. One sees a pronounced voice bar in /ʒ/ that, of course, is absent in /ʃ/. The eigenfrequencies of the two systems are similar but not exactly the same because of the difference in glottal termination. Lower resonances are not strongly evident in the /ʃ/ output, because its transmission function, from point of constriction to mouth, exhibits low-frequency zeros.

The dynamics of continuous synthesis can be illustrated by a consonant–vowel syllable. Figure 9 shows the syllable /ʒi/ synthesized by the system. In this case, the subglottal pressure and cord tension are held constant and the area function is changed linearly from the configuration for /ʒ/ to that for /i/. Heavy noise excitation is apparent during the tightly constricted /ʒ/, and the noise diminishes as the articulation shifts to /i/. Also in this case, the high, front vowel /i/ is characterized by a relatively tight constriction and a small amount of noise excitation continues in the /i/. This same effect can be seen in human speech.

The present model also appears capable of treating sounds such as glottal stops and the glottal aspiration that accompanies /h/. In the former, the tension control can cause an abrupt glottal closure and cessation of voicing. Restoration to a normal tension and quiescent glottal opening permits voicing to again be initiated. In the latter, the flow velocity and area at the glottis can be monitored just as is done along the tract. When conditions suitable for turbulence exist, a noise excitation can be introduced at the glottal location.

In summary, the work suggests that nonnasal speech (including vocal system and its excitation) can be described in terms of the physiological factors subglottal pressure, cord tension, and tract shape. Voiced sounds are produced from these data by an oscillator model of the cords. Voiceless sounds are obtained automatically, in a physiologically realistic way, from the same data. The excitation technique appears to have potential for vocal-tract synthesizers and for speech synthesis in computer answer back systems.

[9] M. Sawashima, "Observation of the Glottal Movements," Proc. Speech Symp., Kyoto (Aug. 1968), Paper C-2-1.

[10] M. Sawashima, H. Hirose, S. Kiritani, and O. Fujimura, "Articulatory Movements of the Larynx," Proc. Int. Congr. Acoust., 6th, Tokyo (Aug. 1968), Paper B-1.