A number of features seem to distinguish one speaker from another. The size and shape of the vocal tract vary considerably among persons. Characteristic damping, mouth and glottal dimensions also vary. Individual nasal coupling, size and damping of the nasal tract are other relevant features. Temporal patterns of intensity (stress) and pitch (inflection) are still others. Vocal obstructions and variations in dental work may contribute still further differences. Some or all these factors might be used to recognize or verify a speaker. It is probable that machine and human do not use the same features to equal effect. The machine, for example, might make use of data the human ear cannot assimilate.

As suggested earlier, the speech-recognition and speaker-identification experiments described here tell us little about the perceptual processing which the human accomplishes. They do not, for example, suggest the temporal span of the recognition unit used by the human. Neither do they indicate subjective techniques for measuring whether the unit is the phoneme, word, sentence, or something larger. The automatic machine methods deal mainly with advantageous processings of essentially the acoustic signal, and not with perception as the human practices it.

The mechanism of human perception of speech is difficult to analyze and present understanding is meager. The discussion of Chapter IV showed that for signals with simple temporal and spectral structure, reasonably close correlations can be made between subjective behavior and the known physiology of the peripheral ear. To a modest extent, similar relations can be established for speech signals. (For example, one can identify features such as voice pitch, formant frequency and voiced-unvoiced excitation in terms of the basilar membrane motion.) But how the neural data are stored and processed after leaving the periphery is a completely open question. Continued research on the electrophysiology of the auditory tract, and on human response to meaningful speech signals, will hopefully provide some of the answers.

VI. Speech Synthesis

Ancient man often took his ability of speech as a symbol of divine origin. Not unnaturally, he sometimes ascribed the same ability to his gods. Pagan priests, eager to fulfill great expectations, frequently tried to make their idols speak directly to the people. Talking statues, miraculous voices and oracles were well known in the Greek and Roman civilizations – the voice usually coming to the artificial mouth via cleverly concealed speaking tubes. Throughout early times the capacity of

"artificial speech" to amaze, amuse and influence its listeners was remarkably well appreciated and exploited.

As the civilized world entered the Renaissance scientific curiosity developed and expanded. Man began to inquire more seriously into the nature of things. Human life and physiological functions were fair targets of study, and the physiological mechanism of speech belonged in this sphere. Not surprisingly, the relatively complex vocal mechanism was often considered in terms of more tractable models. These early models were invariably mechanical contrivances, and some were exceedingly clever in design.

6.1. Mechanical Speaking Machines; Historical Efforts

One of the earliest documented efforts at speech synthesis was by KRATZENSTEIN in 1779. The Imperial Academy of St. Petersburg offered its annual prize for explaining the physiological differences between five vowels, and for making apparatus to produce them artificially. As the winning solution, KRATZENSTEIN constructed acoustic resonators similar in shape to the human vocal tract. He activated the resonators with vibrating reeds which, in a manner analogous to the human vocal cords, interrupted an air stream.

A few years later (1791), VON KEMPELEN constructed and demonstrated a more elaborate machine for generating connected utterances. [Apparently VON KEMPELEN's efforts antedate KRATZENSTEIN's, since VON KEMPELEN purportedly began work on his device in 1769 (VON KEMPELEN; DUDLEY and TARNÓCZY).] Although his machine received considerable publicity, it was not taken as seriously as it should have been. VON KEMPELEN had earlier perpetrated a deception in the form of a mechanical chess-playing machine. The main "mechanism" of the machine was a concealed, legless man – an expert chess player.

The speaking machine, however, was a completely legitimate device. It used a bellows to supply air to a reed which, in turn, excited a single, hand-varied resonator for producing voiced sounds. Consonants, including nasals, were simulated by four separate constricted passages, controlled by the fingers of the other hand. An improved version of the machine was built from VON KEMPELEN's description by Sir CHARLES WHEATSTONE (of the Wheatstone Bridge, and who is credited in Britain with the invention of the telegraph). It is shown in Fig. 6.1.

Briefly, the device was operated in the following manner. The right arm rested on the main bellows and expelled air through a vibrating reed to produce voiced sounds. (See the lower diagram in Fig. 6.1.) The fingers of the right hand controlled the air passages for the fricatives /ʃ/ and /s/, as well as the "nostril" openings and the reed on-off control.

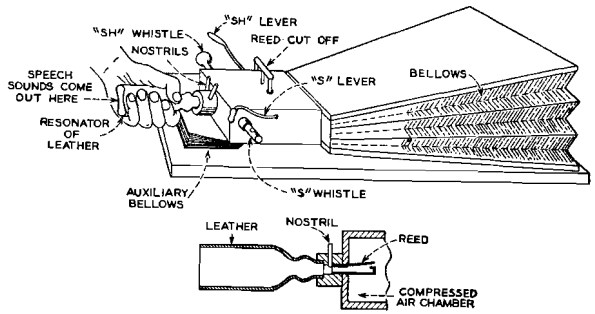

Fig. 6.1. WHEATSTONE's construction of VON KEMPELEN's speaking machine

For vowel sounds, all the passages were closed and the reed turned on. Control of vowel resonances was effected with the left hand by suitably deforming the leather resonator at the front of the device. Unvoiced sounds were produced with the reed off, and by a turbulent flow through a suitable passage. In the original work, VON KEMPELEN claimed that approximately 19 consonant sounds could be made passably well.

VON KEMPELEN's efforts probably had a more far-reaching influence than is generally appreciated. During ALEXANDER GRAHAM BELL's boyhood in Edinburgh, Scotland (latter 1800's), BELL had an opportunity to see the reproduction of VON KEMPELEN's machine which had been constructed by WHEATSTONE. He was greatly impressed with the device. With stimulation from his father (ALEXANDER MELVILLE BELL, an elocutionist like his own father), and his brother MELVILLE's assistance, BELL set out to construct a speaking automaton of his own.

Following their father's advice, the boys attempted to copy the vocal organs by making a cast from a human skull and molding the vocal parts in gutta-percha. The lips, tongue, palate, teeth, pharynx, and velum were represented. The lips were a frame-work of wire, covered with rubber which had been stuffed with cotton batting. Rubber checks enclosed the mouth cavity, and the tongue was simulated by wooden sections—likewise covered by a rubber skin and stuffed with batting. The parts were actuated by levers controlled from a keyboard. A larynx "box" was constructed of tin and had a flexible tube for a windpipe. A vocal cord orifice was made by stretching a slotted rubber sheet over tin supports.

BELL says the device could be made to say vowels and nasals and could be manipulated to produce a few simple utterances (apparently well enough to attract the neighbors). It is tempting to speculate how this boyhood interest may have been decisive in leading to U.S. patent

No. 174,465, dated February 14, 1876—describing the telephone, and which has been perhaps one of the most valuable patents in history.

BELL's youthful interest in speech production also led him to experiment with his pet Skye terrier. He taught the dog to sit up on his hind legs and growl continuously. At the same time, BELL manipulated the dog's vocal tract by hand. The dog's repertoire of sounds finally consisted of the vowels /ɑ/ and /u/, the diphthong /ou/ and the syllables /mɑ/ and /gɑ/. His greatest linguistic accomplishment consisted of the sentence, "How are you Grandmamma?" The dog apparently started taking a "bread and butter" interest in the project and would try to talk by himself. But on his own, he could never do better than the usual growl. This, according to BELL, is the only foundation to the rumor that he once taught a dog to speak.

Interest in mechanical analogs of the vocal system continued to the twentieth century. Among those who developed a penetrating understanding of the nature of human speech was Sir RICHARD PAGET. Besides making accurate plaster tube models of the vocal tract, he was also adept at simulating vocal configurations with his hands. He could literally "talk with his hands" by cupping them and exciting the cavities either with a reed, or with the lips made to vibrate after the fashion of blowing a trumpet.

Around the same time, a different approach to artificial speech was taken by people like HELMHOLTZ, D. C. MILLER, STUMPF, and KOENIG. Their view was more from the point of perception than from production. HELMHOLTZ synthesized vowel sounds by causing a sufficient number of tuning forks to vibrate at selected frequencies and with prescribed amplitudes. MILLER and STUMPF, on the other hand, accomplished the same thing by sounding organ pipes. Still different, KOENIG synthesized vowel spectra from a siren in which air jets were directed at rotating, toothed wheels.

At least one more-recent design for a mechanical talker has been put forward (RIESZ, unpublished, 1937). The arrangement is shown in Fig. 6.2. Air under pressure is brought from a reservoir at the right. Two valves, V_1 and V_2 control the flow. Valve V_1 admits air to a chamber L_1 in which a reed is fixed. The reed vibrates and interrupts the air flow much like the vocal cords. A spring-loaded slider varies the effective length of the reed and changes its fundamental frequency. Unvoiced sounds are produced by admitting air through valve V_2. The configuration of the vocal tract is varied by means of nine movable members representing the lips (1 and 2), teeth (3 and 4), tongue (5, 6, and 7), pharynx (8), and velar coupling (9).

To simplify the control, RIESZ constructed the mechanical talker with finger keys to control the configuration, but with only one control

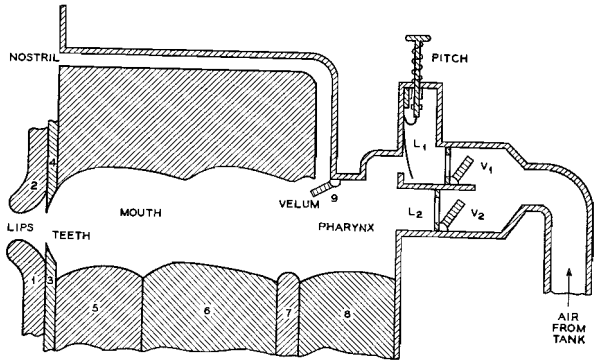

Fig. 6.2. Mechanical vocal tract of RIESZ

each for lips and teeth (i.e., members 1–2 and 3–4 of Fig. 6.2 worked as opposing pairs). The simplified arrangement with control keys is shown in Fig. 6.3. The dark surface regions indicate soft rubber linings to accomplish realistic closures and dampings. Keys 4 and 5 operate excitation valves V_4 and V_5, arranged somewhat differently from V_1 and V_2 in Fig. 6.2. Valve V_4 admits air through a hole forward in the tract (below element 6) for producing unvoiced sounds. Valve V_5 supplies air to the reed chamber for voiced excitation. In this case pitch is controlled by the amount of air passed by valve V_5. When operated by a skilled person, the machine could be made to simulate connected speech. One of its particularly good utterances was reported to be "cigarette"[1].

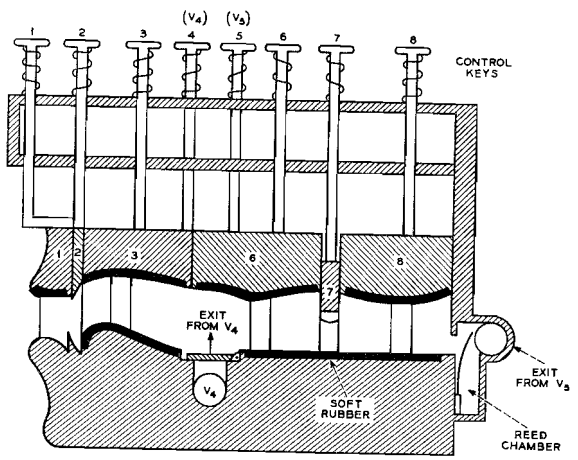

Fig. 6.3. Key control of RIESZ's mechanical talker

[1] Personal communication, R. R. RIESZ.

Interest in mechanical analogs continues to the present day. The motivation is mainly that of simulating and measuring nonlinear vocal effects. The latter are generally difficult to analyze computationally and cannot, of course, be represented with linear circuitry. One of the difficult parameters to measure in the real vocal tract is the location, intensity, spectrum, and internal impedance of the sound source for unvoiced sounds. One way of gaining knowledge about this source is with a mechanical analog. The technique for making such measurements is shown in Fig. 6.4a (HEINZ, 1958).

The size of the spherical baffle is taken to represent the human head. A constricted tube in the baffle represents the vocal tract. Air is blown through the constriction to produce turbulence. The sound radiated is measured with a spectrum analyzer. A typical spectrum obtained when the constriction is placed 4 cm from the "mouth", is plotted in Fig. 6.4b. The sound is roughly similar to the fricative /ʃ/. Because the constriction size for fricative consonants tends to be small, the spectral resonances are conditioned primarily by the cavities in front of the constriction. The antiresonances occur at frequencies where the impedance looking into the constriction from the mouth side is infinite. (Recall the discussion of Section 3.75, Chapter III.) The spectrum of the source is

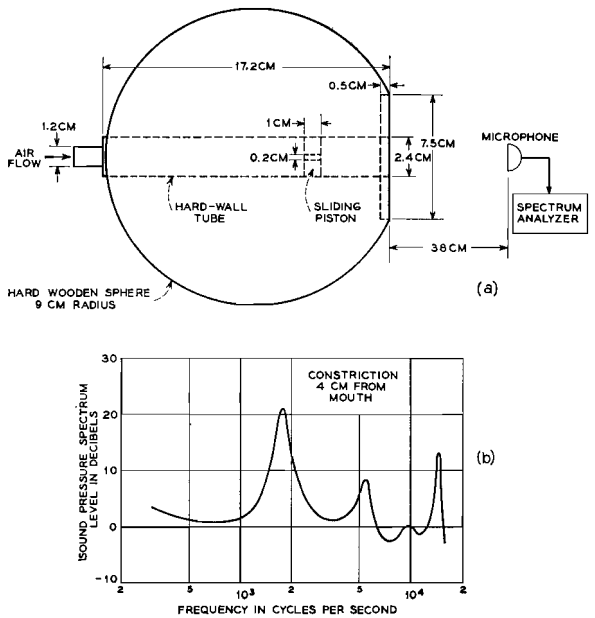

Fig. 6.4a and b. (a) Mechanical model of the vocal tract for simulating fricative consonants. (b) Measured sound spectrum for a continuant sound similar to /ʃ/. (After HEINZ, 1958)

deduced to be relatively flat. Its total power is found to be roughly proportional to the fifth power of the flow velocity.

A number of other mechanical analogs have been used in recent studies of nonlinear flow phenomena in the vocal tract (VAN DEN BERG, ZANTEMA and DOORNENBAL; MEYER-EPPLER, 1953; WEGEL). At least two of these have simulated the air flow in the glottis.

6.2. Electrical Methods for Speech Synthesis

With the evolution of electrical technology, interest in speech synthesis assumed a broader basis. Academic interest in the physiology and acoustics of the signal-producing mechanism was supplemented by the potential for communicating at a distance. Although "facsimile waveform" transmission of speech was the first method to be applied successfully (i.e., in the telephone), many early inventors appreciated the resonance nature of the vocal system and the importance to intelligibility of preserving the short-time amplitude spectrum[1]. Analytical formulation and practical application of this knowledge were longer in coming.

[1] Prominent among this group was ALEXANDER GRAHAM BELL. The events—in connection with his experiments on the "harmonic telegraph"—that led BELL, in March of 1876, to apply the facsimile waveform principle are familiar to most students of communication. Less known, perhaps, is BELL's conception of a spectral transmission method remarkably similar to the channel vocoder.

BELL called the idea the "harp telephone". It consisted of an elongated electromagnet with a row of steel reeds in the magnetic circuit. The reeds were to be arranged to vibrate in proximity to the pole of the magnet, and were to be tuned successively to different frequencies. BELL suggested that "—they might be considered analogous to the rods in the harp of Corti in the human ear". Sound uttered near the reeds would cause to vibrate those reeds corresponding to the spectral structure of the sound. Each reed would induce in the magnet an electrical current which would combine with the currents produced by other reeds into a resultant complex wave. The total current passing through a similar instrument at the receiver would, BELL thought, set identical reeds into motion and reproduce the original sound (WATSON).

The device was never constructed. The reason, WATSON says, was the prohibitive expense! Also, because of the lack of means for amplification, BELL thought the currents generated by such a device might be too feeble to be praticable. (BELL found with his harmonic telegraph, however, that a magnetic transducer with a diaphragm attached to the armature could, in fact, produce audible sound from such feeble currents.)

The principle of the "harp telephone" carries the implication that speech intelligibility is retained by preserving the short-time amplitude spectrum. Each reed of the device might be considered a combined electro-acoustic transducer and band-pass filter. Except for the mixing of the "filter" signals in a common conductor, and the absence of rectifying and smoothing means, the spectrum reconstruction principle bears a striking resemblance to that of the channel Vocoder (see Section 8.1).

6.21. Spectrum Reconstruction Techniques

Investigators such as HELMHOLTZ, D. C. MILLER, R. KOENIG and STUMPF had earlier noted that speech-like sounds could be generated by producing an harmonic spectrum with the correct fundamental frequency and relative amplitudes. In other words, the signal could be synthesized with no compelling effort at duplicating the vocal system, but mainly with the objective of producing the desired percept. Among the first to demonstrate the principle electrically was STEWART, who excited two coupled resonant electrical circuits by a current interrupted at a rate analogous to the voice fundamental. By adjusting the circuit tuning, sustained vowels could be simulated. The apparatus was not elaborate enough to produce connected utterances. Somewhat later, WAGNER devised a similar set of four electrical resonators, connected in parallel, and excited by a buzz-like source. The outputs of the four resonators were combined in the proper amplitudes to produce vowel spectra.

Probably the first electrical synthesizer which attempted to produce connected speech was the Voder (DUDLEY, RIESZ and WATKINS). It was basically a spectrum-synthesis device operated from a finger keyboard. It did, however, duplicate one important physiological characteristic of the vocal system, namely, that the excitation can be voiced or unvoiced. A schematic diagram of the device is shown in Fig. 6.5.

The "resonance control" box of the device contains 10 contiguous band-pass filters which span the speech frequency range and are connected in parallel. All the filters receive excitation from either the noise source or the buzz (relaxation) oscillator. The wrist bar selects the excitation source, and a foot pedal controls the pitch of the buzz oscillator. The outputs of the band-pass filters pass through potentiometer gain controls and are added. Ten finger keys operate the potentiometers. Three additional keys provide a transient excitation of selected filters to simulate stop-consonant sounds.

This speaking machine was demonstrated by trained operators at the World's Fairs of 1939 (New York) and 1940 (San Francisco). Although the training required was quite long (on the order of a year or more), the operators were able to "play" the machines—literally as though they were organs or pianos—and to produce intelligible speech[1]. More recently, further research studies based upon the Voder principle have been carried out (OIZUMI and KUBO).

[1] H. W. DUDLEY retired from Bell Laboratories in October 1961. On the completion of his more than 40 years in speech research, one of the Voder machines was retrieved from storage and refurbished. In addition, one of the original operators was invited to return and perform for the occasion. Amazingly, after an interlude of twenty years, the lady was able to sit down to the console and make the machine speak.

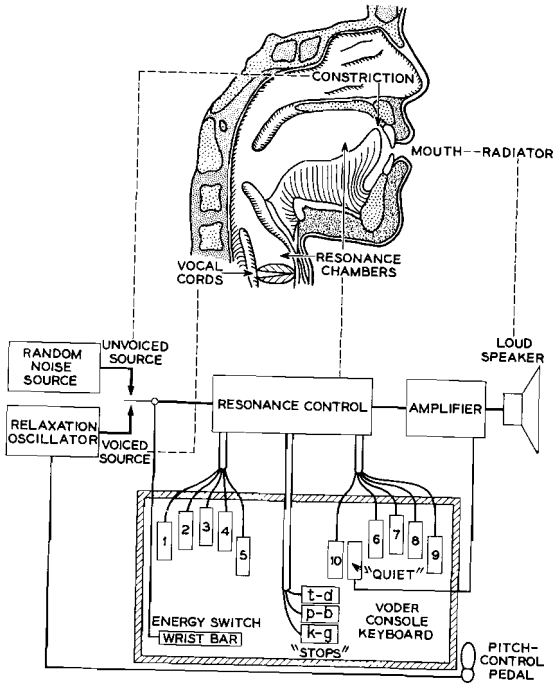

Fig. 6.5. Schematic diagram of the Voder synthesizer. (After DUDLEY, RIESZ and WATKINS)

Speech analysis by the sound spectrograph was described at some length in Chapter V. Since—as HELMHOLTZ and others observed—intelligibility is largely preserved in the short-time amplitude spectrum, speech synthesis from spectrographic plots is immediately suggested. Coupled with this notion is the question of the extent to which spectrograms of real speech might be abstracted or "caricatured" without destroying intelligibility. Several devices for automatically "playing" sound spectrograms have been designed. One uses a line source of light, parallel to the frequency axis of the spectrogram, to illuminate a variable density spectrographic pattern (SCHOTT). Contiguous photocells behind the pattern develop amplitude control signals for a set of band-pass filters (such as in the Voder). Voiced-unvoiced selection and pitch control information are represented in additional tracks. A similar scheme has been used to control a Voder-type synthesizer in an arrangement called Voback (BORST and COOPER).

A somewhat different type of spectrogram playback has been used in extensive studies on speech synthesis (COOPER; COOPER, LIBERMAN, and BORST). The speech wave is effectively simulated by a Fourier

series $\sum_n A_n \cos(n\omega_0 t + \Phi_n)$. The coefficients A_n are time varying and are determined by the spectrogram intensity at a given instant. The sound generation is accomplished by the arrangement shown in Fig. 6.6a.

The regular time-frequency-intensity pattern is illuminated by 50 contiguous light spots. The spots are sinusoidally modulated in intensity at harmonically related frequencies. The contiguous spots are produced by illuminating a "tone-wheel" with a line source. The tone wheel has 50 concentric, variable-density bands. The innermost band has four sinusoidal cycles, the next 8, the next 12, and on up to 200 for the 50th band. The tone wheel is rotated at 1800 rpm so the fundamental frequency is 120 cps. Light from the tone wheel can be either reflected from the spectrographic pattern or transmitted by it. The reflected (or transmitted) light is sensed by a collector and photocell which effectively

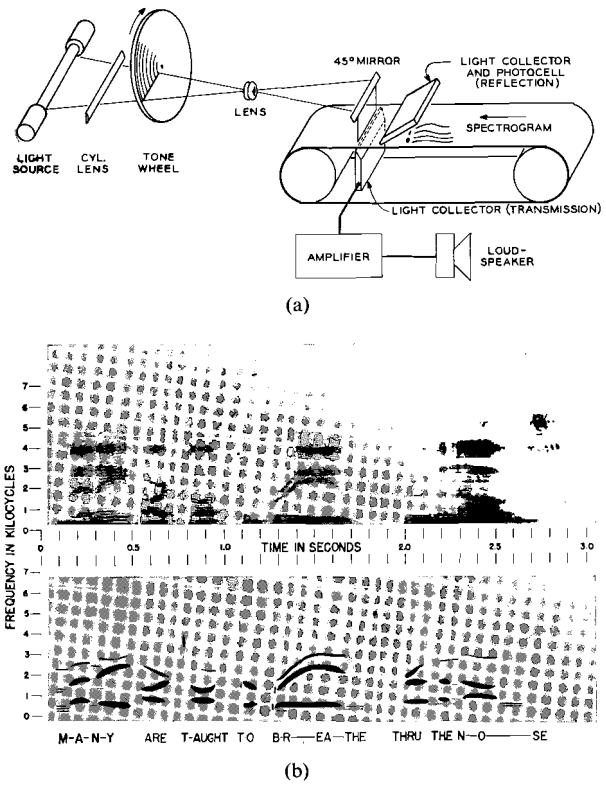

Fig. 6.6a and b. (a) Functional diagram of a spectrogram play-back device. (After COOPER.) (b) Spectrograms of real speech and an abstracted, hand-painted version of the same. Both displays can be synthesized on the pattern play-back machine. (After BORST)

sums the fifty terms of the Fourier series. The collected components are amplified and transduced.

Because of the constant rotation of the tone wheel, the pitch is monotone. Also, the phase relations of the harmonic components are fixed by the tone-wheel bands. Unvoiced sounds are simulated from a random time and intensity modulation of the frequency components – similar to the spectrographic representation of a noise burst. Spectrograms of both real speech and its abstracted version can be played on the machine. A sample of each is shown in Fig. 6.6b. In the abstracted spectrogram, in the lower part of the figure, the dark bars represent the speech formants, and the patches of fine, irregular dots produce the noise bursts. Intelligible monotone speech can be produced by the machine, and it has been used in extensive perceptual studies. Some of these results will discussed in Chapter VII.

6.22. "Terminal Analog" Synthesizers

In Chapter III linear circuit theory was applied to the acoustic analysis of the vocal tract. The results show that for simple geometries the transmission properties can be stated in a straightfoward form. Complex geometries, on the other hand, may be approximated by quantizing the vocal tube as short, abutting cylindrical sections. Effects of losses and yielding walls can be included as discussed in Section 3.73.

The tract behavior can be considered either in terms of its over-all transmission, or in terms of its detailed distributed properties. Speech synthesis may be based upon either view. The former approach attempts to duplicate – usually with a unilateral electrical circuit – the transmission properties of the tract as viewed from its input and output terminals. Synthesizers designed in this manner have, for lack of a better term, been named "terminal-analogs" (FLANAGAN, 1957c). The second view attempts to duplicate, on a one-for-one basis, the geometry and distributed properties of the tract. Electrical synthesizers designed according to this approach are bilateral, nonuniform transmission-line models of the system. The present section proposes to discuss the terminal analog approach, while the following section will treat the transmission-line device.

Both approaches to synthesis must take account of sound radiation and the vocal sources of excitation. These factors, common to both modellings of speech production, will be discussed subsequently.

6.221. Terminal Properties of the Vocal Tract. The unconstricted, glottally-excited tract can be approximated as a straight pipe, closed at the vocal cords ($Z_g=\infty$) and open at the mouth ($Z_r=0$). For such a case the results of Chapter III show that the ratio of mouth and glottal

volume velocities has a frequency-domain representation

$$\frac{U_m}{U_g}=\frac{1}{\cosh\gamma l}, \tag{6.1}$$

where l is the length of the tube, $\gamma=(\alpha+j\beta)=[(R_a+j\omega L_a)(G_a+j\omega C_a)]^{\frac{1}{2}}$ and R_a, L_a, G_a and C_a are the per-unit-length acoustical parameters of the pipe (see Fig. 3.22 and Eq. (3.61)].

It will be convenient in the subsequent discussion to treat frequency as a complex variable. Let $j\omega\to s=\sigma+j\omega$ and rewrite γ as

$$\gamma(s)=[(R_a+sL_a)(G_a+sC_a)]^{\frac{1}{2}},$$

which for low-loss conditions is

$$\gamma(s)\cong\left(\alpha+\frac{s}{c}\right),$$

where $c=1/\sqrt{L_a C_a}$ is the sound velocity [see Eq. (3.8)].

Since the vocal tract is a distributed system, its transmission characteristics involve transcendental functions. However, to represent the terminal behavior by lumped-constant electrical networks, it is necessary to describe the vocal transmission in terms of rational, meromorphic functions. Because the transcendental transfer functions for the vocal tract are meromorphic, and because their numerator and denominator components are generally integral functions (i.e., analytic for all finite values of the complex variable), it is possible to approximate the transmission by rational functions.

A relation in function theory (TITCHMARSH) says that if $f(z)$ is an integral function of the complex variable z, and meets certain restrictions, it can be represented by the product series

$$f(z)=f(0)\,e^{z\frac{f'(0)}{f(0)}}\prod_{m=1}^{\infty}\left(1-\frac{z}{a_m}\right)e^{z/a_m}, \tag{6.2}$$

where the a_m's are the ordered, simple zeros of $f(z)$.

For the vocal transmission (6.1), the zeros of the denominator (or the poles of the transmission) occur for

$$\gamma(s)=\pm j\frac{(2n-1)\pi}{2l},\qquad n=1,2,\ldots^{1}$$

[1] In Chapter III this result was written

$$\gamma=\pm j\frac{(2n+1)\pi}{2l},\qquad n=0,1,2,\ldots$$

[see Eq. (3.62)]. For the present discussion it will be convenient to write $(2n-1)$ $n=1,2,\ldots$. This has the mnemonic nicety that n may also represent the formant number.

or,

$$\gamma^2(s) = -\frac{(2n-1)^2\pi^2}{4l^2} = (R_a + sL_a)(G_a + sC_a),$$

or, dropping the subscript a's,

$$s_n = -\left(\frac{R}{2L} + \frac{G}{2C}\right) \pm j\left[\frac{(2n-1)^2\pi^2}{4l^2LC} - \left(\frac{R}{2L} - \frac{G}{2C}\right)^2\right]^{\frac{1}{2}}, \quad n = 1, 2, \ldots.$$
$$= -\sigma_n \pm j\omega_n. \tag{6.3}$$

For small loss

$$s_n \cong -\alpha c \pm j\frac{(2n-1)\pi c}{2l}, \quad n = 1, 2, \ldots. \tag{6.4}$$

which [except for the change to $(2n-1)$, $n = 1, 2, \ldots$] is the same as Eq. (3.63) in Chapter III. Substituting the result (6.3) in (6.2) gives

$$\cosh z = \prod_{n=1}^{\infty}\left[1 - \frac{z}{\pm j\frac{(2n-1)\pi}{2}}\right], \tag{6.5}$$

where $z = \gamma(s)l$. [The initial two terms of (6.2) yield unity, and the final term multiplies to unity because the roots of $f(z)$ are conjugate imaginaries.] For small loss $\gamma(s)l \cong (\alpha + s/c)l$ and

$$\frac{1}{\cosh\gamma(s)l} = \prod_n \frac{\pm j(2n-1)\pi c/2l}{s + \alpha c \pm j\frac{(2n-1)\pi c}{2l}}$$
$$= \prod_n \frac{\omega_n^2}{(s-s_n)(s-s_n^*)} \tag{6.6}$$
$$\cong \prod_n \frac{s_n s_n^*}{(s-s_n)(s-s_n^*)}$$

which is Eq. (3.64) in Chapter III.

As (6.4) indicates, the poles for the straight pipe are uniformly spaced at $\pi c/l$ intervals along the $j\omega$-axis. In this particular case, a very simple electrical circuit will realize the transmission function, namely the feedback circuit shown in Fig. 6.7. Its transmission is

$$\frac{e_0}{e_i} = H(s) = 1 - a\,e^{-sD} + a^2\,e^{-2sD} - \ldots$$
$$= \frac{1}{1 + a\,e^{-sD}}, \tag{6.7}$$

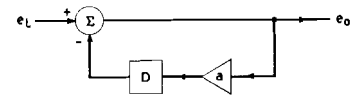

Fig. 6.7. Feedback circuit for producing a transmission having uniformly spaced complex conjugate poles

where a is a positive-real gain less than unity, and D is a simple delay equal to twice the sound transit time through the pipe. The impulse response therefore simulates the multiple reflections, with some loss, that occur at the ends of the pipe. The poles of $H(s)$ occur at

$$s_n = -\frac{1}{D}\ln\frac{1}{a} \pm j\frac{(2n-1)\pi}{D}, \quad n = 1, 2, \ldots. \tag{6.8}$$

If $D = 2l/c$ and $a = e^{-2\alpha l}$, the poles are identical to (6.4).

For a nonuniform pipe, the transmission (6.6) will generally have its poles spaced nonuniformly in frequency. In such a case, one simple way to realize the vocal transmission with electrical circuits is by "building up" the function in terms of the individual pole-pairs. This can be done by cascading individual, isolated electrical resonators, suitably tuned. This approach has the advantage of a one-to-one relation between speech formants and resonator poles, and it provides for non-interacting control of the resonances.

6.221a. Spectral Contribution of Higher-Order Poles. On perceptual grounds it is usually sufficient to simulate only the first several (three to five) modes of the tract. The remaining modes can be accounted for by a single multiplicative term representing their summated influence upon the amplitude (magnitude) spectrum (FANT, 1960). This factor, following the technique of FANT, then becomes simply a frequency-equalizing network. Assuming the higher modes to be approximately those of a straight pipe, the nature of the equalizer can be set down directly.

Write Eq. (6.6) as two product series:

$$P(s) = \prod_{n=1}^{k}\frac{s_n s_n^*}{(s-s_n)(s-s_n^*)} \cdot \prod_{n=k+1}^{\infty}\frac{s_n s_n^*}{(s-s_n)(s-s_n^*)} \tag{6.9}$$
$$= P_k(s) \cdot Q_k(s),$$

where $s_n = (-\sigma_n + j\omega_n)$. For $s = j\omega$,

$$Q_k(j\omega) = \prod_{n=k+1}^{\infty}\frac{\omega_{0n}^2}{(\omega_{0n}^2 - \omega^2) + j2\sigma_n\omega}, \tag{6.10}$$

where $\omega_{0n}^2 = (\sigma_n^2 + \omega_n^2)$.

Taking the magnitude,

$$|Q_k(j\omega)| = \prod_{n=k+1}^{\infty} \frac{\omega_{0n}^2}{[(\omega_{0n}^2-\omega^2)^2+(2\sigma_n\omega)^2]^{\frac{1}{2}}}. \tag{6.11}$$

For low loss $\sigma_n \ll \omega_n$, and

$$|Q_k(j\omega)| \cong \prod_{n=k+1}^{\infty} \frac{1}{\left(1-\frac{\omega^2}{\omega_n^2}\right)}. \tag{6.12}$$

Taking the logarithm of both sides gives

$$\ln|Q_k(j\omega)| = -\sum_{n=k+1}^{\infty} \ln\left(1-\frac{\omega^2}{\omega_n^2}\right).$$

Expanding the logarithm as a series and taking only the first term (to approximate the behavior at frequencies $\omega<\omega_n$) yields

$$\ln|Q_k(j\omega)| \cong \omega^2 \sum_{n=k+1}^{\infty} \frac{1}{\omega_n^2},$$

where

$$\omega_n = (2n-1)\omega_1 = \frac{(2n-1)\pi c}{2l}, \qquad n=1,2,\ldots$$

(that is, the modes for the straight pipe of length l). Alternatively, the logarithm may be written

$$\ln|Q_k| \cong \left(\frac{\omega}{\omega_1}\right)^2 \sum_{n=k+1}^{\infty} \frac{1}{(2n-1)^2}. \tag{6.13}$$

But

$$\sum_1^{\infty} \frac{1}{(2n-1)^2} = \frac{\pi^2}{8},$$

and the sum in (6.13) may be written

$$\sum_1^{\infty} \frac{1}{(2n-1)^2} = \frac{\pi^2}{8} - \sum_1^{k} \frac{1}{(2n-1)^2}. \tag{6.14}$$

Therefore,

$$\ln|Q_k| \cong \left(\frac{\omega}{\omega_1}\right)^2 \left[\frac{\pi^2}{8} - \sum_1^{k} \frac{1}{(2n-1)^2}\right] = \left(\frac{\omega}{\omega_1}\right)^2 [R(k)]$$

or,

$$|Q_k| \cong e^{(\omega/\omega_1)^2 R(k)}, \tag{6.15}$$

where $R(k)$ is a positive-real function of k, the highest pole accounted for on an individual basis.

6.221b. Non-Glottal Excitation of the Tract. The discussion of Chapter III showed that if the vocal excitation occurs at some point other than at the end of the tract, the transmission function will exhibit zeros as well as poles. This can be simply illustrated for front excitation of a straight pipe by a pressure source, as shown in Fig. 6.8. The ratio of

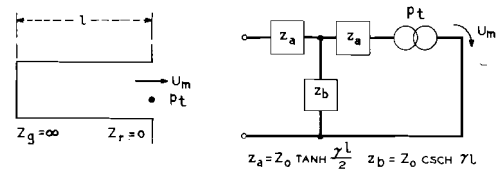

Fig. 6.8. Front excitation of a straight pipe by a pressure source

mouth current to the source pressure is simply the driving point impedance of the mouth, or

$$\begin{aligned}\frac{U_m(s)}{p_t(s)} &= \frac{1}{Z_0}\tanh\gamma(s)\,l \\ &= \frac{1}{\cosh\gamma(s)\,l}\cdot\frac{\sinh\gamma(s)\,l}{Z_0} \\ &= P(s)\cdot Z(s).\end{aligned} \tag{6.16}$$

Since $P(s)$ has no zeros, the zeros of the transmission are the zeros of $Z(s)$ and occur for

$$(e^{2\gamma l}-1)=0$$

$$\gamma = \pm j\frac{m\pi}{l}, \qquad m=0,1,2,\ldots \tag{6.17}$$

$$\gamma^2 = \frac{-m^2\pi^2}{l^2} = [(R+sL)(G+sC)].$$

The zeros therefore lie at

$$s_m = -\left(\frac{R}{2L}+\frac{G}{2C}\right) \pm j\left[\frac{m^2\pi^2}{l^2LC} - \left(\frac{R}{2L}-\frac{G}{2C}\right)^2\right]^{\frac{1}{2}},$$

or, again for small losses,

$$s_m \cong \left(-\alpha c \pm j\frac{m\pi c}{l}\right) \qquad m=0,1,2,\ldots. \tag{6.18}$$

The poles of the transmission are the same as given in Eq. (6.4), and the poles and zeros in this instance alternate in the $j\omega$-direction.

Applying the product series formula in Eq. (6.2) gives

$$\sinh z = z \prod_{m=1}^{\infty} \left(1 - \frac{z}{\pm j m \pi}\right),$$

where

$$z = \gamma l \cong \left(\alpha l + s \frac{l}{c}\right). \tag{6.19}$$

Then

$$\begin{aligned}\sinh \gamma l &= \left(\alpha l + s\frac{l}{c}\right) \prod_{m=1}^{\infty} \left(1 - \frac{\alpha l + s\frac{l}{c}}{\pm j m \pi}\right) \\ &= \frac{l}{c}(\alpha c + s) \prod_{m=1}^{\infty} \left(\frac{s + \alpha c \pm j m \pi c}{\pm j \frac{m \pi c}{l}}\right) \\ &\cong \frac{l}{c}(s + s_0) \prod_{m=1}^{\infty} \frac{(s - s_m)(s - s_m^*)}{s_m s_m^*},\end{aligned} \tag{6.20}$$

where $s_0 = -\alpha c$.

6.221c. Spectral Contribution of Higher-Order Zeros. The series for the zero terms can be "truncated" as described previously for pole terms, and a spectral correction factor can be obtained for higher-order zeros. Following the technique of Eq. (6.9),

$$Z(s) \cong \frac{l}{c Z_0}(s + s_0) \prod_{m=1}^{k} \frac{(s - s_m)(s - s_m^*)}{s_m s_m^*} \cdot |Y_k(s)|,$$

where

$$\begin{aligned}\ln |Y_k(j\omega)| &\cong -\sum_{m=k+1}^{\infty} \frac{\omega^2}{\omega_m^2} \\ &\cong -\frac{\omega^2}{\omega_1^2} \sum_{m=k+1}^{\infty} \frac{1}{m^2},\end{aligned} \tag{6.21}$$

and where $\omega_1 = \pi c / l$.

The summation may be rewritten as

$$\ln |Y_k(j\omega)| \cong -\frac{\omega^2}{\omega_1^2}\left[\frac{\pi^2}{6} - \sum_{m=1}^{k} \frac{1}{m^2}\right],$$

or

$$|Y_k(j\omega)| \cong e^{-(\omega^2/\omega_1^2)\, T(k)}, \tag{6.22}$$

where $T(k)$ is a positive-real function of the zero number k. Except for the sign of the exponent, this is the same form as (6.15). The factor $|Y_k(j\omega)|$ can therefore be realized by a frequency-equalizing network in conjunction with the variable poles and zeros of a formant synthesizer.

This simple example of front excitation illustrates that the vocal transmission, in general, involves poles $[P(s)]$ as well as zeros $[Z(s)]$. In the example, the zeros (like the poles) are uniformly distributed in frequency. For the nonuniform vocal tract, the mode frequencies will generally be irregularly distributed. Besides being dependent upon source location, zeros of transmission can also arise from side-branch paths coupled to the main transmission path. Cases in point are nasal consonants, nasalized vowels and perhaps liquids such as /l/[1]. In all cases where the sound radiation is from a single port (i.e., either mouth or nostril), the vocal transmission is minimum phase. For simultaneous radiation from mouth and nostril (as in a nasalized vowel) the transmissions to individual ports are minimum phase, but the combined response at a fixed point in front of the speaker may be nonminimum phase.

6.221d. Effects of a Side-Branch Resonator. The effect of a nasal or oral side branch can be simply illustrated by the circuit of Fig. 6.9a. For very low frequencies the circuit may be treated in terms of lumped-constant approximations to the major cavities and constrictions, as illustrated in Fig. 6.9b. The poles occur at frequencies where the sum of the admittances at any network node is zero. The velar junction is a convenient point to consider. Neglecting losses, the respective admit-

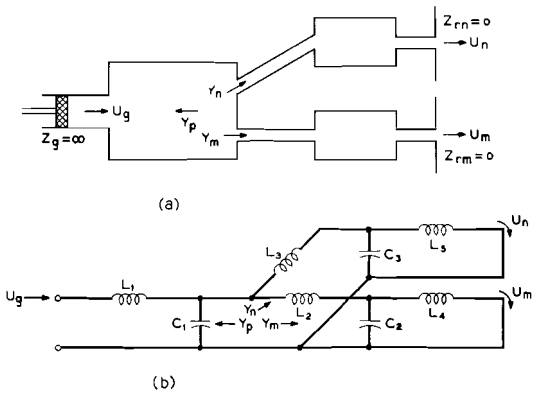

Fig. 6.9a and b. Simplified configuration illustrating coupling between oral and nasal cavities

[1] The cul-de-sac formed by the tongue can act as a side-branch resonator.

tances for the low-frequency approximation are

$$Y_n=\frac{s^2+\dfrac{1}{L_5 C_3}}{s L_3\left[s^2+\dfrac{1}{C_3}\left(\dfrac{1}{L_3}+\dfrac{1}{L_5}\right)\right]}$$

$$Y_m=\frac{s^2+\dfrac{1}{L_4 C_2}}{s L_2\left[s^2+\dfrac{1}{C_2}\left(\dfrac{1}{L_2}+\dfrac{1}{L_4}\right)\right]} \tag{6.23}$$

$$Y_p=s C_1,$$

or for real frequencies $s \to j\omega$,

$$Y_n=\frac{\omega_{n0}^2-\omega^2}{j\omega L_3(\omega_{np}^2-\omega^2)}$$

$$Y_m=\frac{\omega_{m0}^2-\omega^2}{j\omega L_2(\omega_{mp}^2-\omega^2)} \tag{6.24}$$

$$Y_p=j\omega C_1,$$

where ω_{n0} and ω_{m0} are the zeros of the nasal and mouth admittances respectively, and ω_{np} and ω_{mp} are the poles of the nasal and mouth admittances.

The poles of the system occur at frequencies for which

$$\Sigma Y = Y_n + Y_m + Y_p = 0$$

or

$$\omega^2 C_1=\frac{\omega_{n0}^2-\omega^2}{L_3(\omega_{np}^2-\omega^2)}+\frac{\omega_{m0}^2-\omega^2}{L_2(\omega_{mp}^2-\omega^2)}. \tag{6.25}$$

The low-frequency zero of U_n/U_g is ω_{mp}, and the zero of U_m/U_g is ω_{np}.

It is instructive to consider the loci of the low frequency modes for a highly simplified situation. Suppose the pharyngeal, oral and nasal cavities (C_1, C_2, C_3) are held fixed in size, and the mouth and velar constrictions (L_2, L_3, L_4) are varied. Suppose the velar areas are such that $(A_n+A_m)=A_0=$constant, so that L_2 and L_3 are inversely related. Assume that all tube lengths are held fixed so that area variations alone constitute the lumped element variation. Consider the low frequency mode behavior corresponding to the sequence: vowel → nasalized vowel → nasal, as in /ɑm/. The simplified articulatory sequence is: vowel, with the nasal tract decoupled and sealed off and the mouth open; nasalized vowel, with the velum partially open and the mouth still open; and nasal, with the velum full open and the mouth closed.

For the vowel, the nasal coupling is nil and $L_3 \cong \infty$. The frequencies ω_{n0} and ω_{np} are equal (i.e., the pole and zero are coincident) and $Y_n=0$. The poles of the glottis-to-mouth transmission occur at frequencies where $Y_m=Y_p$. As the vowel is nasalized, the velum opens, L_3 diminishes and L_2 increases. ω_{n0} remains fixed, but ω_{np} parts from ω_{n0} and moves up in frequency. ω_{np} becomes the zero of glottis-to-mouth transmission. In a similar manner ω_{m0} remains fixed, but ω_{mp} moves down. The exact trajectories of the system modes depend upon the relative sizes of the nasal and oral cavities, but, in general, the original vowel poles move up in frequency. A new pole is introduced in the region above ω_{n0} by the parting of ω_{n0} and ω_{np}.

As the mouth closes to produce the nasal, L_4 becomes infinite and all sound radiation transfers to the nostril. The closed oral cavity now acts as a side branch resonator for the glottis-to-nostril transmission. ω_{m0} now goes to zero, and ω_{mp} becomes lower. ω_{mp} is the zero of glottis-to-nostril transmission. The first system pole is relatively low in frequency, and the second resides in the vicinity of ω_{mp}. The third is generally somewhat higher than ω_{np}. A more detailed computation, using an idealized vocal configuration, has been given previously in Fig. 3.37. Representative frequency positions for a nasal such as /m/ are approximately 250, 1100, 1350 and 2000 cps for the first four poles and 1300 cps for the zero. More extensive analyses of nasals can be found in the literature (FUJIMURA, 1962a).

So long as the radiation is from a single port, the *dc* transmission to that port is essentially unity. For simultaneous radiation from mouth and nostril, the sound energy divides according to the oral and nasal admittances, and the *dc* transmission to a single port is determined by the respective branch losses.

6.222. Cascade Type Synthesizers. The intent of these elementary considerations is to indicate that for all configurations and excitations, the vocal transmission $T(s)$ may be approximated in terms of its first few (low-frequency) poles and zeros, that is, the first several roots of $P(s)$ and $Z(s)$. A straightforward means for simulating the vocal transmission electrically is to build up the product functions in terms of the individual poles and zeros by cascading individual electrical resonators. As the preceding discussion showed, the transmission function for a vowel sound can be written

$$T(s)=P(s)=\prod_n \frac{s_n s_n^*}{(s-s_n)(s-s_n^*)}.$$

Such a function can be represented in terms of its individual poles by the isolated, cascaded, series RLC resonators shown in Fig. 6.10a. Here the

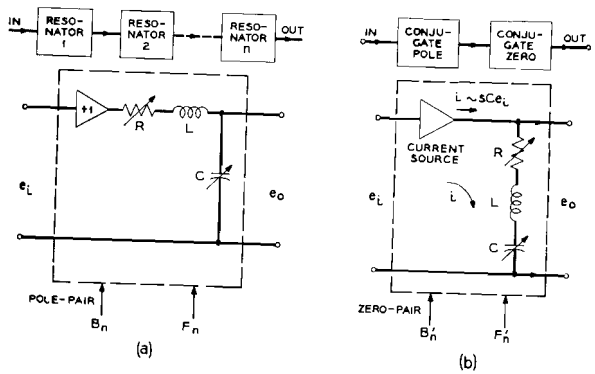

Fig. 6.10a and b. (a) Cascade connection of isolated RLC resonators for simulation of vocal transmission for vowel sounds. Each pole-pair or vocal resonance is simulated by a series circuit. (b) Cascaded pole and zero circuit for simulating low frequency behavior of a side branch resonator. The zero pair is approximated by the transmission of a simple series circuit

transmission of a single resonant circuit is

$$\frac{e_0(s)}{e_i(s)}=\frac{\frac{1}{LC}}{s^2+\frac{R}{L}s+\frac{1}{LC}}=\frac{s_n s_n^*}{(s-s_n)(s-s_n^*)},$$

where

$$\omega_n=\sqrt{\frac{1}{LC}-\frac{R^2}{4L^2}},$$
$$\sigma_n=\frac{R}{2L}, \tag{6.26}$$

and

$$s_n=-\sigma_n+j\omega_n.$$

Control of the formant tuning is effected by changes in the tuning capacitor C. Control of formant bandwidth is accomplished by variation of R. For the serial connection of resonators, specification of the pole frequencies s_n implies specification of the spectral peaks, or formant amplitudes, as well. This point has been treated in some detail in the literature (FANT, 1956; FLANAGAN, 1957c).

The results of Chapter III and the preceding discussion (Fig. 6.9) suggest that sounds such as unvoiced consonants, nasals, nasalized vowels, and perhaps liquids, may have at least one low-frequency zero

which might be perceptually significant[1]. In particular, a pole-zero pair additional to the usual vowel formants is commonly associated with nasals and nasalized vowels. The transmission of the vowel resonator string of Fig. 6.10a can be simply modified to accomodate this condition. A resonance and an antiresonance—as shown in the upper part of Fig. 6.10b—can be included in the synthesizer circuit (FLANAGAN, COKER and BIRD). So long as a pure vowel is to be produced, the added pole and zero are made coincident in frequency, and their transmission is unity. For nasal production they are pulled apart and set to appropriate values corresponding to the relations for the side branch resonator.

Practically, the complex conjugate zero can be approximated by the electrical circuit shown in the lower part of Fig. 6.10b. Its transmission is

$$\frac{e_0(s)}{e_i(s)}=LC\left(s^2+s\frac{R}{L}+\frac{1}{LC}\right), \tag{6.27}$$

which is the reciprocal of the conjugate pole. As in the pole-pair resonator, the low frequency (dc) gain is made unity—which is proper so long as radiation occurs from a single port, and is approximately correct for the mouth radiation of nasalized vowels.

The front-excited voiceless sounds can also be approximated in terms of their poles and zeros. Following the results of the previous discussion and of Chapter III, a reasonable approximation is given by

$$T(s)=P(s)\cdot Z(s)=K\cdot s\cdot\frac{\prod_m(s-s_m)(s-s_m^*)}{\prod_n(s-s_n)(s-s_n^*)}, \tag{6.28}$$

where an m and n of order 1 or 2 often suffice perceptually (in addition to higher-order pole and zero corrections). The zero at zero frequency arises because of the essentially closed back cavity (see Fig. 3.31). The amplitude scale factor K is accounted for by an over-all amplification.

6.223. Parallel Synthesizers. The vocal tract transmission has been represented as a ratio of product series which, when truncated, produce rational meromorphic functions. Because the poles are simple, the transmission can be expanded as a partial fraction with first-degree terms

$$T(s)=P(s)Z(s)=\sum_n\frac{A_n}{(s-s_n)}+\frac{A_n^*}{(s-s_n^*)},\qquad n=1,2,\ldots$$
$$=\sum_n\frac{2a_n s+2(\sigma_n a_n+\omega_n b_n)}{s^2+s\sigma_n s+(\sigma_n^2+\omega_n^2)}, \tag{6.29}$$

[1] The perceptual effects of spectral zeros—both of the excitation and of the system—have not been throughly established. The extent to which the quality of synthetic speech depends upon these factors is a current problem in research. It will be discussed further in a later section.

where $s_n=(-\sigma_n+j\omega_n)$, and $A_n=(s-s_n)T(s)|_{s\to s_n}=(a_n+jb_n)$ is the residue in the n-th pole and is a function of all the poles and zeros. The inverse transform is

$$h(t)=\sum_n 2|A_n|e^{-\sigma_n t}\cos(\omega_n t+\varphi_n),$$

where

$$A_n=|A_n|e^{j\varphi_n}.$$

Expanding the cosine term, $h(t)$ may be rewritten

$$h(t)=\sum_n 2|A_n|e^{-\sigma_n t}[\cos\varphi_n\cos\omega_n t-\sin\varphi_n\sin\omega_n t]. \tag{6.30}$$

Each term of the latter expression can be realized by the operations shown in Fig. 6.11, where the filters represented by the boxes are simple resonant circuits.

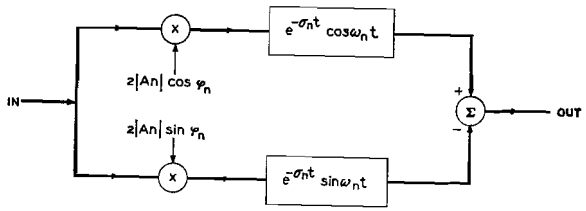

Fig. 6.11. Circuit operations for simulating the time-domain response of Eq. (6.30)

If the transmission function is for pure vowels, $Z(s)\to 1$ and $T(s)\to P(s)$ and the transmission has only poles. Its numerator is not a function of s, but only of the s_n, that is, $\prod_n s_n s_n^*=f(s_n)$. The residue in the q-th pole is then

$$A_q=\frac{f(s_n)}{2j\omega_q\prod_{n\neq q}[(\sigma_n-\sigma_q)^2+(\omega_n^2-\omega_q^2)+2j\omega_q(\sigma_q-\sigma_n)]}. \tag{6.31}$$

If the σ's are essentially equal (a reasonable approximation for the lower modes of the vocal tract), then

$$A_q\cong\frac{f(s_n)}{2j\omega_q\prod_{n\neq q}(\omega_n^2-\omega_q^2)},$$

or

$$A_q\cong\frac{f(s_n)}{2j\omega_q(-1)^{q-1}}\frac{1}{\prod_{n\neq q}|\omega_n^2-\omega_q^2|}. \tag{6.32}$$

The residues are therefore pure imaginary (i.e., $\cos\varphi_n=0$) and their signs alternate with pole number. The inverse transform (impulse response) for this transmission is

$$h(t)=\sum_n(-1)^{n-1}2|A_n|e^{-\sigma_n t}\sin\omega_n t, \tag{6.33}$$

where each term can by synthesized by the electrical operations in Fig. 6.12. This circuit is essentially the lower branch of the previous circuit where now $-\sin\varphi_n=-\sin[(-1)^n(\pi/2)]=(-1)^{n-1}$, and the RCL resonator has an impulse response $(\omega_n e^{-\sigma_n t}\sin\omega_n t)$. Summation of the outputs of similar circuits, one for each n, produces the response (6.33).

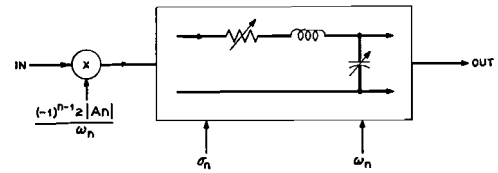

Fig. 6.12. Circuit for simulating the vowel function impulse response [see Eq. (6.33)]

The magnitude of the residue bears a simple approximate relation to the spectral magnitude at the formant frequency. Recall the residue magnitude is

$$|A_n|=|(s-s_n)T(s)|_{s\to s_n},$$

which for small damping ($\sigma\ll\omega_n$) is approximately

$$|(s-s_n)T(s)|_{s\to j\omega_n}=|(j\omega_n-s_n)T(j\omega_n)|\approx|A_n|,$$

or,

$$\sigma_n|T(j\omega_n)|\approx|A_n|. \tag{6.34}$$

If the transmission function exhibits zeros, as exemplified by Eq. (6.28), the residues are then

$$\begin{aligned}A_q'&=(s-s_q)T(s)|_{s\to s_q}=Z(s)\cdot(s-s_q)P(s)|_{s\to s_q}\\&=Z(s_q)\cdot A_q=Ks_q[\prod_m(s_q-s_m)(s_q-s_m^*)]A_q\\&=A_qKs_q\cdot\prod_m[(\sigma_q-\sigma_m)^2+(\omega_m^2-\omega_q^2)+j2\omega_q(\sigma_q-\sigma_m)].\end{aligned} \tag{6.35}$$

Again, if the σ's are nearly the same,

$$A_q'=A_qKs_q\cdot\prod_m(\omega_m^2-\omega_q^2), \tag{6.36}$$

and the sign of A'_q is determined by the relative magnitudes ω_m and ω_q. Or,

$$A'_q = A_q(-1)^p K s_q \prod_m |\omega_m^2 - \omega_q^2|, \tag{6.37}$$

where p is the number of zeros lying below the pole ω_p. Or, substituting for A_q from Eq. (6.32),

$$A'_q = \frac{f(s_n)(-1)^p K s_q \prod_m |\omega_m^2 - \omega_q^2|}{2j\,\omega_q(-1)^{q-1} \prod_{n \neq q} |\omega_n^2 - \omega_q^2|}, \tag{6.38}$$

and the net sign of the residue is determined by the difference between the numbers of poles and zeros lying below the q-th pole.

Again the residue bears a simple approximate relation to the real-frequency spectrum evaluated at the pole frequency. That is,

$$A_n = (s - s_n)\, T(s)|_{s \to s_n},$$

but for low damping $s_n \to j\omega_n$.

$$\begin{aligned} A_n &\cong (j\,\omega_n - s_n)\, T(j\,\omega_n) \\ A_n &\cong \sigma_n\, T(j\,\omega_n) = \sigma_n |T(j\,\omega_n)| \underline{/T(j\,\omega_n)} \\ A_n &= |A_n|\, e^{j\varphi_n}. \end{aligned} \tag{6.39}$$

A number of terminal-analog synthesizers, both of the parallel and cascade types, have been constructed and operated. (See for example, FANT, 1959a; STEVENS, BASTIDE and SMITH; LAWRENCE, 1953; STEAD and JONES; CAMPANELLA; CHANG; FLANAGAN, 1956a, 1960b.) Most of the devices utilize one or more of the relations discussed – either by overt recognition of the principles or by qualitative implication. The transmission relations commonly exploited involve the formant frequency and the magnitude of the residue, or the formant frequency and amplitude.

At least one study has considered use of the complex residue, that is, the angle or sign of the residue. In this case, analysis of the short-time phase spectrum of speech[1] – in conjunction with the short-time amplitude spectrum – is used to gain naturalness. Specification of the complex residues, as implied by Eq. (6.29), is equivalent to specification of spectral zeros. A parallel formant synthesizer, implemented as described by Eq. (6.30) and using pitch-synchronous spectral analysis to obtain formant frequency and *complex* residue, produced speech of improved quality (FLANAGAN, 1965).

[1] See Eq. (5.4), Chapter 5, for a definition of the short-time phase spectrum.

6.23. Transmission-Line Analogs of the Vocal System

A different method for simulating the vocal transmission is the nonuniform electrical transmission line. The discussion in Chapter III indicated how the nonuniform acoustic tubes of the vocal and nasal tracts can be represented by abutting right-circular cylinders (see Fig. 3.35). The approximation to the nonuniform tract is better the more numerous the cylindrical elements.

Each cylindrical section of length l can be represented by its T-equivalent as shown in Fig. 6.13a, where $z_a = Z_0 \tanh \gamma l/2$ and $z_b = Z_0 \operatorname{csch} \gamma l$.

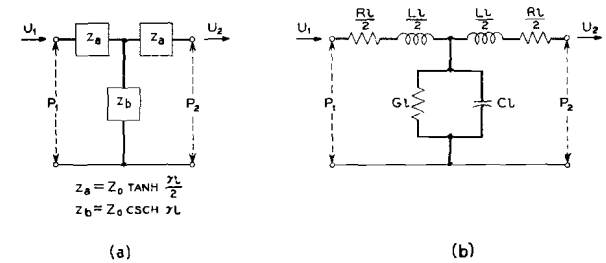

Fig. 6.13a and b. T-circuit equivalents for a length l of uniform cylindrical pipe. (a) Exact circuit, (b) first-term approximations to the impedance elements

A practical electrical realization of the individual T-section is obtained by taking the first terms in the series expansions of the hyperbolic quantities. For a hard-walled tube this gives $z_a \cong \frac{1}{2}(R + j\omega L)\, l$ and $z_b \cong 1/(G + j\omega C)\, l$ where the R, L, G and C are the per-unit-length acoustic parameters of the tube, as previously discussed. The resulting network is Fig. 6.13b[1].

For practical realization, the characteristic impedance of the analogous electrical line may be scaled from the acoustic value by a convenient constant, i.e., $Z_{0e} = k Z_{0a}$, where the superscripts e and a distinguish electrical and acoustical quantities. For low-loss conditions, $Z_{0a} \cong \sqrt{L_a/C_a} = \rho c/A$. Since $L_a = \rho/A$ and $C_a = A/\rho c^2$, a given simulated cross-sectional area is equal $\rho c \sqrt{C_a/L_a}$. The losses R and G require knowledge of the circumference as well as the cross-sectional area of the tract [see Eq. (3.33)]. They can also be introduced into the electrical circuit and their impedances scaled after the fashion just indicated. Given the shape factor, all analogous electrical elements can be determined from the A and l data-pairs, or from area data for a given number of fixed-length cylindrical sections.

[1] Section 3.73 derives a technique for including the effects of a yielding wall.

A vocal tract representation in terms of equivalent electrical sections forms the ladder networks of Fig. 6.14. The upper circuit is for glottal excitation of the tract by a volume-velocity source U_g and with internal impedance Z_g. The lower circuit is for forward fricative excitation by a pressure source P_t with internal impedance Z_t. Both circuits can be solved—at least in principle—by straightforward matrix methods. If voltage (pressure) equations are written for each circuit loop, beginning at the glottis and ending at the mouth and nose, the number of independent equations is equal the number of loops. The transmissions from

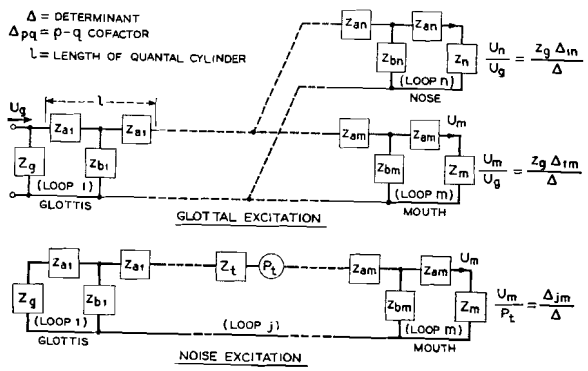

Fig. 6.14. Ladder network approximations to the vocal tract. The impedance elements of the network are those shown in Fig. 6.13b

glottis to mouth, from glottis to nostril, and from front noise source to mouth are, respectively,

$$\frac{U_m}{U_g}=\frac{Z_g\Delta_{1m}}{\Delta}$$

$$\frac{U_n}{U_g}=Z_g\frac{\Delta_{1n}}{\Delta} \tag{6.40}$$

$$\frac{U_m}{P_t}=\frac{\Delta_{jm}}{\Delta},$$

where Δ is the impedance determinant (characteristic equation) for the network having impedance members z_{11}, z_{12}, etc., where z_{11} is the self-impedance of loop 1, z_{12} is the mutual impedance between loops 1 and 2, etc., and Δ_{xy} is the cofactor of the x-th row and y-th column of the determinant Δ. As mentioned earlier, all the transmissions of Eq. (6.40) are minimum phase functions[1].

[1] The functions are the responses of passive ladder networks. They can have zeros of transmission only for zeros of a shunt element or for poles of a series element. All these poles and zeros must lie in the left half of the complex-frequency plane.

Several electrical transmission-line synthesizers have been constructed. The first such device consisted of 25 uniform T-sections (DUNN, 1950). Each section represented a tract length of 0.5 cm and a nominal area of 6 cm^2. A variable inductance could be inserted between any two sections to simulate the tongue construction. Another variable inductance at the mouth end of the line represented the lip constriction. Radiation from the mouth was simulated by taking the output voltage across a small series inductance. For voiced sounds, the synthesizer was excited by a high-impedance sawtooth oscillator whose fundamental frequency could be controlled. The source spectrum was adjusted to fall at about -12 db/octave (recall Fig. 3.17). To simulate unvoiced and whispered sounds, a white noise source was applied at an appropriate point along the line.

At least two other passive line analogs, similar to DUNN's device, have been constructed (STEVENS, KASOWSKI and FANT; FANT, 1960). These synthesizers incorporate network sections which can be varied independently to simulate the tract geometry in detail. At least one effort has been made to develop a continuously-controllable transmission-line analog. Continuous variation of the network elements by electronic means permits the device to synthesize connected speech (ROSEN; HECKER). This device utilizes saturable-core inductors and electronically-variable capacitors as the line elements. A nasal tract is also provided. The number of network sections and their control points are shown in Fig. 6.15. Control of the synthesizer can be effected either from an electronic data-storage circuit (ROSEN) of from a digital computer (DENNIS).

The transmission-line synthesizer has outstanding potential for directly incorporating the constraints that characterize the vocal mechanism. Success in this direction, however, depends directly upon deriving a realistic model for the area and for the dynamic motions of the vocal tract. Research on one such model has been described in Section 5.4. Also, the usefulness of a transmission-line synthesizer in a

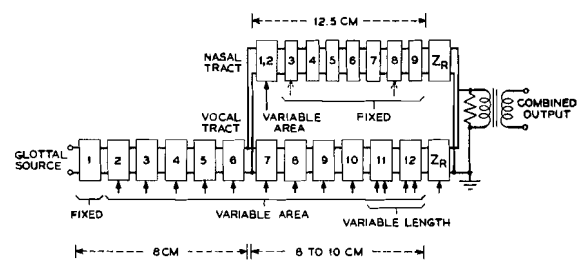

Fig. 6.15. Continuously controllable transmission line analog of the vocal system. (After ROSEN; HECKER)

complete analysis-synthesis system depends upon how accurately vocal tract area data, or its equivalent, can be derived automatically from connected speech. Some progress has been made toward analyzing speech signals in articulatory terms from which area and length numbers can be derived (see Section 5.4, Chapter 5).

Besides obvious application in a bandwidth compression system, the transmission-line synthesizer, along with other synthesis devices, has potential use as a computer output device for man-machine communication; as a stimulus generator for psychoacoustic and bioacoustic experimentation; or, as a standard sound generator for speech pathology, therapy or linguistics studies. The possibility of specifying the control functions in articulatory terms makes applications such as the latter particularly attractive.

All transmission-line synthesizers of early design have been implemented as analog network devices. Digital techniques, on the other hand, offer many advantages in stability and accuracy. One of the first digital transmission-line synthesizers was programmed on a computer in terms of the reflection coefficients at the junctions of cylindrical tube-elements (KELLY and LOCHBAUM).

Another computer implementation has duplicated the bilateral properties of the transmission line by a difference-equation equivalent. Because absolute impedance relations are preserved in this formulation, it has been useful in studying the acoustic interaction between the vocal tract and the vocal cords. The same formulation has also been used as a complete synthesizer for voiced and unvoiced sounds (FLANAGAN and LANDGRAF; FLANAGAN and CHERRY).

Further discussion of digital representation of transmission-line synthesizers is given in Section. 6.26.

6.24. Excitation of Electrical Synthesizers

The preceding sections have discussed simulation of the vocal transmission both from the transfer-function point of view and from the transmission-line approach. Having implemented one or the other for electrical synthesis of speech, the system must be excited from signal sources analogous to those of the vocal tract. This section considers vocal source characteristics that appear relevant in synthesis.

6.241. Simulation of the Glottal Wave. The results of Chapter III suggested that the vocal cord source is approximately a high-impedance, constant volume-velocity generator. Hence, to a first-order approximation, the vocal tract and glottal source can be assumed not to interact greatly. To the extent that this is true (and we shall subsequently discuss

this matter further), the source and system can be analyzed independently, and their characteristics can be simulated individually.

The shape and periodicity of the vocal cord wave can vary considerably. This is partially illustrated by the single periods of glottal area and volume-velocity waves shown in Fig. 6.16. The extent to which variability in period and shape affect speech naturalness and quality is an important research question. In many existing electrical synthesizers, the properties of the vocal cord source are approximated only in a gross form. It is customary to specify the vocal pitch as a smooth, continuous time function and to use a fixed glottal wave shape whose amplitude spectrum falls at about −12 db/octave. In many synthesizers the source is produced by repeated impulse excitation of a fixed, spectral-shaping network. Such lack of fidelity in duplicating actual glottal characteristics

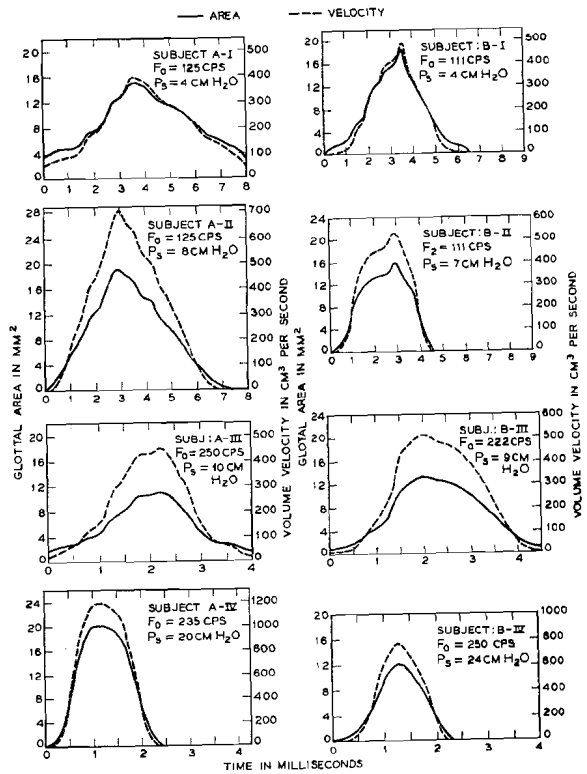

Fig. 6.16. Single periods of measured glottal area and calculated volume velocity functions for two men (A and B) phonating the vowel /æ/ under four different conditions of pitch and intensity. F_0 is the fundamental frequency and P_s the subglottal pressure. The velocity wave is computed according to the technique described in Section 3.52. (After FLANAGAN, 1958)

undoubtedly detracts from speech naturalness and the ability to simulate a given voice.

6.241a. Spectral Properties of Triangular Waves. Under some conditions of voicing (commonly, mid-range pitch and intensity), the glottal wave is roughly triangular in shape. The spectral properties of triangular waves therefore appear to have relevance to voiced excitation. They have been studied in some detail with a view toward better understanding the relations between waveform and spectrum in real glottal waves (DUNN, FLANAGAN and GESTRIN)[1].

Fig. 6.17 shows a triangular approximation to the glottal wave. The opening time is τ_1, the closing time $\tau_2 = k\tau_1$, and the total open time

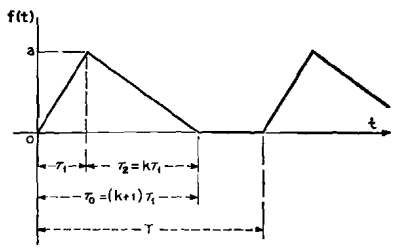

Fig. 6.17. Triangular approximation to the glottal wave. The asymmetry factor is k

$\tau_0 = (1+k)\tau_1$. The amplitude of the wave is a and its period T. Its Laplace transform is

$$F(s) = \frac{a}{s^2}\left[\frac{1}{\tau_1} - \left(\frac{1}{\tau_1} + \frac{1}{\tau_2}\right) e^{-s\tau_1} + \frac{1}{\tau_2} e^{-s(\tau_1+\tau_2)}\right]. \tag{6.41}$$

The spectral zeros are the complex values of s which make $F(s)=0$. Except for the $s=0$ root, the zeros are the roots of the bracketed expression, or the roots of

$$[e^{-(k+1)s\tau_1} - (k+1)e^{-s\tau_1} + k] = 0. \tag{6.42}$$

Because the equation is transcendental it can be solved exactly only for special values of the asymmetry constant, k. In particular, solutions are straightforward for values of k which can be expressed as the ratio of small whole numbers. In less simple cases, the roots can be obtained by numerical solution.

[1] It should be emphasized again that the implication here is not that the glottal pulse is a neat triangular wave, but only that this analytical simplification permits tractable and informative calculations. These data are included because they are not avaliable elsewhere.

Let

$$x = e^{-s\tau_1} = e^{-(\sigma+j\omega)\tau_1} = e^{-\sigma\tau_1}(\cos\omega\tau_1 - j\sin\omega\tau_1). \tag{6.43}$$

The (6.42) becomes

$$x^{(k+1)} - (k+1)x + k = 0. \tag{6.44}$$

When k is an integer, (6.44) will yield $(k+1)$ values of x. These can then be put into (6.43), and both $\sigma\tau_1$ and $\omega\tau_1$ found by equating real and imaginary parts in separate equations.

For integers up to $k=5$, (6.44) can be solved by straightforward algebraic methods. In the case $k=5$, (6.44) is a sixth degree equation in x, but a double root exists at $x=1$, and a fourth degree equation is left when these are removed. For higher values of k, roots can be approximated by known methods.

However, k need not be an integer. Suppose only that it is a rational number (and it can always be approximated as such). Then $(k+1)$ is also rational. Let

$$k+1 = \frac{p}{q}, \tag{6.45}$$

where p and q are positive integers, and $p \geqq q$, since k cannot be less than zero. Then (6.44) can be written

$$x^{\frac{p}{q}} - \frac{p}{q}x + \frac{p-q}{p} = 0. \tag{6.46}$$

Let $y = x^{1/q}$, so that (6.46) becomes

$$y^p - \frac{p}{q}y^q + \frac{p-q}{q} = 0, \tag{6.47}$$

and by (6.43)

$$y = e^{-\frac{1}{q}\sigma\tau_1}\left(\cos\frac{1}{q}\omega\tau_1 - j\sin\frac{1}{q}\omega\tau_1\right). \tag{6.48}$$

Eq. (6.47) has integer exponents, and can be solved for y. Then (6.48) can be solved for

$$\frac{1}{q}\sigma\tau_1 \quad \text{and} \quad \frac{1}{q}\omega\tau_1,$$

which need only to be multiplied by p to get $\sigma\tau_0$ and $\omega\tau_0$.

The preceding methods become awkward when p is larger than 6. The following is more suitable for numerical approximation by digital computer. Equating the real and imaginary parts of (6.42) separately to zero gives the equations

$$e^{-(k+1)\sigma\tau_1}\cos(k+1)\omega\tau_1 - (k+1)e^{-\sigma\tau_1}\cos\omega\tau_1 + k = 0, \tag{6.49}$$

$$e^{-(k+1)\sigma\tau_1}\sin(k+1)\omega\tau_1 - (k+1)e^{-\sigma\tau_1}\sin\omega\tau_1 = 0. \tag{6.50}$$

Both of these equations must be satisfied by the pair of values of $\sigma\tau_1$ and $\omega\tau_1$ which represent a zero. Eq. (6.50) can be solved for $\sigma\tau_1$

$$\sigma\tau_1=\frac{1}{k}\log\frac{\sin(k+1)\omega\tau_1}{(k+1)\sin\omega\tau_1}. \tag{6.51}$$

A series of values of $\omega\tau_1$ is put into (6.51) and the $\sigma\tau_1$ computed for each. Each pair of values is substituted into (6.49) to find those which satisfy it. The solutions can be approximated as closely as desired by choosing suitably small increments of $\omega\tau_1$, and by interpolation. A modest amount of computation time on a digital computer produces the first half-dozen roots.

6.241b. Repetition and Symmetry of the Zero Pattern. Let ω be the imaginary part of a zero that (together with its real part σ) simultaneously satisfies (6.49) and (6.50). Also let k be related to integers p and q as in (6.45). Consider another imaginary part ω' such that

$$\omega'\tau_1=2q\pi+\omega\tau_1.$$

Then

$$\omega'\tau_0=(k+1)\omega'\tau_1=\frac{p}{q}\omega'\tau_1=2p\pi+(k+1)\omega\tau_1. \tag{6.52}$$

Both sines and cosines of $\omega'\tau_1$ and $(k+1)\omega'\tau_1$ are the same as those of $\omega\tau_1$ and $(k+1)\omega\tau_1$. Hence, with no change in σ, ω' also represents a zero. The pattern of zeros between $\omega\tau_0=0$ and $\omega\tau_0=2p\pi$ will be repeated exactly in each $2p\pi$ range of $\omega\tau_0$, to infinity, with an unchanged set of σ's.

Again supposing ω is the imaginary part of a zero, let ω' be a frequency such that

$$\omega'\tau_1=2q\pi-\omega\tau_1. \tag{6.53}$$

Then

$$\omega'\tau_0=(k+1)\omega'\tau_1=2p\pi-(k+1)\tau_1. \tag{6.54}$$

Now the cosines of $\omega'\tau_1$ and $(k+1)\omega'\tau_1$ are the same as those of $\omega\tau_1$ and $(k+1)\omega\tau_1$, while the sines are both of opposite sign. Both (6.49) and (6.50) will still be satisfied, and ω' represents a zero having the same σ as that of ω. In each $2p\pi$ interval of $\omega\tau_0$, the zeros are symmetrically spaced about the center of the interval (an odd multiple of $p\pi$), each symmetrical pair having equal values of σ. There may or may not be a zero at the center of symmetry, depending upon whether p is odd or even.

6.241c. Zeros of the Reversed Triangle. If $f(t)$ is the triangular wave, then $f(-t)$ is the wave reversed in time, and

$$\mathscr{L}[f(t)]=F(s)$$

and,

$$\mathscr{L}[f(-t)]=F(-s). \tag{6.55}$$

Therefore, the zeros of the reversed triangle are the negatives of those for the original triangle. Since the zeros of the original triangle occur in complex conjugate pairs, the reversed triangle has the same zeros as the original triangle, but with the signs of the real parts reversed.

Also, the asymmetry constant for the reversed triangle, is $1/k$, where k is the asymmetry of the original triangle.

6.241d. Zeros of the Right Triangle. When $k=0$, the triangle is right and has a transform

$$F(s)=\frac{a}{s^2\tau_0}[1-e^{-s\tau_0}(1+s\tau_0)]. \tag{6.56}$$

Its zeros occur for

$$(1+s\tau_0)=e^{s\tau_0}. \tag{6.57}$$

Equating real and imaginary parts,

$$1+\sigma\tau_0=e^{\sigma\tau_0}\cos\omega\tau_0, \tag{6.58}$$

$$\omega\tau_0=e^{\sigma\tau_0}\sin\omega\tau_0. \tag{6.59}$$

[Note the solution $\omega=0$, $\sigma=0$ cannot produce a zero because of the s^2 in the denominator of (6.56).]

As before, the roots can be approximated numerically with the computer. Note that with σ and ω real, and taking only positive values of ω, $\sin\omega\tau_0$ is positive according to (6.59). Also, since $\omega\tau_0$ is larger than $\sin\omega\tau_0$, $\sigma\tau_0$ must be positive and the real parts of the zeros must be positive, or they must lie in the right half s-plane. Then by (6.58) $\cos\omega\tau_0$ is also positive which means that all zeros must occur for $\omega\tau_0$ in the first quadrant.

For $k=\infty$, the triangle is also right, but reversed in time. Its zeros are therefore the same as those for $k=0$, but with the signs of the real parts reversed.

6.241e. Loci of the Complex Zeros. Using the foregoing relations, enough zeros have been calculated to indicate the low-frequency behavior of the triangular wave. A complex-frequency plot of the zero loci—normalized in terms of $\omega\tau_0$ and $\sigma\tau_0$ and with the asymmetry k as the parameter—is shown in Fig. 6.18. In this plot the asymmetry is restricted to the range $0\leq k\leq 1$. For $k>1$, these loci would be mirrored in the vertical axis, that is, the signs of σ would be reversed.

For the symmetrical case ($k=1$), the zeros are double and fall on the $j\omega$-axis at even multiples of 2π; i.e., at 4π, 8π, 12π, etc. They are re-

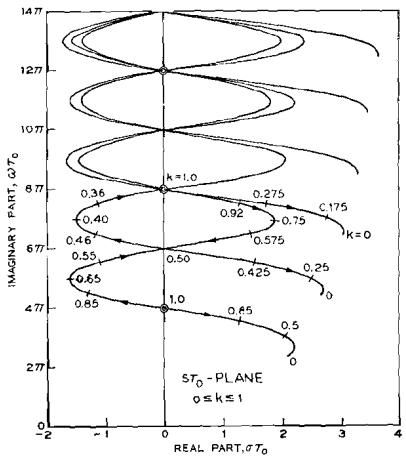

Fig. 6.18. Complex frequency loci of the zeros of a triangular pulse. The s-plane is normalized in terms of $\omega\tau_0$ and $\sigma\tau_0$. The asymmetry constant k is the parameter. (After DUNN, FLANAGAN and GESTRIN)

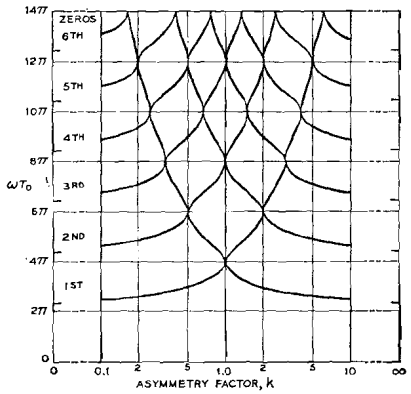

Fig. 6.19. Imaginary parts of the complex zeros of a triangular pulse as a function of asymmetry. The imaginary frequency is normalized in terms of $\omega\tau_0$ and the range of asymmetry is $0 \leqq k \leqq \infty$. (After DUNN, FLANAGAN and GESTRIN)

presented by the small concentric circles at these points. In terms of cps, the double zeros lie at $2/\tau_0$, $4/\tau_0$, etc., and the amplitude spectrum is $(\sin^2 x/x^2)$. As k is made smaller than unity, the double zeros part—one moving initially into the right half plane and the other into the left. Their paths are plotted.

As the order of the zero increases, the s-plane trajectory also increases in length and complexity for a given change in k. A given reduction in k from unity causes the first zero to move into the right half plane where it remains. The same change in k may cause a higher order zero, say the sixth, to make several excursions between right and left half planes. For the first, second and third zeros, values of k from 1.0 to 0.0 are laid off along the paths. For $k=0$, the triangle is right, with zero closing time, and all zeros have terminal positions in the right half plane. Note, too, that in the vicinity of the $j\omega$-axis, a relatively small change in symmetry results in a relatively large change in the damping of the zeros.

All imaginary-axis zeros are double and the degree of the zeros never exceeds two. This point is further emphasized in a plot of the loci of the imaginary parts of the zeros as a function of the asymmetry factor k. The pattern is shown in Fig. 6.19. It is plotted for values of k between 0.1 and 10. All points of tangency represent double $j\omega$-axis

zeros. The average number of zeros is one per every 2π interval of $\omega\tau_0$. The pattern of imaginary parts is symmetrical about the $k=1$ value, with the right and left ordinates showing the zeros of the right triangles, i.e., for $k=0$ and $k=\infty$.

To illustrate the sensitivity of the amplitude spectrum to a specific change in the asymmetry constant, Fig. 6.20 shows amplitude spectra $|F(j\omega)|$ for two values of asymmetry, namely, $k=1$ and $k=11/12$ (or 12/11). For $k=1$ the zeros are double and are spaced at cps frequencies of $2/\tau_0$, $4/\tau_0$, $6/\tau_0$, etc. The spectrum is $\sin^2 x/x^2$ in form. A change in k to 11/12 (or to 12/11) causes each double zero to part, one moving into the right half plane and the other into the left. Their $j\omega$-positions

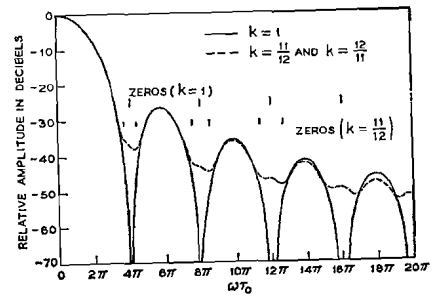

Fig. 6.20. Amplitude spectra for two triangular pulses, $k=1$ and $k=11/12$. (After DUNN, FLANAGAN and GESTRIN)

are indicated by the ticks on the diagram. The increase in real parts is such as to provide the spectral "fill" indicated by the dotted curve. In this case a relatively small change in symmetry results in a relatively large spectral change.

6.241 f. Other approximations to the Glottal Pulse. The preceding comments have exclusively concerned triangular approximations to the glottal wave. In reality the glottal wave can take on many forms, and it is instructive to consider the zero patterns for other simple approximations. The triangle has three points where slope is discontinuous. What, for example, might be the effect of eliminating one or more of these discontinuities by rounding or smoothing the wave?

There are several symmetrical geometries that might be considered realistic approximations to glottal waves with more rounding. Three, for example, are pulses described respectively by a half (rectified) sine wave, a half ellipse, and a raised cosine. The waveforms are plotted in the top part of Fig. 6.21. The first two have two points of discontin-

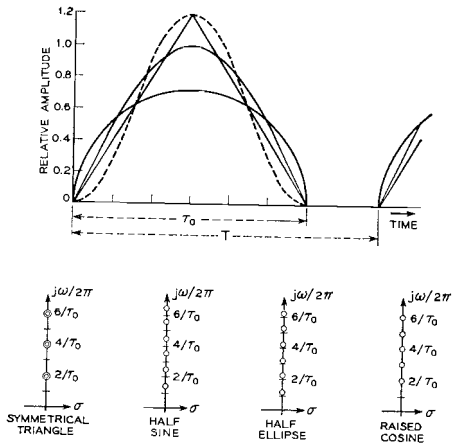

Fig. 6.21. Four symmetrical approximations to the glottal pulse and their complex zeros

uous slope; the latter has none. They can be described temporally and spectrally as follows.

Half-sine wave

$$f(t)=a\sin\beta t, \quad 0\leqq t\leqq\frac{\pi}{\beta}, \quad \beta=\frac{\pi}{\tau_0}$$
$$=0, \quad \text{elsewhere} \tag{6.60}$$
$$F(\omega)=\left(\frac{\beta a}{\beta^2-\omega^2}\right)(1+e^{-j\pi\omega/\beta}),$$

where the zeros occur at:

$$\omega=\pm\frac{(2n+1)\pi}{\tau_0}=\pm(2n+1)\beta, \quad n=1,2,\ldots^{1}.$$

Half-ellipse

$$f(t)=\frac{4}{\pi\tau_0}\left[1-\left(\frac{2t}{\tau_0}\right)^2\right]^{\frac{1}{2}}, \quad |t|\leqq\tau_0/2$$
$$=0, \quad \text{elsewhere} \tag{6.61}$$
$$F(\omega)=\frac{2J_1(\omega\tau_0/2)}{\omega\tau_0/2},$$

where, except for $\omega=0$, the zeros occur at the roots of $J_1(\omega\tau_0/2)$.

[1] For all these symmetrical waves, the zeros lie on the $j\omega$-axis.

Raised Cosine

$$f(t)=a(1-\cos\beta t), \quad 0\leqq t\leqq\frac{2\pi}{\beta}, \quad \beta=\frac{2\pi}{\tau_0}$$
$$=0, \quad \text{elsewhere} \tag{6.62}$$
$$F(\omega)=a\left[\frac{\beta^2}{j\omega(\beta^2-\omega^2)}\right][1-e^{-j2\pi\omega/\beta}],$$

and the zeros occur at:

$$\omega=\pm n\beta=\pm\frac{2n\pi}{\tau_0}, \quad n=2,3,\ldots.$$

The complex zeros for these functions are plotted in the lower part of Fig. 6.21. The plots suggest that relatively small changes in rounding and pulse shape can have appreciable influence upon the zero pattern and upon the low-frequency behavior of the glottal spectrum. Although the zeros may shift around, the average number of zeros in a given frequency interval (above a frequency of about $1/\tau_0$) still remains the same for all the waves, namely one per $1/\tau_0$ cps[1].

6.241g. Asymptotic Density of Source Zeros. This average density of zeros also holds at high frequencies. Consider an arbitrary glottal pulse, $f(t)$, which is finite and nonzero in the interval $0<t<\tau_0$ and zero elsewhere. Since $\int_0^{\tau_0} f(t)\,e^{-st}dt$ must be finite, the function can have no poles. Suppose the second derivative of $f(t)$ is bounded inside the same interval and that the slope is discontinuous at $t=0$ and $t=\tau_0$. Except at $s=0$, two differentiations of $f(t)$ leave the zeros the same, and produce impulses of areas $f'(0_+)$ and $f'(\tau_{0-})$ at the leading and trailing edges of the pulse. The transform of the twice-differentiated pulse is therefore

$$s^2F(s)=\int_0^{\infty} f''(t)\,e^{-st}\,dt=f'(0_+)+f'(\tau_{0-})\,e^{-s\tau_0}+\int_{0+}^{\tau_{0-}} f''(t)\,e^{-st}\,dt.$$

Since $f''(t)$ is bounded in $0<t<\tau_0$, the integral of the third term must be of order $1/s$ or less. At high frequencies it becomes small compared to the first two terms and the transform is approximately

$$s^2F(s)\cong[f'(0_+)+f'(\tau_{0-})\,e^{-s\tau_0}],$$

[1] The spectra given here are for single pulses, that is, continuous spectra given by the Laplace or Fourier transforms of the pulses. For periodically repeated pulses, the spectra are discrete harmonic lines whose amplitudes are given by $(1/T)\,F(m\Omega_0)$, where $F(m\Omega_0)$ is the Fourier transform of a single pulse evaluated at the harmonic frequencies $m\Omega_0=m\,2\pi/T$, $m=1,2,3\ldots$.

with zeros at

$$s=-\frac{1}{\tau_0}\ln\left|\frac{f'(0_+)}{f'(\tau_{0-})}\right|\pm j\frac{(2n+1)\pi}{\tau_0},\qquad n=0,1,\ldots. \tag{6.63}$$

At low frequencies, however, the zero positions may be much more irregular, as the previous computations show.

6.241h. Perceptual Effects of Glottal Zeros. A relevant question concerns the effect of glottal zeros in real speech. Are they perceptually significant? Should they be taken into account in speech analysis techniques such as spectral pattern matching? Are they important for synthesizing natural speech? The complete answers to these questions are not clear and comprehensive subjective testing is needed. It is clear, however, that under particular conditions (which can sometimes be identified in sound spectrograms), a glottal zero may fall proximate to a speech formant and may alter both the spectrum and the percept.

The formant nullifying potential of a glottal zero can easily be demonstrated in synthetic speech. Fig. 6.22 shows a four-resonance vowel synthesizer circuit. The circuit is excited by an approximately symmetrical, triangular glottal wave. The amplitude spectra actually measured with a wave analyzer are shown for two conditions of open time of the glottal wave. The vowel is /ʌ/. In case (*A*), the open time is chosen to position the first double glottal zero near to the first formant ($\tau_0\cong 4$ msec). In case (*B*), the first glottal zero is positioned between the first and second formants ($\tau_0\cong 2.5$ msec). The relative pole-zero positions are shown for the first two formants in the *s*-plane diagrams. The first formant peak is clearly suppressed and flattened in the first

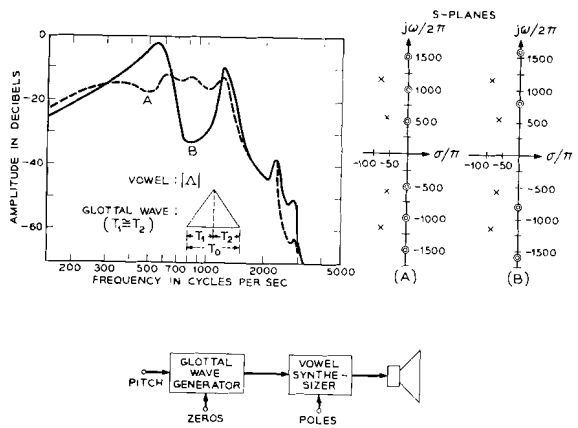

Fig. 6.22a and b. Effect of glottal zeros upon the measured spectrum of a synthetic vowel sound. (a) $\tau_0=4.0$ msec. (b) $\tau_0=2.5$ msec. (After FLANAGAN, 1961b)

case[1]. A significant difference in vowel quality is obvious in listening to the two conditions.

If an even more artificial situation is posed, the effect of source zeros can be made still more dramatic. For example, suppose the synthesizer is set for the vowel /ə/ which has nearly uniformly-spaced poles. Suppose also that the excitation is brief, double pulses described by $f(t)=a(t)+b(t-\delta)$, where $a(t)$ and $b(t)$ are impulses with areas a and b, respectively. The frequency transform of $f(t)$ is $F(s)=(a+be^{-s\delta})$ which has zeros at

$$s=\left[-\frac{1}{\delta}\ln\frac{a}{b}\pm j\frac{(2n+1)\pi}{\delta}\right],\qquad n=0,1,\ldots. \tag{6.64}$$

That is, this excitation produces the same zero pattern as the asymptotic high frequency spacing given in Eq. (6.63). By suitable choice of a/b and δ, the source zeros can be placed near the formants. Three different excitation conditions (including a single pulse) are shown in three columns in Fig. 6.23. The input excitation and the resulting synthetic sound waveforms are also shown. In the first case the vowel is clearly heard and identified as /ə/. In the second and third cases, the vowel quality and color are substantially altered. Cases 2 and 3 differ very little perceptually, although the sound waveforms are greatly different. From the perceptual standpoint there appears to be a relatively narrow vertical strip, centered about the $j\omega$-axis, in which a glottal zero has the potential for substantially influencing the percept[2]. The double pulse excitation provides a simple means for manipulating the zero pattern for subjective testing. Also, to a very crude approximation, it is somewhat similar to the phenomenon of diplophonia (SMITH, S.).

As emphasized earlier in this section, the perceptual importance of glottal wave detail and of source zeros has not been thoroughly established. At least one speech analysis procedure, however, has taken glottal zeros into account to obtain more precise spectral analyses (MATHEWS, MILLER and DAVID, 1961b). A pole-zero model, with an average zero density of one per $1/\tau_0$ cps, is fitted in a weighted-least-square sense to real speech spectra (see Section 5.21). A typical pole-zero fit to the spectrum of a single pitch period of a natural vowels is shown in Fig. 6.24. The analysis procedure does not discriminate between right and left half-plane zeros, and all zeros are plotted in the left

[1] In neither case does the measured amplitude spectrum go to zero at the frequency of the zeros. The laboratory-generated glottal wave was not precisely symmetrical and its zeros did not lie exactly on the $j\omega$-axis.

[2] Symmetric glottal pulses produce zeros on the $j\omega$-axis, as described in the preceding discussion. In natural speech this region appears to be largely avoided through vocal-cord adjustments.

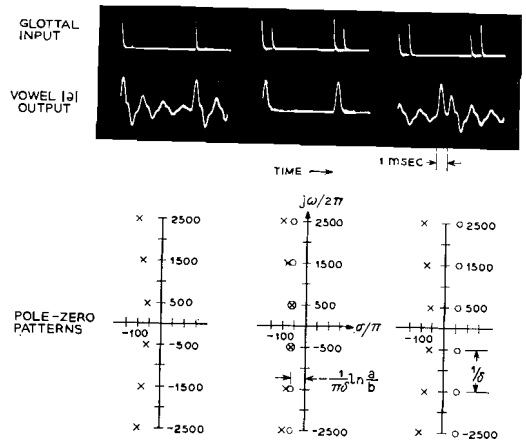

Fig. 6.23

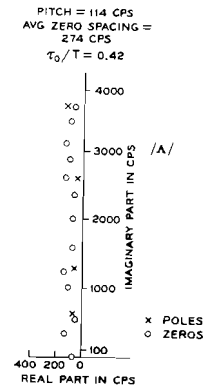

Fig. 6.24

Fig. 6.23. Method for manipulating source zeros to influence vowel quality. Left column, no zeros. Middle column, left-half plane zeros. Right column, right-half plane zeros. (After FLANAGAN, 1961 b)

Fig. 6.24. Best fitting pole-zero model for the spectrum of a single pitch period of a natural vowel sound. (After MATHEWS, MILLER and DAVID, 1961 b)

half-plane. An open time of the glottal wave of about 0.4 times the pitch period is suggested by the result.

Whether the precise positions of source zeros are perceptually significant remains a question for additional study. Only their influence on over-all spectral balance and gross shape may be the important factor. The vocal excitation may vary in waveform so rapidly in connected speech that the zero pattern is not stationary long enough to influence the percept. A speaker also might adjust his glottal wave by auditory feedback to minimize unwanted suppression of formant frequencies.

One experiment leads to the view that the glottal wave can be represented by a fixed analytical form, and that period-to-period irregularities in the pitch function can be smoothed out (ROSENBERG, 1971). Natural speech was analyzed pitch-synchronously. Pitch, formant frequencies and an inverse-filter approximation to the glottal wave were determined for each period. The glottal wave shape was "cartoonized" and characterized by fixed, smooth, analytical functions, whose glottis-open times depended only upon pitch period[1]. Using the analyzed pitch and formant data, the speech was synthesized with this artificial characterization of the glottal wave. Listening tests were then conducted.

[1] Note that the spectral zeros of such waves vary in frequency position as the fundamental frequency changes. Only for monotone pitch are the spectral zeros constant in position.

Subjects preferred asymmetric wave characterizations with one slope discontinuity (corresponding to cord closure) and with opening and closing times equal to 40% and 16% of the pitch period. The subjects were relatively insensitive to variations in the precise shape and open-close times. Very small opening or closing times, and approximately equal opening and closing times were clearly not preferred. The latter, as discussed above, leads to spectral zeros near the $j\omega$-axis. The results also demonstrated that elimination of fine temporal detail in the glottal wave shape does not degrade speech quality. These results appear consistent with data on factors found important in formant-synthesized speech (HOLMES, 1961).

Another experiment, using the same analysis techniques, determined the amount of averaging of pitch and formant data that is perceptually tolerable in synthetic speech (ROSENBERG, 1971). In the vowel portions of syllables in connected speech, averaging over as much as four to eight pitch periods did not degrade quality. This averaging completely eliminated fine detail (period-to-period fluctuations) in the pitch and formant data. Longer averaging, which modified the underlying pitch and formant trajectories, did definitely impair quality.

Acoustic interaction between the vocal cords and vocal tract contributes some temporal details to the glottal volume flow waveform. This interaction also influences the temporal variation of voice pitch. These experiments suggest that the fine structure, both in wave shape and in pitch-period variation, is not perceptually significant, but that variations in values averaged over several pitch periods are significant.

One point should perhaps be emphasized in considering inverse-filter estimates of glottal wave shape. The fundamental hypothesis is that the source and system are linearly separable, and that the acoustic properties of each can be uniquely assigned. The glottal wave is usually obtained from the inverse filter according to some criterion such as minimum ripple. Such criteria are completely acceptable within the frame of a particular analysis model; that is, by specifically defining non-interactive source and system. On the other hand, if the objective is an accurate estimate of the real glottal flow, which in fact may have substantial ripple and detail, then the inverse-filter method can be treacherous. Properties justly belonging to the source might be assigned to the system, and vice versa.

6.242. Simulation of Unvoiced Excitation. The discussion of Chapter III pointed out the uncertainties in our present knowledge of unvoiced sources of excitation. Existing measurements (HEINZ, 1958) suggest that the source for voiceless continuants (fricatives) has a relatively flat spectrum in the mid-audio frequency range, and that the source impedance

is largely resistive. In electrical synthesis of speech, these sounds are commonly generated by having a broadband random noise source excite the simulated vocal resonances. Stop sounds, on the other hand, are often produced by a transient excitation of the resonators, either with electrical pulses or brief noise bursts. Voiced fricatives, since they are excited by pitch-synchronous noise bursts in the real vocal tract, can be simulated by multiplying the simulated glottal wave with an on-going broadband noise signal.

6.243. Models for Sound Generation in the Vocal Tract. Increased insight into vocal-tract excitation can be obtained from efforts to model the acoustics of human sound generation (FLANAGAN and LANDGRAF; FLANAGAN and CHERRY; ISHIZAKA; FLANAGAN, 1969). Such efforts are also directly relevant to speech synthesis by vocal-tract simulation.

6.243a. Model for Voiced Excitation. Following the analyses of Chapter III, voiced excitation of the vocal system can be represented as in Fig. 6.25. The lungs are represented by the air reservoir at the left. The force of the rib-cage muscles raises the air in the lungs to subglottal pressure P_S. This pressure expells a flow of air with volume velocity U_g through the glottal orifice and produces a local Bernoulli pressure. The vocal cords are represented as a symmetric mechanical oscillator, composed of mass M, spring K and viscous damping, B. The cord oscillator is actuated by a function of the subglottal pressure and the glottal Bernoulli pressure. The sketched waveform illustrates the pulsive form of the U_g flow during voiced sounds. The vocal tract and nasal tract are shown as tubes whose cross-sectional areas change with distance. The acoustic volume velocities at the mouth and nostrils are U_m and U_n, respectively. The sound pressure P in front of the mouth is approximately a linear superposition of the time derivatives $\dot{U}_m$ and $\dot{U}_n$.

Following the transmission-line relations derived in Chapter III, the acoustic system of Fig. 6.25 can be approximated by the network of Fig. 6.26. The lung volume is represented by a capacity and loss

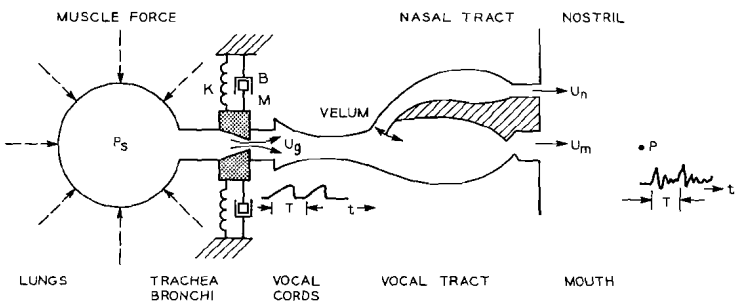

Fig. 6.25. Schematic diagram of the human vocal mechanism. (After FLANAGAN *et al.*, 1970.)

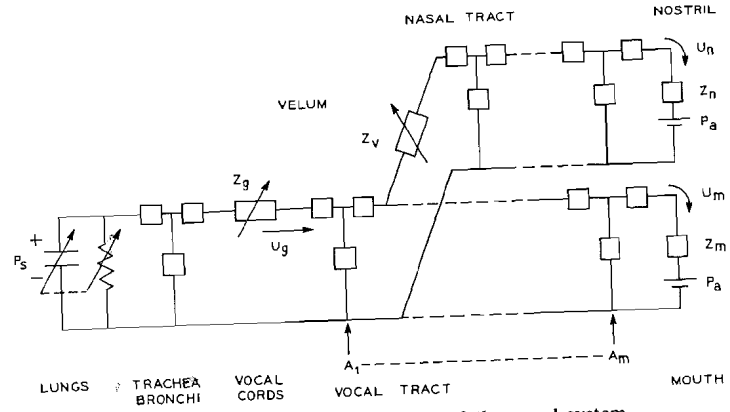

Fig. 6.26. Network representation of the vocal system

whose sizes depend upon the state of inflation. The lungs are connected to the vocal cords by the trachea and bronchi tubes, represented in the figure as a single T-section. The impedance of the vocal cords Z_g is both time-varying and dependent upon the glottal volume velocity U_g. The vocal tract is approximated as a cascade of T-sections in which the element impedances are determined by the cross-sectional areas $A_1 \ldots A_n$. The value of N is determined by the precision to which the area variation is to be represented. The line is terminated in a radiation load at the mouth Z_m, which is taken as the radiation impedance of a circular piston in a plane baffle. U_m is the mouth current and, for simulation of d.c. quantities, a battery P_a represents atmospheric pressure.

The nasal tract is coupled by the variable velar impedance Z_v. The nasal tract is essentially fixed in shape, and the nostril current U_n flows through the radiation impedance Z_n.

This formulation of the vocal system can simulate respiration as well as phonation. The glottis is opened (Z_g is reduced), the rib cage muscles enlarge the lung capacitor (volume), and the atmospheric pressure forces a charge of air through the tract and onto the capacitor. The glottis is then clenched and increased in impedance; the rib cage muscles contract, raising the voltage (pressure) across the lung capacity, and force out a flow of air. Under proper conditions, the vocal-cord oscillator is set into stable vibration, and the network is excited by periodic pulses of volume velocity. The lung pressure, cord parameters, velar coupling, and vocal tract area all vary with time during an utterance. A difference-equation specification of the network, with these variable coefficients, permits calculation of the Nyquist samples of all pressures and volume velocities, including the output sound pressure (FLANAGAN and LANDGRAF).

To simplify computation and to focus attention on the properties of the vocal-cord oscillator, the cords can be represented by a single moveable mass as shown in Fig. 6.27 (it being understood that the normal movement is bilaterally symmetric with the opposing cord-mass experiencing identical displacement). 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}.

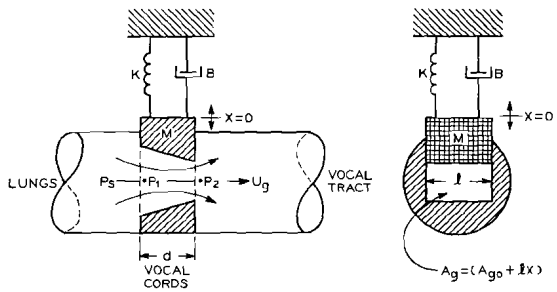

Fig. 6.27. Acoustic oscillator model of the vocal cords. (After FLANAGAN and LANDGRAF)

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 subglottal 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.

The undamped natural frequency of the oscillator is proportional to $(K/M)^{\frac{1}{2}}$. It is convenient to define a vocal-cord tension parameter Q, which scales the natural frequency by multiplying the stiffness and dividing the mass. This is loosely analogous to the physiological tensing of the cords, which stiffens them and reduces their distributed mass. Since the trachea-bronchi impedance is relatively low (compared to that of the glottal orifice), and since the large lung volume is maintained at nearly constant pressure over short durations, a source of constant pressure can approximate the subglottal system. For voiced, non-nasal sounds, this modification to the network is shown in Fig. 6.28.

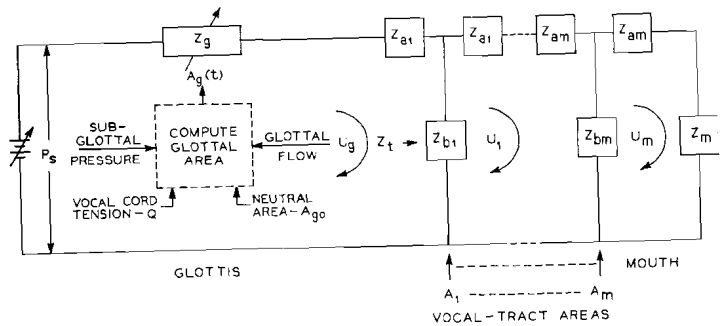

Fig. 6.28. Simplified network of the vocal system for voiced sounds. (After FLANAGAN and LANDGRAF)

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, 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 values of the tension parameter Q and of the phonation-neutral area A_{g0} are also introduced into this equation. In other words, the dashed box of Fig. 6.28 represents iterative solutions to the differential equation for the system described in Fig. 6.27.

This continuous system can be represented by $(m+2)$ differential equations, which, in turn, can be approximated by difference equations. These difference equations are programmed for simultaneous solution on a digital computer. The program accepts as input data time-varying samples of the subglottal pressure P_s, the cord tension Q, the neutral area A_{g0} and the vocal tract areas $(A_1 \ldots A_m)$, and it computes sampled values of all volume velocities, including the glottal flow and mouth output. The resulting functions can be digital-to-analog converted and led to a display scope or loud-speaker. A typical glottal area and volume velocity, plotted by the computer for a vocal-tract shape corresponding to the vowel /a/, is shown in Fig. 6.29. 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, a neutral area of 0.05 cm^2 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

[1] See Section 3.52.

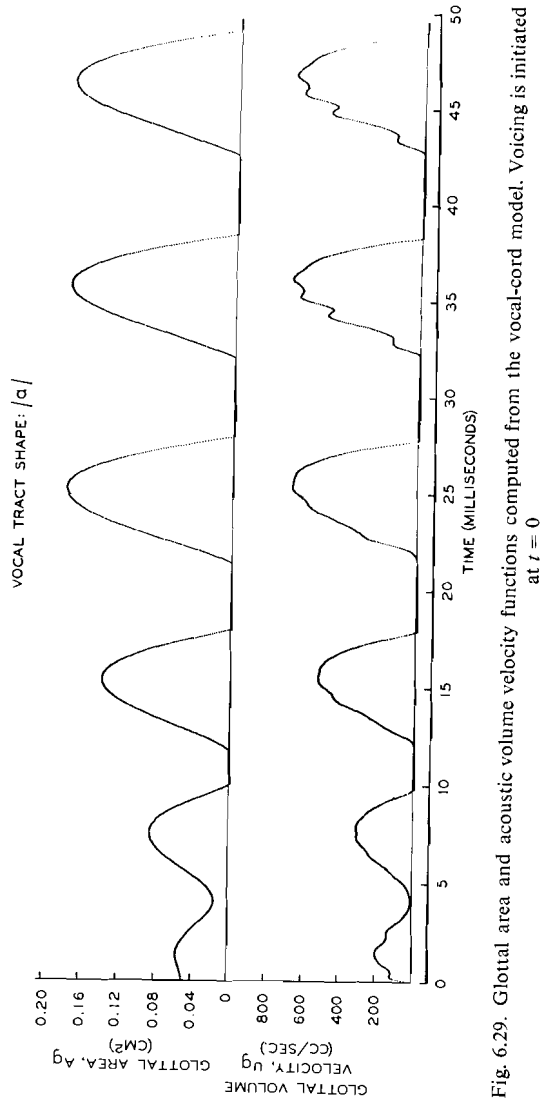

Fig. 6.29. Glottal area and acoustic volume velocity functions computed from the vocal-cord model. Voicing is initiated at $t = 0$

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.

The behavior of the vocal-cord model over a range of glottal conditions suggests that it duplicates many of the features of human speech. Furthermore, the parameters necessary for complete synthesis of voiced sounds are now reduced to the articulatory quantities: tract areas,

$A_1 \ldots A_m$; subglottal pressure, P_s; cord tension, Q; and phonation neutral area A_{g0}. A spectrogram of the audible output for a linear transition in vocal tract shape from the vowel /i/ to the vowel /a/ is shown in Fig. 6.30. The glottal conditions in this case are constant and are: $P_s = 8$ cm H_2O, $A_{g0} = 0.05$ cm² 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; this is, a function of the acoustic loading that the tract presents to the vocal cords. The spectral sections indicate realistic formant and pitch values.

The single-mass model of the cords, because of its simplicity and because it produces many features of human speech, is attractive for use in transmission-line synthesizers. It does not, however, represent physiological details such as phase differences between the upper and lower edges of the real cords. Also, its acoustic interaction with the vocal system is critically dependent upon the relations assumed for intraglottal pressure distribution. (The values determined by VAN DEN BERG were used in the above simulations.) If a more detailed simulation of the physiology and the acoustic interaction is needed, the acoustic oscillator concept can be extended to multiple mass-spring representations of the cord mass and compliance (FLANAGAN and LANDGRAF). A two-mass oscillator, stiffness coupled, has been found to represent with additional accuracy the real-cord behavior (ISHIZAKA and MATSUDAIRA; DUDGEON; ISHIZAKA and FLANAGAN). Continuing research aims to use this additional sophistication in synthesis.

6.243b. Voiceless Excitation. With slight modification, and with no additional control data, the system of Fig. 6.28 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 (see Section 3.6). To the extent that a one-dimensional wave treatment is valid, the noise sound pressure can be taken as 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 (HEINZ, 1958).

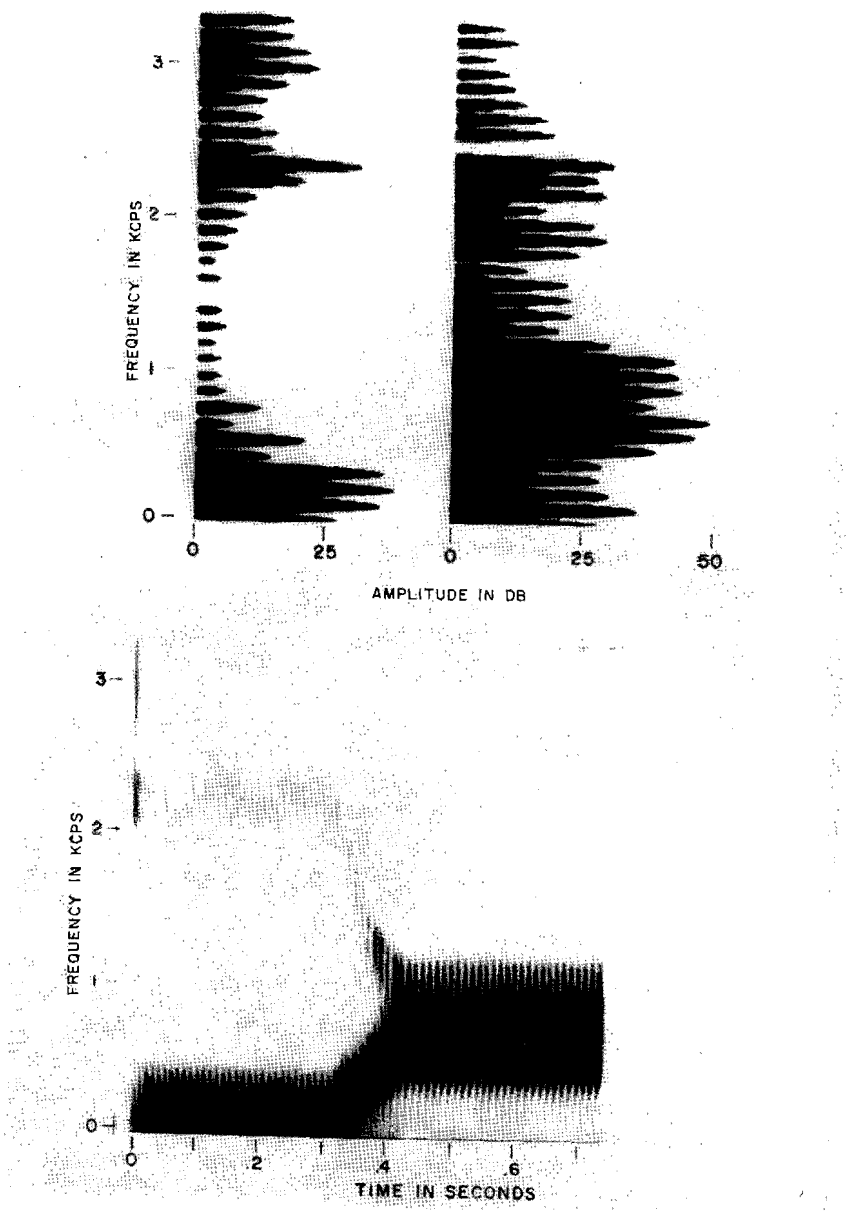

Fig. 6.30. Spectrogram of a vowel-vowel transition synthesized from the cord oscillator and vocal tract model. The output corresponds to a linear transition from the vowel /i/ to the vowel /a/. Amplitude sections are shown for the central portion of each vowel

The transmission-line vocal tract, including the vocal-cord model, can be modified to approximate the nonlinearities of turbulent flow (FLANAGAN and CHERRY). Fig. 6.31 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

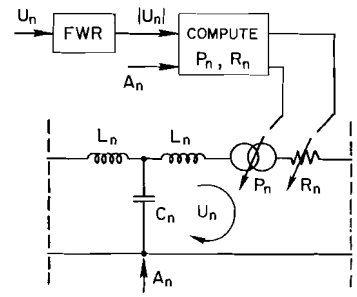

Fig. 6.31. Modification of network elements for simulating the properties of turbulent flow in the vocal tract. (After FLANAGAN and CHERRY)

is a flow-dependent loss similar to the glottal 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. In the computer simulation they are calculated 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, neutral area and cord tension, in effect, determine the place of the constriction and the loss and noise introduced there.

The P_n source is taken as 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 words, the noise is produced by the low-frequency components of U_n, including the *dc* flow.

This noise excitation works equally well for voiced and unvoiced sounds. The operation for voiced fricatives includes all features of the formulation, and is a good vehicle for illustration. For example, consider what happens in a vowel when the constriction is made substantially smaller than normal, giving rise to conditions favorable for turbulent flow. Since we have already shown results for the vowel /a/, 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. 6.32. 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 cm^2).

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. 6.29 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

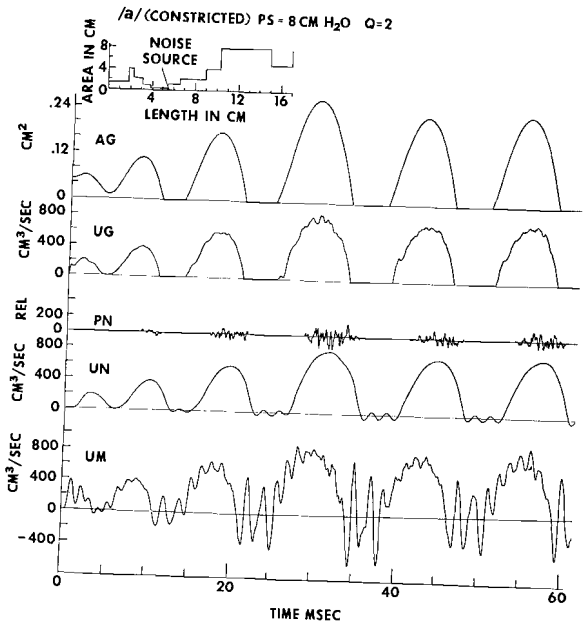

Fig. 6.32. Waveforms of vocal functions. The functions are calculated for a voiced fricative articulation corresponding to the constricted vowel /a/. (After FLANAGAN and CHERRY)

flow. Note, too, that noise influence on the mechanical oscillator motion (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 spectrogram of this audible output is compared with that for a normal /a/ in Fig. 6.33. The normal vowel is shown on the left; the constricted vowel on the right. Note in the constricted, noisy /a/ 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.

Voiceless sounds are produced in this cord-tract model simply by setting the neutral area of the vocal cords (A_{g0}) to a relatively large value, for example 1 cm^2. As this is done, the Bernoulli pressure in the glottal orifice diminishes, the oscillations of the vocal cords decay, and the cord displacement assumes a steady large value. Control of A_{g0} therefore corresponds to the voiced-voiceless distinction in the model. Measurements on real speech suggest this kind of effect in passing from voiced to voiceless sounds (SAWASHIMA; SAWASHIMA *et al.*). Corresponding exactly to this change, spectrograms of the audible output for the voiced-voiceless cognates /ʒ/ and /ʃ/ are compared in Fig. 6.34. The vocal-tract shape is the same for both sounds. One sees a pronounced voice bar in /ʒ/ (left spectrogram) that, of course, is absent in /ʃ/ (right spectrogram). 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. Fig. 6.35 shows the syllable /ʒi/ synthesized by the system. In this case, the subglottal pressure the phonation neutral area and cord tension are held constant and the vocal tract 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

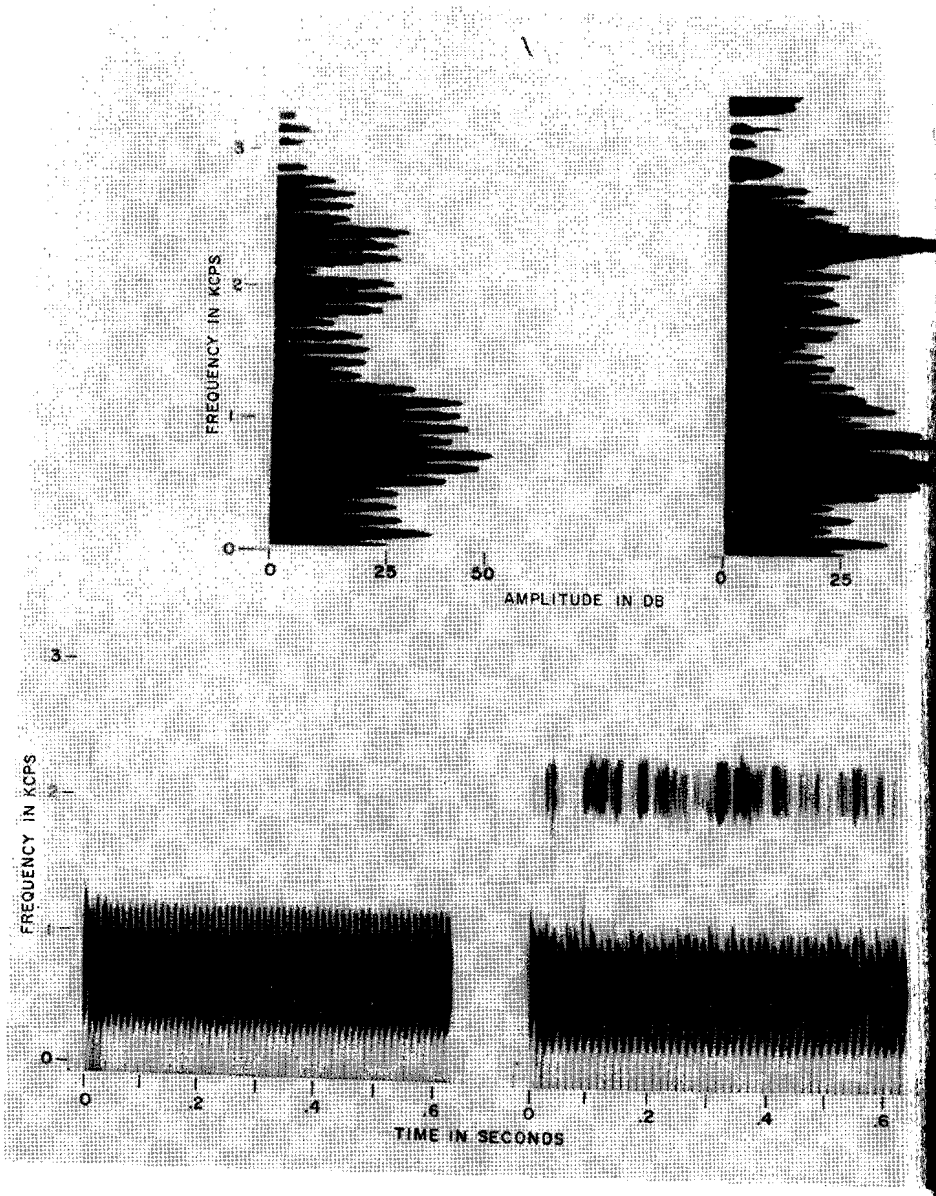

Fig. 6.33. Sound spectrograms of the synthesized output for a normal vowel /a/ (left) and the constricted /a/ shown in Fig. 6.32 (right). Amplitude sections are shown for the central portion of each vowel

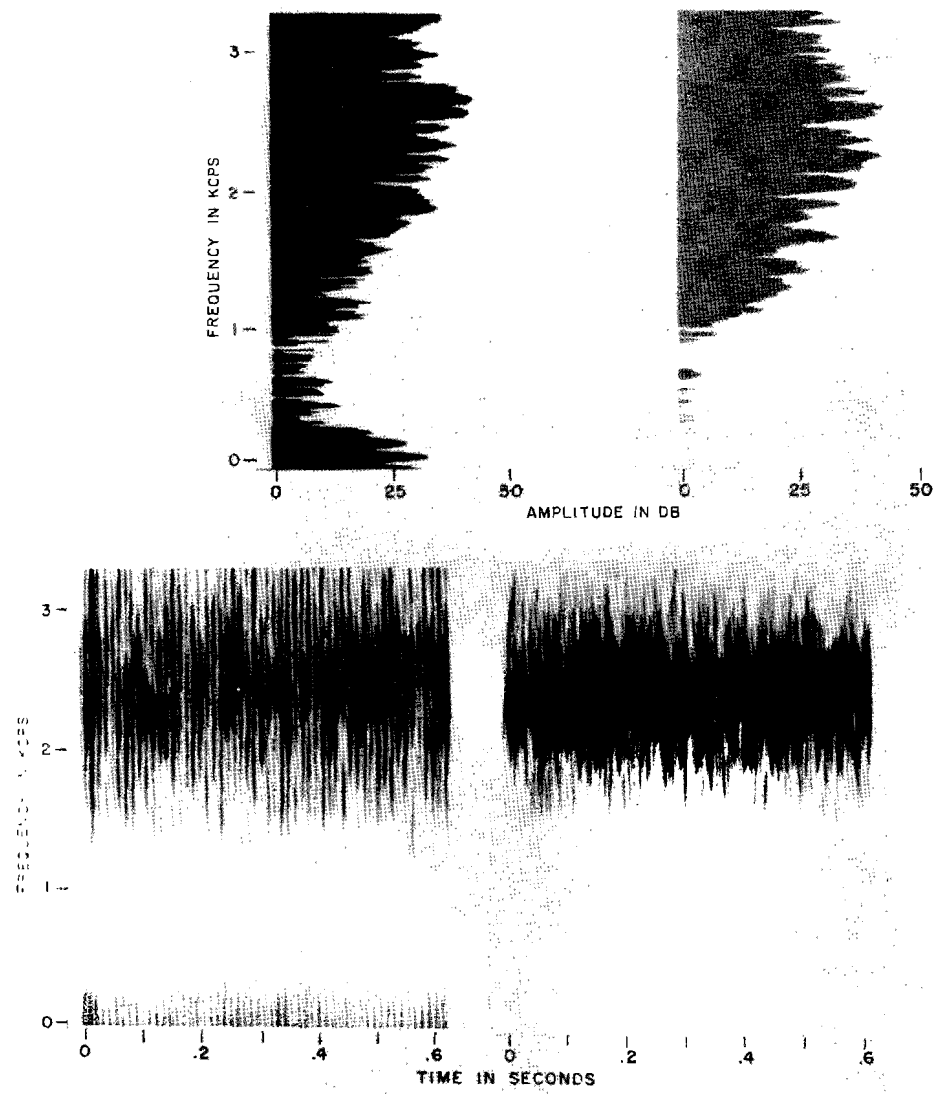

Fig. 6.34. Spectrograms for the voiced-voiceless cognates /ʒ/ and /ʃ/. Amplitude sections are shown for the central portion of each sound

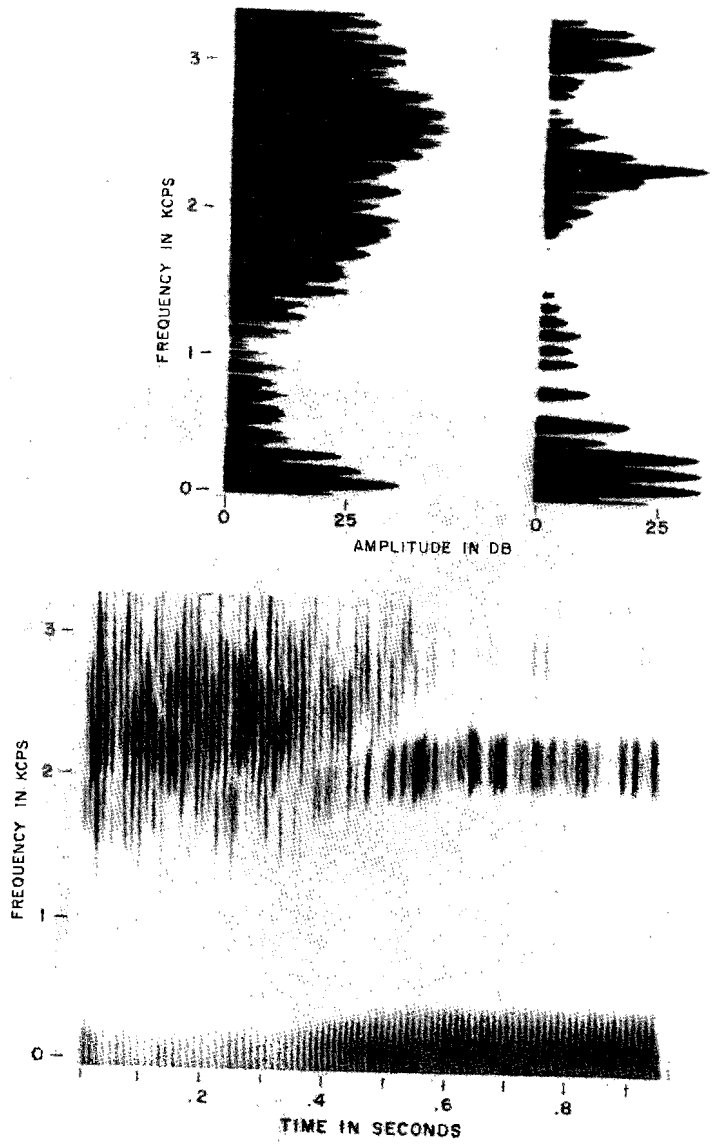

Fig. 6.35. Sound spectrogram for the synthesized syllable /ʒi/. Amplitude sections are shown for the central portion of each sound. (After FLANAGAN and CHERRY)

and a small amount of noise excitation continues in the /i/. This same effect can be seen in human speech.

This 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. Note, too, that the central parameters for the voiceless synthesis are exactly the same as for voiced synthesis; namely, $A_1 \ldots A_m, P_s, Q$ and A_{g0}. No additional control data are necessary. Place and intensity of voiceless excitation are deduced from these data.

Although crude in its representations of acoustic non-linearities, this model for voiced and voiceless excitation appears to give realistic results. It is applicable to speech synthesis by vocal tract simulation and it provides a point of departure for further study of sound generation in the human vocal tract.

6.25. Vocal Radiation Factors

Electrical synthesizers usually attempt to account for source characteristics, vocal transmission and mouth-nostril radiation. In a terminal-analog synthesizer, the radiation factor is essentially the functional relation between sound pressure at a point in space and the acoustic volume current passing the radiating port. A transmission-line analog, on the other hand, should be terminated in an impedance actually analogous to the acoustic load on the radiating port. For most speech frequencies, the latter is adequately approximated as the radiation load on a piston in a large baffle (see Section 3.3). The former, for frequencies less than about 4000 cps, is adequately approximated by the relations for a small spherical source (see Section 3.4). That is, the pressure at a point in front of the speaker is proportional to the time derivative of the mouth volume velocity.

To simulate the radiation function in terminal-analog synthesizers, a frequency equalization proportional to frequency (i.e., a 6 db/oct boost) can be applied to the vocal transmission function. Similarly in the transmission-line analog, the current through the radiation load can be differentiated to represent the output sound pressure (alternatively, the voltage directly across the radiation load can be taken as the pressure). Because the mouth and nostrils are spatially proximate (a fraction of a wavelength apart at the lower speech frequencies), the effect of simultaneous radiation from these two points can be approximated by linearly superposing their volume currents or sound pressures.

6.26. Speech Synthesis by Computer Simulation

6.261. Digital Techniques for Formant Synthesis. The approximations made of vocal transmission in Section 6.22 can be represented by linear differential equations with constant coefficients. In turn, such equations can be approximated as linear difference equations. The difference equations can be programmed in a digital computer as arithmetic operations upon discrete values of the variables[1]. As an example, the input and output voltages for the series electrical resonator shown in Fig. 6.10a are related by

$$e_i = LC\frac{d^2 e_0}{dt^2} + RC\frac{de_0}{dt} + e_0 . \tag{6.65}$$

If the derivatives are approximated by differences between successive values of the dependent variable – sampled at uniform, discrete values of the independent variable – the equation can be written as

$$e_i = e_0 + RC\Delta e_0 + LC\Delta^2 e_0 ,$$

where Δ is the first backward difference divided by the sampling interval. Explicitly,

$$\begin{aligned} e_i(t_n) = e_0(t_n) &+ RC\left[\frac{e_0(t_n) - e_0(t_{n-1})}{(t_n - t_{n-1})}\right] \\ &+ LC\left[\frac{e_0(t_n) - 2e_0(t_{n-1}) + e_0(t_{n-2})}{(t_n - t_{n-1})(t_{n-1} - t_{n-2})}\right] . \end{aligned} \tag{6.66}$$

Collecting terms

$$\begin{aligned} e_{i_n} &= e_{0_n}\left[1 + \frac{RC}{D} + \frac{LC}{D^2}\right] - e_{0_{n-1}}\left[\frac{RC}{D} + \frac{2LC}{D^2}\right] + e_{0_{n-2}}\left[\frac{LC}{D}\right] , \\ &= a\, e_{0_n} + b\, e_{0_{n-1}} + c\, e_{0_{n-2}} , \end{aligned} \tag{6.67}$$

where $D = (t_n - t_{n-1})$ is the sampling interval and $e_{0_n} = e_0(t_n)$.

The theory of linear difference equations (HILDEBRAND, 1952) shows that the unforced homogeneous solution ($e_{i_n} = 0$) of Eq. (6.67) is a linear combination of exponential terms

$$e_{0_n} = K_1 \beta_1^n + K_2 \beta_2^n , \tag{6.68}$$

where β_1 and β_2 are the roots of the determinantal equation

$$a\beta^2 + b\beta + c = 0 ,$$

[1] Alternatively, special purpose digital hardware can accomplish the arithmetic operations.

K_1 and K_2 are arbitrary constants, and a, b and c are defined in (6.67). In the present instance the roots will be complex conjugate, and

$$\beta = -\frac{b \pm j\sqrt{4ac - b^2}}{2a} = e^{r_1 \pm j r_2} , \tag{6.69}$$

where

$$e^{r_1} = \sqrt{\frac{c}{a}}$$

and

$$r_2 = \tan^{-1}\frac{\sqrt{4ac - b^2}}{-b} .$$

Therefore,

$$e_{0_n} = e^{r_1 n}(K_1' \cos r_2 n + K_2' \sin r_2 n) ,$$

where K_1' and K_2' are linear combinations of K_1 and K_2, and the response samples are those of a damped sinusoid. Following through the arithmetic gives

$$e^{r_1} = \left[\frac{1}{1 + 2\alpha D + \omega_0^2 D^2}\right]^{\frac{1}{2}} ,$$

where

$$\alpha = \frac{R}{2L} \quad \text{and} \quad \omega_0^2 = \frac{1}{LC} ,$$

and

$$r_1 = -\tfrac{1}{2}\ln[1 + 2\alpha D + \omega_0^2 D^2] . \tag{6.70}$$

Expanding the logarithm as a series for $\ln(1+x)$, $-1 < x < 1$, and taking the first term yields

$$r_1 \cong -D\left(\alpha + \frac{\omega_0^2 D}{2}\right) .$$

For a sufficiently small sampling interval D,

$$\frac{\omega_0^2 D}{2} \ll \alpha$$

and

$$r_1 \cong -\alpha D ,$$

and the response samples are damped approximately as $e^{-\alpha n D}$, which is similar to the solution for the continuous equation.

In the same fashion

$$r_2=\tan^{-1} D\left\{\frac{\left(\frac{1}{LC}-\frac{R^2}{4L}\right)}{1+\frac{RD}{L}+\frac{R^2D^2}{4L^2}}\right\}^{\frac{1}{2}}$$

$$r_2=\tan^{-1} D\left\{\frac{(\omega_0^2-\alpha^2)}{1+2\alpha D+\alpha^2 D^2}\right\}^{\frac{1}{2}} \tag{6.71}$$

$$r_2=\tan^{-1}\frac{D\omega}{(1+\alpha D)},$$

so that for small values of sampling interval

$$r_2\cong\frac{D\omega}{1+\alpha D}$$

and for small damping $r_2\cong D\omega$. The response samples are then approximately those of a damped sinusoid with angular frequency ω, which is the continuous equation solution. One notices, however, that if the sampling is coarse the solution to the difference equation begins to depart substantially from the sampled values of the continuous system. This situation can be improved by more sophisticated approximations to the derivative (which of course require additional computation). The trades which can be made between sampling rate and derivative approximation is a topic area worthy of study.

A different approach permits one to compute exact samples of the continuous impulse response. If, in addition, the sampling rate exceeds twice the bandwidth of the continuous signal, the continuous response can be reconstructed by low-pass filtering. The approach employs the z-transform. Consider the same series RLC formant resonator used in the preceding discussion [see Fig. 6.10a). Its transfer function, in terms of a Laplace transform, is

$$\frac{e_0(s)}{e_i(s)}=F(s)=\frac{s_1 s_1^*}{(s-s_1)(s-s_1^*)}=\frac{A_1}{(s-s_1)}+\frac{A_1^*}{(s-s_1^*)} \tag{6.72}$$

where

$s_1=-\sigma_1+j\omega_1$ is the pole frequency,

$A_1=\lim_{s\to s_1}(s-s_1)F(s)$ is the complex residue in pole s_1,

and the asterisk denotes complex conjugate. The inverse transform of $F(s)$ is the impulse response $f(t)$. Sampled values of the latter can be described as impulses with areas equal to the function at the sampling

instants, that is,

$$f^\dagger(t)=\sum_{n=0}^{\infty} f(t)\delta(t-nD) \tag{6.73}$$

where $\delta(t)$ is a unit area impulse and $f^\dagger(t)$ is a periodic impulse train with period D representing the sample values $f(nD)$. The transform of $f^\dagger(t)$ is the complex convolution of the transform of its components, or

$$\mathscr{L}[f^\dagger(t)]=F^\dagger(s)=F(s)*\mathscr{L}\{\sum_n \delta(t-nD)\}.$$

But

$$\mathscr{L}[\sum_n \delta(t-nD)]=1+e^{-sD}+e^{-2sD}\ldots$$

$$=\Delta(s)=\frac{1}{1-e^{-sD}},$$

which has poles at $s=\pm j2m\pi/D$, $m=0, 1, \ldots$. The convolution to be computed is

$$F^\dagger(s)=\frac{1}{2\pi j}\int_{c-j\infty}^{c+j\infty} F(\lambda)\Delta(s-\lambda)\,d\lambda. \tag{6.74}$$

Using the residue theorem and recognizing that the circuit is linear and passive so that the poles of $F(s)$ lie in the left half plane, the integral can be evaluated for a contour of integration enclosing only the poles of $F(s)$.

$$F^\dagger(s)=\sum_{\substack{k \text{ poles}\\ \text{of } F(\lambda)}} \operatorname{Res}[F(\lambda)\Delta(s-\lambda)]_{\lambda=\lambda_k},$$

or

$$F^\dagger(s)=\sum_k\left[\frac{1}{1-e^{-D(s-\lambda_k)}}\right]\operatorname{Res}[F(\lambda)]_{\lambda=\lambda_k}. \tag{6.75}$$

Making the substitution $e^{sD}=z$, Eq. (6.75) can be rewritten

$$F(z)=\sum_k\frac{1}{1-e^{\lambda_k D}z^{-1}}\operatorname{Res}[F(\lambda)]_{\lambda=\lambda_k}. \tag{6.76}$$

For the example at hand (that is, the single formant resonator)

$$\operatorname{Res}[F(s)]_{s=s_1}=A_1=\left(\frac{\sigma_1^2+\omega_1^2}{j2\omega_1}\right),$$

and

$$F(z)=\frac{\sigma_1^2+\omega_1^2}{\omega_1}\left\{\frac{e^{-\sigma_1 D}z^{-1}(\sin\omega_1 D)}{1-2e^{-\sigma_1 D}(\cos\omega_1 D)z^{-1}+e^{-2\sigma_1 D}z^{-2}}\right\}. \tag{6.77}$$

Notice also that Eq. (6.74) can be written

$$F^{\dagger}(s)=\frac{1}{2\pi j}\int_{-c-j\infty}^{-c+j\infty} F(s-\lambda)\Delta(\lambda)\,d\lambda,$$

and that the poles of $\Delta(\lambda)$ are

$$\lambda=\pm j\frac{2m\pi}{D}, \qquad m=0,1,2,\ldots,\infty.$$

If the integration contour is selected to enclose the $j\omega$-axis poles of $\Delta(\lambda)$, then the integral is

$$F^{\dagger}(s)=\frac{1}{D}\sum_{m=-\infty}^{+\infty} F\left(s-j\frac{2m\pi}{D}\right), \tag{6.78}$$

because the residue in any pole of $\Delta(\lambda)$ is $1/D$.

The system function represented by Eq. (6.75), or by Eq. (6.78), is a transform relating discrete samples of the input and output of the continuous system. Since $z^{-1}=e^{-sD}$ is a delay of one sample interval, D, the digital operations necessary to simulate the sampled response of the single formant resonator, given by Eq. (6.77), involve only delays, multiplications and summations. They are shown in Fig. 6.36a. If the $F(z)$ function in Eq. (6.77) is thought of in terms of the transmission of a common negative feedback amplifier,

$$G=\frac{K}{1+\beta K},$$

the return circuit connections in Fig. 6.36a become apparent.

The resonator of Fig. 6.36a has an impulse response equal to the sampled impulse response of the continuous function of Eq. (6.72).

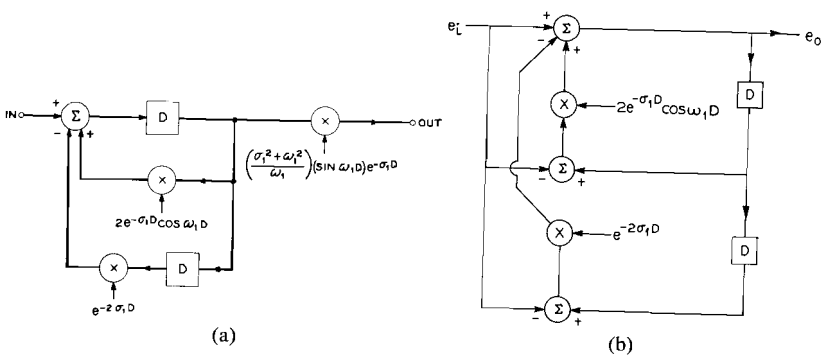

Fig. 6.36a and b. Digital operations for simulating a single formant resonance (pole-pair) (a) implementation of the standard z-transform; (b) practical implementation for unity dc gain and minimum multiplication

The frequency behavior for the two relations, however, are somewhat different. For example, their dc gains are

$$F(s)|_{s\to 0}=1$$

and

$$F^{\dagger}(s)|_{s\to 0}=\frac{1}{D}\sum_{m=-\infty} F\left(-j\frac{2m\pi}{D}\right),$$

respectively. Digital resonators can, however, be specified in terms of their frequency behavior and without direct reference to continuous resonators (GOLD and RADER). Since the formant resonance must correspond to prescribed bandwidth and frequency, and since its dc gain must be essentially unity, it is convenient in practice to modify (6.77) to

$$\frac{e_0(z)}{e_i(z)}=F(z)=\frac{1-2e^{-\sigma_1 D}\cos\omega_1 D+e^{-2\sigma_1 D}}{1-2e^{-\sigma_1 D}(\cos\omega_1 D)z^{-1}+e^{-2\sigma_1 D}z^{-2}}. \tag{6.79}$$

This relation can be programmed for a minimum of two multiplications as shown in Fig. 6.36b. The origin of the configuration of Fig. 6.36b is easily seen by noting the output $e_0(z)$ is given by

$$e_0(z)=(2e^{-\sigma_1 D}\cos\omega_1 D)(z^{-1}e_0-e_i)$$
$$-(e^{-2\sigma_1 D})(z^{-2}e_0-e_i)+e_i,$$

where, as before, z^{-1} is the delay operator e^{-sD}.

The reciprocal of $F(z)$ has zeros where $F(z)$ has poles, so that the sampled-data equivalent of a simple complex conjugate zero is the reciprocal of Eq. (6.77)

$$\frac{1}{F(z)}=\left(\frac{\omega_1}{\sigma_1^2+\omega_1^2}\right)\left[\frac{1-2e^{-\sigma_1 D}(\cos\omega_1 D)z^{-1}+e^{-2\sigma_1 D}z^{-2}}{e^{-\sigma_1 D}z^{-1}\sin\omega_1 D}\right]. \tag{6.80}$$

This response is physically unrealizable because the z^{-1} in the denominator implies an output prior to an input. Multiplication by z^{-1} to incur a unit sample delay does not alter the s-plane zero positions and makes the transmission function realizable by the digital operations shown in Fig. 6.37. As in the sampled data pole-pair, the frequency data ω_1 and the bandwidth control σ_1 are supplied to the multipliers. As with the digital resonator, it is practically convenient to have unity gain at zero frequency. The final gain multiplication in Fig. 6.37 can therefore be alternatively made $(1-2e^{-\sigma_1 D}\cos\omega_1 D+e^{-2\sigma_1 D})^{-1}$ to correspond to the reciprocal of the practical resonator shown in Fig. 6.36b and in Eq. (6.79).

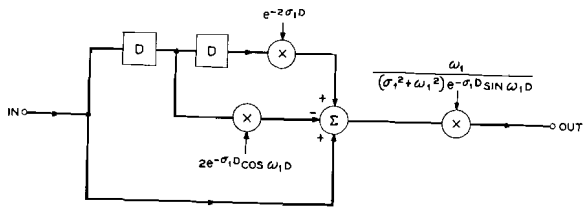

Fig. 6.37. Digital operations for simulating a single anti-resonance (zero-pair)

These basic pole and zero operations have been used to simulate a complete formant-vocoder synthesizer on a digital computer. One configuration of the synthesizer is shown in Fig. 6.38 (FLANAGAN, COKER, and BIRD). Voiced sounds are generated by the top branch which contains four variable poles and one variable zero. A fixed pole, not shown in the diagram, is included for high-frequency compensation. For vowels the final pole-zero pair is tuned coincidently so that its combined transmission is unity. Three poles therefore represent vowel spectra, in accordance with the acoustic relations developed in Section 3.7. For voiced nonvowels, such as the nasals, the final pole-zero pair is parted and positioned to represent relations given in Section 3.76. In general the pole-zero pair does not critically influence perception, provided the formant data are accurate, but is largely important to obtain realistic overall shape of the synthesized spectrum. Fundamental frequency, $F0$, and amplitude of voicing, A_v, are also controlled.

The unvoiced sounds are produced by the lower branch composed of one zero and either one or two poles. The amplitude of the noise is controlled by A_n. As Figs. 6.36 and 6.37 indicate, control of frequencies ω_n and bandwidths σ_n is effected by supplying these values to the multiplying elements in the digital circuits. Image poles, produced at multiples of the sampling frequency [see Eq. (6.78)] make further correction for higher vocal resonances unnecessary. This feature, which must be

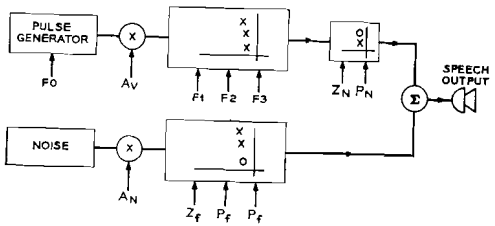

Fig. 6.38. Block diagram of a computer-simulated speech synthesizer. (After FLANAGAN, COKER and BIRD)

treated explicitly in analog synthesizers (see Section 6.221), comes free in the digital representation.

A typical listing of control data—as supplied to the computer on punched cards—is shown in Table 6.1. The data represent approximately 1 sec of synthetic speech. The first column is time in tens of milliseconds; the second, pitch in cps; the next two columns, relative amplitudes of buzz and hiss; and finally, the pole and zero frequencies in cps. Each value entered in the table is held by the circuit until a new value is specified. The control functions are interpolated between specified values in 2.5 msec steps. The sampling rate for the simulation is

Table 6.1. *Typical listing of control data for the computer-simulated synthesizer of Fig. 6.38*

Time	Pitch	A_V	A_N	F_1	F_2	F_3	P_N	Z_N	Z_F	P_F
−20	107		0	170	1290	2190	750	1000	1750	3500
4										
5		100								
7				180	1260	2170	850	950		
8				210	1470	2270	900	900		
9				390	1550	2300				
10				400	1620	2380				
11						1690	2410			
12					1700	2460				
19					1690	2500				
23				410	1510	2430				
24				350	1490	2410				
25				300	1475	2400				
26				250	1490					
28				230	1510					
32				215	1620	2390				
35				210	1700	2330				
36		0	25							
37							610	610		
41									1655	3310
46									1500	2950
47									1400	2800
48			0	320	1420	1800				
51										
52			25						975	1950
54									960	1920
56	120								925	1850
57		100	0							
58	118				1390	1750				
61	112			450	1200	1700				
65	107			600	1140	1710				
70				690	1115	1910				
72				700	1150	2000				
78					1305	2070				

$1/D = 10$ KC. A spectrogram of synthetic speech produced from such data is shown in Fig. 6.39. Also shown is the original speech from which the control functions were derived.

Digitally-simulated formant synthesizers—implemented either by programmed operations in general-purpose computers or as special-purpose digital hardware—have been used in a variety of forms (for example, KELLY and GERSTMAN; FLANAGAN, COKER and BIRD; RABINER;

Fig. 6.39. Spectrograms of synthetic speech produced by a computer-simulated formant synthesizer and of the original utterance. (After FLANAGAN, COKER and BIRD)

RABINER, SCHAFER and FLANAGAN). Analog hardware synthesizers, controlled by digital computers, have over the past had even more extensive use (for example, COKER and CUMMISKEY; HOLMES, MATTINGLY and SHEARME; DIXON and MAXEY; LEE; KATO; NAKATA; FUJISAKI). Digital implementations, however, have distinct advantages in stability and accuracy, and current advances in digital circuitry make commitment to full digital operation irresistible (RABINER *et al.*, 1971).

Much of the formant synthesis work over the past several years has made extensive use of interactive laboratory computers (see, for example, various work referenced in FLANAGAN *et al.*, 1970). Expecially valuable have been small interactive computers of integrated circuit design. Their ability for high-speed arithmetic and logic operations, and their ability to store sizeable amounts of information (both in primary and secondary memories) has substantially aided work in speech analysis and synthesis (FLANAGAN, 1971). The interactive computer has become a common laboratory tool, and as digital technology continues to develop, laboratory computers will expand in sophistication and utility.

Formant synthesizers, digitally implemented or controlled, have been used in many studies of speech synthesis-by-rule and in computer synthesis from stored formant data. In synthesis-by-rule, discrete symbols representing each speech phoneme, its duration and pitch are supplied as input. Each specified phoneme calls up from storage a set of formant values appropriate to that phoneme. Transitions of the formant and excitation functions from one phoneme to another are determined by stored rules designed to approximate the constraints of natural speech. The ultimate in synthesis-by-rule is to convert printed language to speech.

Several studies have treated the problem of converting printed English to continuous speech (TERANISHI and UMEDA; COKER and UMEDA; COKER, UMEDA and BROWMAN; LEE; ALLEN). In one of these (COKER, UMEDA and BROWMAN) a computer program uses a pronouncing dictionary to convert printed English text into discrete phonemic symbols, each carrying its own modifiers for pitch and duration. The text conversion is accomplished through a programmed syntax analysis and a prosodic feature determination. A dynamic model of the vocal tract (shown previously in Fig. 5.38) responds to the discrete phoneme commands to produce sequences of area changes similar to the real vocal tract. A difference equation solution of the Webster horn equation is periodically made to obtain continuous formant (eigenfrequency) data, and the latter are used to control a digital formant synthesizer to produce the synthetic speech.

A result of the automatic conversion of printed English into discrete control symbols for the synthesizer is shown in Table 6.2. These con-

Table 6.2. *Discrete control symbols for synthesis from printed text.* (After COKER, UMEDA and BROWMAN)

English text	Syntax and prosodic rules output
the	4dh 4a
north	6n $4aw 2er 6th
wind	6w *qq5i 4n 4d
and	4aa –n –d
the	–dh 4a
sun	6s *qq5uh 6n
were	4w 4er
arguing	4: $q6ah –r –g –y 4uu 4i 6ng
one	4w &5uh 4n
day	6d *q9ay qq9<
,	$,
when	2h 2w &5eh 4n
a	4a
traveler	4t 4tr *q7aa –v 4o –l 4er
came	4k &4ay 4< 4m
along	4a 4l 8aw 4ng
,	$,
wrapped	6r $q8aa 4p 4t
in	4i –n
a	4a
warm	6w $5ah 2er 6m
coat	6k *q2oh qq2oh 6t

trol symbols actuate articulatory motions in the vocal tract model of Fig. 5.38. The resulting synthetic output, compared with a similar human utterance is shown in Fig. 6.40. Formant motions, word durations and pitch are seen to be realistically similar to natural speech.

In synthesis from stored formant data, libraries of formant-analyzed words, phrases or syllables reside in the machine along with rules for concatenating these elements into connected speech (RABINER, SCHAFER

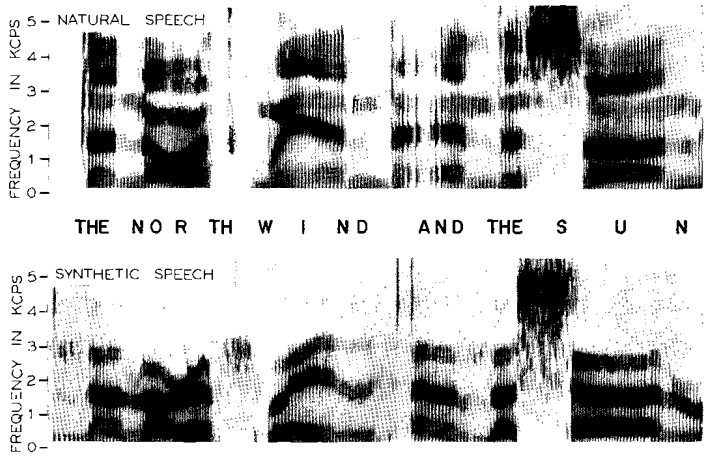

Fig. 6.40. Spectrograms comparing natural speech synthesized directly from printed text. (After COKER, UMEDA and BROWMAN)

and FLANAGAN). This approach has the advantage of using naturally-spoken signals to derive the so-called "segmental" information (i.e., the vocal resonances) rather than calculating these data. Additional storage is the price paid.

Input to the system is the English text for the word string to be generated, as illustrated in Fig. 6.41. From the library of words, stored as control functions for a formant synthesizer, the program selects and concatenates the sequence demanded. Formant functions must be interpolated naturally, word durations must be adjusted and pitch variations must be calculated for the connected utterance. The resulting control parameters are supplied to the formant synthesizer for conversion to speech.

Fig. 6.42 illustrates one technique of concatenating formant-analyzed words. At the top is a naturally-spoken sentence. At the bottom is a sentence produced from the same words spoken in isolation, formant

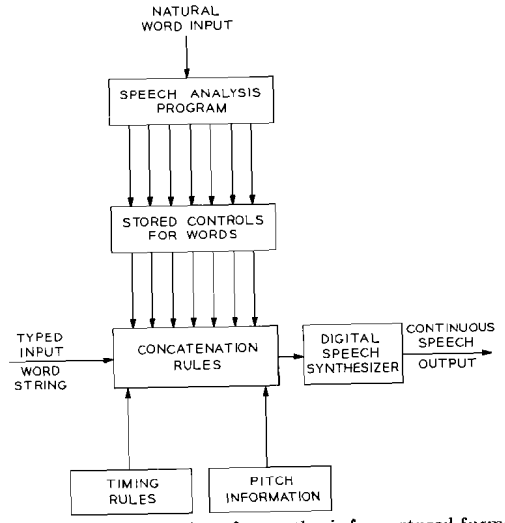

Fig. 6.41. Programmed operations for synthesis from stored formant data. (After RABINER, SCHAFER and FLANAGAN)

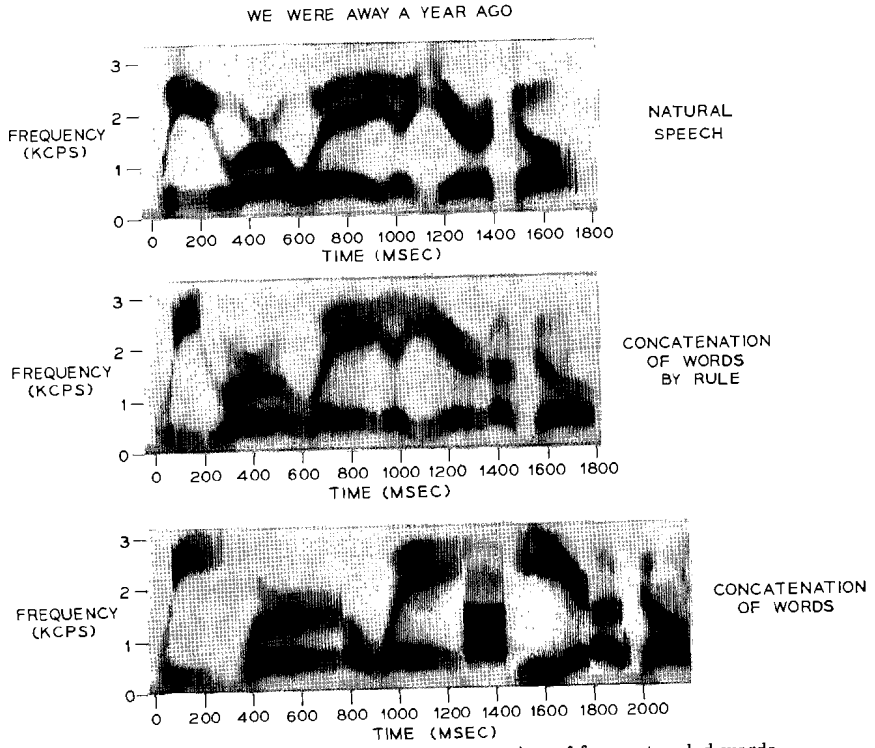

Fig. 6.42. Computer synthesis by concatenation of formant coded words. (After RABINER, SCHAFER and FLANAGAN)

analyzed and synthesized, and merely abutted in sequence. The differences are obvious and marked. The center is the result of concatenation of the same isolated words, but where the program imposes formant interpolation, word duration and pitch according to stored rules. The result is much more like the natural signal. In particular one can examine the effects on the /ə/ vowel in "... away *a* year ...", seen at about 1000 msec in the top spectrogram. Mere abuttment renders this particularly badly at about 1400 msec in the bottom spectrogram. By rule, however, in the middle spectrogram, the sound is produced relatively well at about 1000 msec. This method of concatenation has been used successfully as an automatic answer back system for speaking seven-digit telephone numbers (RABINER, SCHAFER and FLANAGAN).

Synthesis-by-rule and concatenation methods both depend critically upon the adequacy of rules for calculating prosodic information; i.e., duration, pitch and intensity variations. This problem represents a whole field of study in itself, and it is the focus of considerable interest in phonetics research.

6.262. Digital Techniques for Vocal Tract Simulation. Formant synthesizers represent the input-output characteristics of the vocal tract. For this reason they are sometimes called "terminal" analogs. In many instances it is desirable to represent the distributed nature of the vocal system. Such representation is particularly important in efforts to model articulatory dynamics of speech (where the primary input data are physiological factors, such as the area function).

Distributed aspects of the system can be treated directly in terms of the wave-equation for one-dimensional wave propagation in a non-uniform pipe [i.e., WEBSTER's horn equation, Eq. (3.1)], or in terms of bilateral transmission-line models of the system (see Section 6.23). In the first instance, steady-state solutions can be used to obtain the undamped eigen (formant) frequencies of non-nasal sounds, and transient solutions can be used to compute pressure and volume velocity distributions as functions of time. The Webster equation is

$$\frac{\partial^2 p}{\partial x^2}+\frac{1}{A}\frac{\partial p}{\partial x}\frac{\partial A}{\partial x}=\frac{1}{c^2}\frac{\partial^2 p}{\partial t^2}, \tag{6.81}$$

where $p=p(x, t)$ is the sound pressure as a function of distance and time and $A(x)$ is the vocal tract area as a function of distance[1]. For steady-state behavior $p=p(x)\,e^{j\omega t}$.

[1] The volume velocity satisfies an analogous equation

$$A\frac{\partial}{\partial x}\left(\frac{1}{A}\frac{\partial U}{\partial x}\right)=\frac{1}{c^2}\frac{\partial^2 U}{dt^2}.$$

For convenient numerical solution on a computer, the differential equation can be approximated by a difference equation. A number of possibilities exist for making the approximation. Consider space to be quantized into uniform small intervals $\Delta x=l$. Let a central second difference approximate the second derivatives and a first back difference approximate the first derivative, i.e.,

$$\left.\frac{d^2 f(x)}{dx^2}\right|_{x=x_i}=\left(\frac{f_{i+1}-2f_i+f_{i-1}}{l^2}\right)$$

and

$$\left.\frac{df(x)}{dx}\right|_{x=x_i}=\left(\frac{f_i-f_{i-1}}{l}\right). \tag{6.82}$$

Then the steady-state pressure at point $x=x_{i+1}$ can be written as the recursion formula

$$p_{i+1}=\left[p_i\left(1-\frac{\omega^2 l^2}{c^2}+\frac{A_{i-1}}{A_i}\right)-\left(\frac{A_{i-1}}{A_i}\right)p_{i-1}\right]. \tag{6.83}$$

This formulation has been used to calculate the undamped eigenfrequencies of the non-uniform tract (COKER, 1968). Typical boundary conditions for the pressure are $p_{\text{glottis}}\neq 0$, $p_{\text{mouth}}=0$. Assuming non-zero pressure at the glottis, the pressures at successive points along the tract are calculated from the recursion formula. By iteration, the value of ω is found that satisfies the mouth boundary. Convergence to the eigenfrequencies is facilitated by observing the number of pressure nodes along the tract which a given value of ω produces. That is, the first eigenfrequency corresponds to a quarter-wave resonance with no pressure nodes along tract; and the second formant to a three-quarter wave resonance with one node. This computation is repeated periodically as the $A(x)$ function changes with time.

It is relevant to note that the difference equation (6.83), so formulated, corresponds to representing the tract by a simple L-section ladder network with the LC elements shown in Fig. 6.43. The node equation relating the pressures p_{i-1}, p_i, p_{i+1} is identical to Eq. (6.83).

Another technique, useful for digital calculation of the transient sound pressure along the vocal tract, is a representation in terms of reflection coefficients (KELLY and LOCHBAUM). This approach depends upon initially approximating the non-uniform pipe by right-circular elements and assuming plane-wave propagation in each section, as discussed in Chapter 3.

Consider a plane wave moving from the left in the pipe shown in Fig. 6.44a and encountering an impedance discontinuity at $x=0$. The

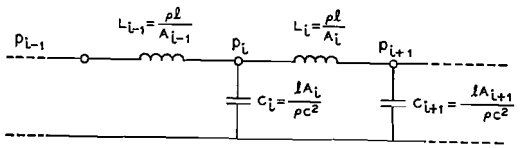

Fig. 6.43. Ladder network corresponding to a difference-equation approximation of the Webster wave equation

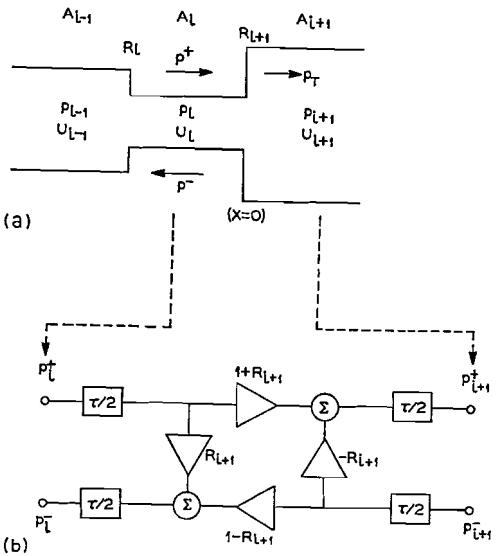

Fig. 6.44a and b. Representation of an impedance discontinuity in terms of reflection coefficients

steady-state pressure and volume velocity in the left tube must satisfy

$$p_i(x)=(p^+ e^{-jkx}+p^- e^{jkx})$$

$$U_i(x)=\frac{1}{Z_i}(p^+ e^{-jkx}-p^- e^{jkx}), \tag{6.83}$$

where p^+ and p^- are the magnitudes of plane progressive waves moving to the right and the left respectively in the tube section with area A_i, $k=\omega/c$ and Z_i is the characteristic impedance of the left tube. (The pressure and particle velocity in a plane wave are linked by $dp/dx=-j\omega\rho u$.) Since pressure and volume velocity are continuous at the boundary,

$$p_i(0)=p_{i+1}(0)=(p^+ + p^-)$$

$$U_i(0)=U_{i+1}(0)=\frac{1}{Z_i}(p^+ - p^-), \tag{6.84}$$

where the subscripts i and $i+1$ correspond to the tube elements A_i and A_{i+1}. If the right-hand tube were infinitely long with characteristic impedance Z_{i+1}, a plane wave transmitted and continuing to the right would have magnitude $p_T=(p^+ + p^-)$ and must satisfy

$$\frac{p_T}{U_{i+1}}=Z_{i+1}=\frac{Z_i(p^+ + p^-)}{(p^+ - p^-)}. \tag{6.85}$$

Then, the left-going wave in the left pipe is

$$p^-=\left(\frac{Z_{i+1}-Z_i}{Z_{i+1}+Z_i}\right)p^+=R_{i+1}\,p^+$$

and

$$p_T=(p^+ + p^-)=(1+R_{i+1})\,p^+, \tag{6.86}$$

where R_{i+1} is the reflection coefficient at the junction of A_i and A_{i+1}. If the tubes are lossless, their characteristic impedances are real

$$Z_i=\rho c/A_i; \qquad Z_{i+1}=\rho c/A_{i+1}$$

and

$$R_{i+1}=\left(\frac{A_i-A_{i+1}}{A_i+A_{i+1}}\right). \tag{6.87}$$

For a plane wave coming originally from the right, instead of the left, the sign of R_{i+1} is changed.

The Eq. (6.86) can therefore be used to represent each junction in a cascade of right-circular elements which approximate the non-uniform tract. The relations for right and left going waves are given in Fig. 6.44b, where the delay τ is the transit time through each section, $\tau=l/c$, and the unilateral amplifier boxes denote multiplication by the indicated parameters. (The $\tau/2$ delays can be lumped into single τ delays, one in the lower branch, one in the upper branch without altering the behavior.)

For any section of the quantized pipe, recursion equations describe the transient values of the (+) and (−) waves. The temporal sampling times correspond to the transit times through the uniform sections. Using i as the spatial index and j as the temporal index, the difference equations are

$$\begin{aligned} p^+_{i,j}&=-R_i\,p^-_{i,j-1}+p^+_{i-1,j-1}(1+R_i)\\ p^-_{i,j}&=R_{i+1}\,p^+_{i,j-1}+p^-_{i+1,j-1}(1-R_{i+1})\\ p_{i,j}&=(p^+_{ij}+p^-_{ij}), \end{aligned} \tag{6.88}$$

or, more conveniently for digital computation,

$$p^+_{i+1,j} = R_{i+1}(p^+_{i,j-1} - p^-_{i+1,j-1}) + p^+_{i,j-1}$$
$$p^-_{i,j} = R_{i+1}(p^+_{i,j-1} - p^-_{i+1,j-1}) + p^-_{i+1,j-1}. \tag{6.89}$$

The last pipe element of the line terminates in a load that is the radiation impedance of the mouth. Let A_L be the area of the last pipe element and Z_L the terminating radiation load. At the load terminals (the end of pipe A_L), the right-going and left-going pressure waves satisfy

$$\frac{p_L}{U_L} = Z_L = \frac{A_L(p^+_L + p^-_L)}{\rho c(p^+_L - p^-_L)}.$$

If Z_L is written in terms of a z-transform, the reflected wave p^-_L can be obtained in terms of weighted and delayed values of p^+_L; that is, a reflection coefficient relation can be set down in which $p^-_L = p^+_L f(z^{-1})$. The load pressure $(p^+_L + p^-_L)$ produces a mouth volume velocity U_L through Z_L, which, when differentiated, represents the radiated pressure. Formulations such as these have been used in a phoneme-driven vocal-tract synthesizer (KELLY and LOCHBAUM) and in a simulation of articulatory activity (MERMELSTEIN).

A further useful approach for digital representation of the distributed vocal tract follows the bilateral transmission line developed in Chapter III. Once the line composed of elemental T or π sections is set down, transient solutions of pressure and volume velcoity along the line may be obtained from difference equation approximations to the differential equations for the network. Area variations are reflected in network element changes. This approach also permits duplication of absolute acoustic impedances. For this reason it has been used in a vocal-tract synthesizer to study the acoustics of vocal-cord vibration and turbulent sound generation (FLANAGAN and CHERRY; see Section 6.243).

VII. Perception of Speech and Speech-Like Sounds

As a general topic, auditory perception can be divided a number of ways. From the standpoint of communication, one separation might be between classical auditory psychophysics, on the one hand, and the recognition of acoustic signals presented within a linguistic framework, on the other. The former relates principally to the abilities and limitations of the hearing organ as a mechano-neural transducer of all acoustic signals. The latter bears mainly upon the identification and classification of auditory patterns which are significant within the communicative experience of the listener.

Classical auditory psychophysics strives to discover the "resolving power" of the hearing mechanism. Discrimination is usually examined along fundamental dimensions of the stimulus—usually along only one dimension at a time. The measurements are generally conducted under conditions which are most favorable for making the relevant discriminations, that is, differential discriminations or close comparisons. Differential thresholds for dimensions such as intensity and frequency fall into this classification. Intuitively one feels that large neural storage and complex central processing probably are not brought into play in such detections. The measures more likely reflect the capacity of the transducer and the peripheral neural net to preserve details about a given stimulus dimension. The discussion in Chapter IV, for example, touched upon these properties of the peripheral system. The apparent relations between physiological and psychoacoustic response were analyzed for several stimulus cases. The acoustic signals were of the "classical" type in that they were either temporally punctate or spectrally simple, or both.

Speech, on the other hand, is a multidimensional signal that elicits a linguistic association. For it to be an effective communication code, some sort of absolute perceptual categorization must be made of its content. That is, the signal must be broken down into a finite number of discrete message elements. The "size" of these perceptual elements, and the manner in which they are processed to yield the percept, are questions of considerable debate and not little speculation. Our present knowledge brings us nowhere near a good understanding of the process. Theorizing about speech perception—cloaked in all of its linguistic and over-learned functions—abounds with pitfalls. An even larger problem, perhaps, is reconciling physiological, psychophysical and linguistic factors. As in other difficult situations, it is tempting to push back to some still-higher center the final decision-making process that is the real seat of perception.

Although a complete theory of speech perception remains in the future, a good deal can be said about auditory discrimination. Some of the "classical" measurements relate strongly to signal dimensions important to speech—even though the measurements are made outside of linguistic or contextual frames. In addition, a respectable amount of information has been accumulated on the acoustic cues associated with synthetic approximants to simple speech elements—for example, syllables and phonemes.

From the practical point of view, articulation tests and intelligibility measures based upon absolute recognition of sentences, words, syllables, and isolated phonemes can be used to good effect in evaluating transmission facilities. For a given processing of the voice signal, these tests often help to identify factors upon which perception depends (although