9.4, 9.8

Received 1 November 1967

Digital-Formant Synthesizer for Speech-Synthesis Studies

LAWRENCE R. RABINER

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

Recent work on speech synthesis by rule has led to the design and evaluation of a new digital-formant synthesizer. The synthesizer was simulated on a digital computer at a 20-kHz sampling frequency. The improvements over previous synthesizers include: an excitation source for the unvoiced component of voiced fricatives; a method for producing a voice bar for the closure period of voiced stops; and a simple design for the higher pole correction network that exploits the properties of sampled data systems. Spectrographic examples of synthetic speech are shown in order to illustrate these improvements. A brief discussion of some advantages and disadvantages of serial and parallel synthesizers is presented.

INTRODUCTION

IN recent years, an extensive amount of time and effort has gone into designing and evaluating machines capable of converting low-information-rate control signals into speech.[1–4] Improvements in channel and formant vocoders have been in part responsible for this resurgence of interest in speech synthesizers.[5–7] Furthermore, the general availability of both large and small digital computers[8,9] and advances in digital and analog hardware have contributed in large measure to the widespread interest in speech synthesizers.

Among the more common speech synthesizers is the terminal-analog synthesizer. Besides serving as the synthesizer for a formant vocoder, a terminal-analog synthesizer is often used to generate material for experiments in speech synthesis and perception.[9,10] The ability of a terminal analog synthesizer to produce high-quality natural-sounding speech has been demonstrated repeatedly,[11,12] thus motivating efforts to increase the capability of this machine and to make it more versatile. Its potential as a speech-research tool is high.

This paper is concerned with a digital simulation of a serial terminal-analog synthesizer. The synthesizer was simulated on a large computer (GE 645) using FORTRAN IV. The general structure of the synthesizer in relation to the production of various classes of speech sounds is examined closely, and improvements over previous synthesizers are discussed in detail. These improvements include a voiced fricative excitation network, a new design for the higher pole-correction network, and a simple method of producing a voice bar for voiced stop consonants. A discussion of some advantages and disadvantages of parallel and serial realizations of a terminal-analog synthesizer conclude this paper.

I. GENERAL DESCRIPTION

A terminal-analog synthesizer models the speech producing mechanism, including the vocal tract, excitation

[1] R. S. Tomlinson, "SPASS—An Improved Terminal-Analog Speech Synthesizer," J. Acoust. Soc. Am. 38, 940(A) (1965).

[2] G. Fant, J. Martony, U. Rengman, and A. Risberg, "OVE II Synthesis Strategy," Paper F5, Speech Commun. Seminar, Stockholm (1962).

[3] J. L. Flanagan, "Note on the Design of Terminal-Analog Speech Synthesizers," J. Acoust. Soc. Am. 29, 306–310 (1957).

[4] C. H. Coker and P. Cummiskey, "On-Line Computer Control of a Formant Synthesizer," J. Acoust. Soc. Am. 38, 940(A) (1965).

[5] C. H. Coker, "Real-Time Formant Vocoder, Using a Filter Bank, a General-Purpose Digital Computer and an Analog Synthesizer," J. Acoust, Soc. Am. 38, 940(A) (1965).

[6] B. Gold, "Techniques for Speech Bandwidth Compression, Using Combinations of Channel Vocoders and Formant Vocoders," J. Acoust. Soc. Am. 38, 2–10 (1965).

[7] M. R. Schroeder, "Vocoders: Analysis and Synthesis of Speech." Proc. IEEE 54, 720–734 (1966).

[8] P. Denes, "On Line Computing in Speech Research," Proc. Intern. Congr. Acoust. 5th Liège (1965), Paper A23.

[9] K. Nakata, "Synthesis of Nasal Consonants by a Terminal Analog Synthesizer," J. Radio Res. Labs. 6, 243–254 (1959).

[10] L. Rabiner, "Speech Synthesis by Rule: An Acoustic Domain Approach," Bell System Tech. J. 47, 17–37 (1968).

[11] J. N. Holmes, "Notes on Synthesis Work," Speech Transmission Lab. Quart. Progr. Rept., Stockholm (April 1961).

[12] W. J. Strong, "Machine-Aided Formant Determination for Speech Synthesis," J. Acoust. Soc. Am. 41, 1434–1442 (1967).

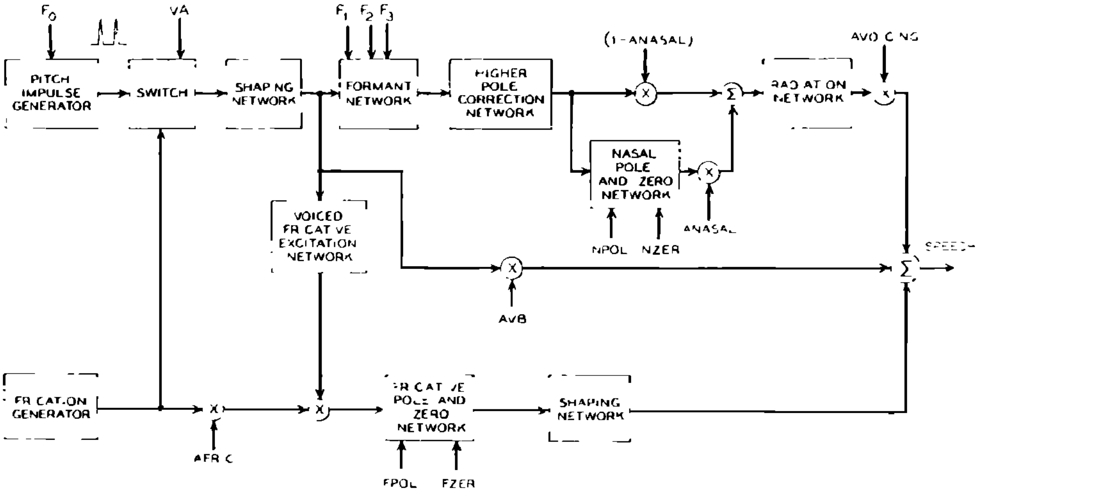

FIG. 1. Block diagram of speech synthesizer.

sources and radiation impedance. The theory of representing vocal-tract behavior, in terms of its transmission properties, by a transfer function has been developed in detail elsewhere.[13,14] The vocal-tract transfer function can be reduced to either a series of complex pole-pair and zero-pair networks, or a parallel addition of complex pole-pair networks. In theory, these realizations are equivalent, but there are practical reasons for choosing one or the other. These reasons are discussed in a later Section. The serial realization is used in the discussion that follows.

A block diagram of the synthesizer is shown in Fig. 1. There are two sources of excitation for the synthesizer: These are the pitch-impulse generator and the frication generator. The pitch-impulse generator emits impulses at a rate specified externally by the F_0 control signal. This generator provides the excitation for voiced speech and for both the voiced and unvoiced components of voiced fricatives. The output of the frication generator is a stationary white Gaussian noise with zero mean and specified standard deviation. The frication generator serves as the source of aspiration, whispered speech, and frication noise for fricatives and voiceless stop consonants. The levels of the sources are set to supply approximately equal power.

The remaining blocks of the synthesizer provide the filtering necessary to produce the various classes of speech sounds. Vocalic sounds are produced using the upper branch of the synthesizer. The upper branch consists of an externally controlled switch (which gates either source to the shaping network); a shaping network; a formant network; a higher-pole correction network; a nasal side branch in parallel with a gain control; and a radiation network. For vocalic sounds, the VA control signal is set to gate the pitch impulse generator output to the shaping network.

The shaping network provides suitably shaped pulses to excite the synthesizer. A complex-conjugate pole-pair resonator providing a high frequency falloff of 12 dB oct was used here. The formant network consists of three complex conjugate pole-pair resonators. The center frequencies and bandwidths of the resonators are externally controlled (F_1, F_2, F_3). (Note that each frequency parameter specifies a complex pole or zero position indicating *both* center frequency and bandwidth.)

The higher pole correction network provides the compensation to the low-frequency spectrum for the higher-frequency speech resonances not directly included in the synthesizer. The characteristics of this network are well known,[13,15] as are electrical realizations. However, a particularly simple network based on the properties of sampled data systems was used in this synthesizer.[16] This network is explained in a later Section.

The external amplitude control ANASAL is set to zero for vocalic nonnasal sounds, thereby eliminating the nasal side branch. For nasal sounds, ANASAL is set to one, eliminating the direct path to the radiation network and introducing an additional complex conjugate pole-pair and zero-pair resonator. The center frequencies and bandwidths of the additional resonances are externally controlled (NPOL, NZER).

The radiation network provides a 6-dB/oct rise in the speech spectrum. It consists simply of a differentiator. The gain control AVOICING provides means for controlling the over all voiced output level.

The lower branch of the synthesizer is used to produce unvoiced sounds. This branch consists of a frication

[13] G. Fant, *Acoustic Theory of Speech Production* (Mouton and Co., 's-Gravenhage, The Netherlands, 1960).

[14] J. L. Flanagan, *Speech Analysis, Synthesis and Perception* (Academic Press Inc., New York, 1965), pp. 175-188.

[15] Ref. 13, pp. 178-179.

[16] B. Gold and L. Rabiner, "Analysis of Digital and Analog Formant Synthesizers," IEEE Trans. on Audio and Electroacoust. (to be published).

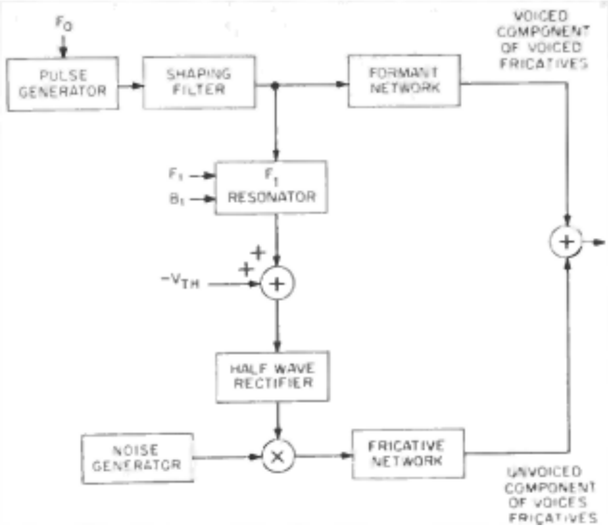

FIG. 2. Excitation network for voiced fricatives.

generator, a gain control, a multiplier, a frication pole and zero network, and a shaping network. The gain control AFRIC provides means for controlling the overall unvoiced output level.

The input to the multiplier from the voiced-fricative excitation network is constant except during a voiced fricative. For a voiced fricative, the input to the multiplier is a signal that pitch synchronously modulates the unvoiced input. The details of this technique are explained in the next Section.

The fricative pole and zero network is simply a cascade of a complex-conjugate pole–pair resonator and a complex-conjugate zero-pair resonator. The pole and zero positions are variable (FPOL, FZER). The shaping network is used to provide high- and low-frequency emphasis, as well as to account for radiation characteristics. A cascade of a differentiator and a low-frequency resonator is used here.

The middle branch of the synthesizer is used to produce a voicebar for the closure interval of voiced stop consonants. This branch consists simply of a gain control (AVB) that adjusts the level of the pitch pulses produced at the output of the shaping network in the upper branch.

The output speech is produced by adding the output from each of the three branches.

II. VOICED FRICATIVE EXCITATION NETWORK

The voiced-fricative excitation network that connects the output of the pulse-shaping filter to the lower branch of the synthesizer is used to model the production of the unvoiced component of voiced fricatives.[17] Figure 2 shows the relevant details of this network. (Certain components of the synthesizer of Fig. 1 have been omitted from Fig. 2 for simplicity.)

The unvoiced excitation is produced as follows. The pitch pulses excite a resonator tuned to the first formant of the voiced component of the fricative. This resonator is the first-order approximation to the transfer function of volume velocity (the signal of interest in Fig. 2) from the glottis through the point of constriction of the vocal tract. A threshold level (V_{TH}) is subtracted from the output of the resonator, and the result is half-wave rectified. These operations model the physical fact that turbulence is not produced until the volume velocity of the airflow exceeds a threshold value.

The output of the half-wave rectifier modulates the output of the noise generator, producing a pitch-synchronous excitation for the unvoiced component of the fricative. The final unvoiced component is produced by feeding this excitation into the fricative network (i.e., the lower branch of the synthesizer). The voiced component is produced by exciting the formant network by the pitch pulses.

Figure 3 shows spectrograms of synthetic and natural versions of the voiced fricatives /z/ and /ʒ/. The synthetic speech was produced entirely by rule.[10] The natural speech is presented for comparison purposes. A careful examination of these spectrograms shows the effects of the pitch-synchronous modulation during the fricatives for both the synthetic and the natural speech. The presence of the low-frequency voiced components for both the synthetic and natural versions can also be seen in Fig. 3. The major difference between the companion spectrograms is that the lower formants of the natural speech are weaker than the lower formants of

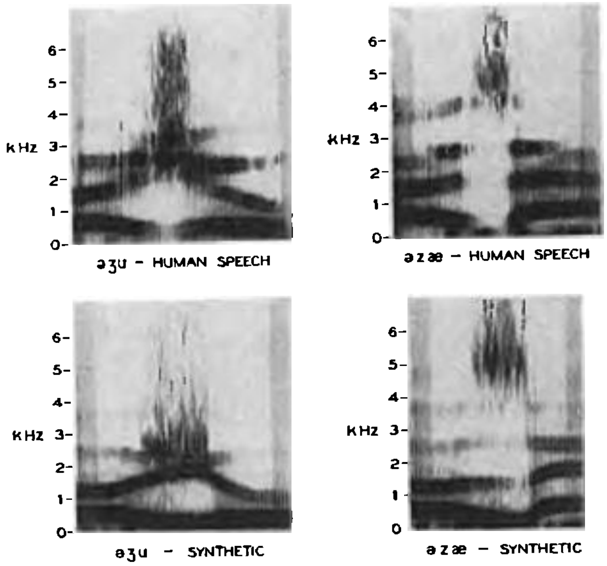

FIG. 3. Spectrographic examples of voiced fricatives.

[17] O. Fujimura (personal communication, 1967) has indicated that G. Rosen used a modulation scheme to produce voiced fricatives on DAVO [G. Rosen, "Dynamic Analog Speech Synthesizer," J. Acoust. Soc. Am. 30, 201–210 (1958)]; however, no written record of his techniques appears to be available.

the synthetic versions, indicating a need for low-frequency de-emphasis.

The intelligibility of the voiced fricatives was measured in a formal test, as part of a larger-scale evaluation of synthetic consonants.[10] The relevant results are that /z/ and /ʒ/ were identified correctly 100% of the time in prestressed and poststressed position in VCV tests. There were 15 possible responses for each stimulus in these tests. The voiced fricatives /v/ and /ð/ are not included in this discussion because the unvoiced component was not found necessary for their synthesis. Hence, they were synthesized in the same way as vocalic sounds.

No formal tests were conducted to determine the quality of the synthetic voiced fricatives, but informal comments by listeners were highly favorable.

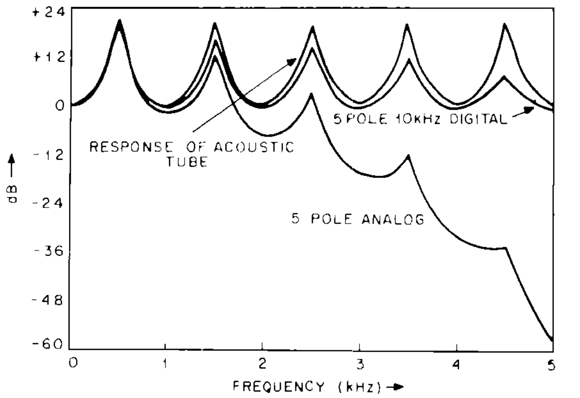

FIG. 5. Digital and analog approximations to transfer function of an acoustic tube, open at one end and closed at the other end.

III. HIGHER-POLE CORRECTION NETWORK

The higher-pole correction network is used to compensate for the effects, on the low-frequency spectrum, of the missing higher-order poles in the usual approximation to the vocal-tract transfer function. In an analog synthesizer, a network approximating the transfer function of the higher-order poles is used. In a digital synthesizer, however, the design of the higher-pole correction is relatively simple. Additional resonators, or formant networks, placed at appropriately chosen frequencies, provide the necessary higher-pole correction.

To illustrate this method, consider a synthesizer with a 10-kHz sampling frequency. Assume the vowel /ə/ is to be synthesized and the pole positions are 500, 1500, 2500, 3500, 4500, ··· $(2n+1)$ 500 Hz. As shown in Fig. 4, we place the three resonators of the formant network at 500, 1500, and 2500 Hz. The higher-pole correction network consists of resonators at 3500 and 4500 Hz. Owing to the periodicity, in frequency, of sampled data systems, the pole at any frequency represents an infinity of poles with a spacing equal to the sampling frequency—i.e., the pole at 500 Hz represents analog poles at 500, 10 500, 20 500 Hz, etc. Hence, the infinity of higher-order poles is guaranteed by using this method. The density of the higher-order poles is also correct, thus automatically ensuring the necessary higher-pole correction. The effectiveness of the digital higher-pole correction is shown in Fig. 5. Here we see five-pole digital and five-pole analog approximations to the transfer function of a straight acoustic tube modeling the vowel /ə/. No analog higher-pole correction is used other than the additional fixed formants at 3500 and 4500 Hz. The bandwidths of the digital and analog poles are increasing with frequency (modeling the observed data for speech), whereas the bandwidths of the poles of the tube transfer function are constant. The high-frequency behavior of the analog approximation is significantly different from the tube transfer function, indicating the need for additional higher-pole correction. The advantages of the automatic higher-pole correction of the digital approximation are apparent.

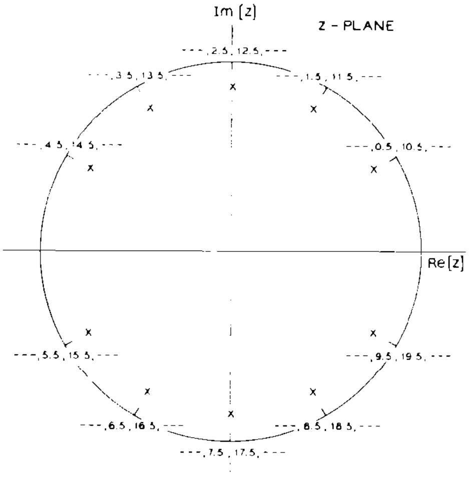

FIG. 4. Pole positions for a five-pole 10-kHz synthesizer.

The modifications of the higher-pole correction network for a 20-kHz sampling frequency are straightforward. Instead of inserting fixed resonators at just 3500 and 4500 Hz, additional resonators at 5500, 6500, 7500, 8500, and 9500 Hz are necessary. There will be 10 poles in the upper half-plane of Fig. 4, instead of five as for a 10-kHz synthesizer; however, the poles are at the same frequencies as before. For a 20-kHz synthesizer,

TABLE I. Resonator frequencies and bandwidths.

Resonator	CF (Hz)	Q	BW (Hz)
F_4	3500	20	175
F_5	4500	16	281
F_6	5500	12	458
F_7	6500	9	722
F_8	7500	6	1250
F_9	8500	4	2125
F_{10}	9500	2	4750

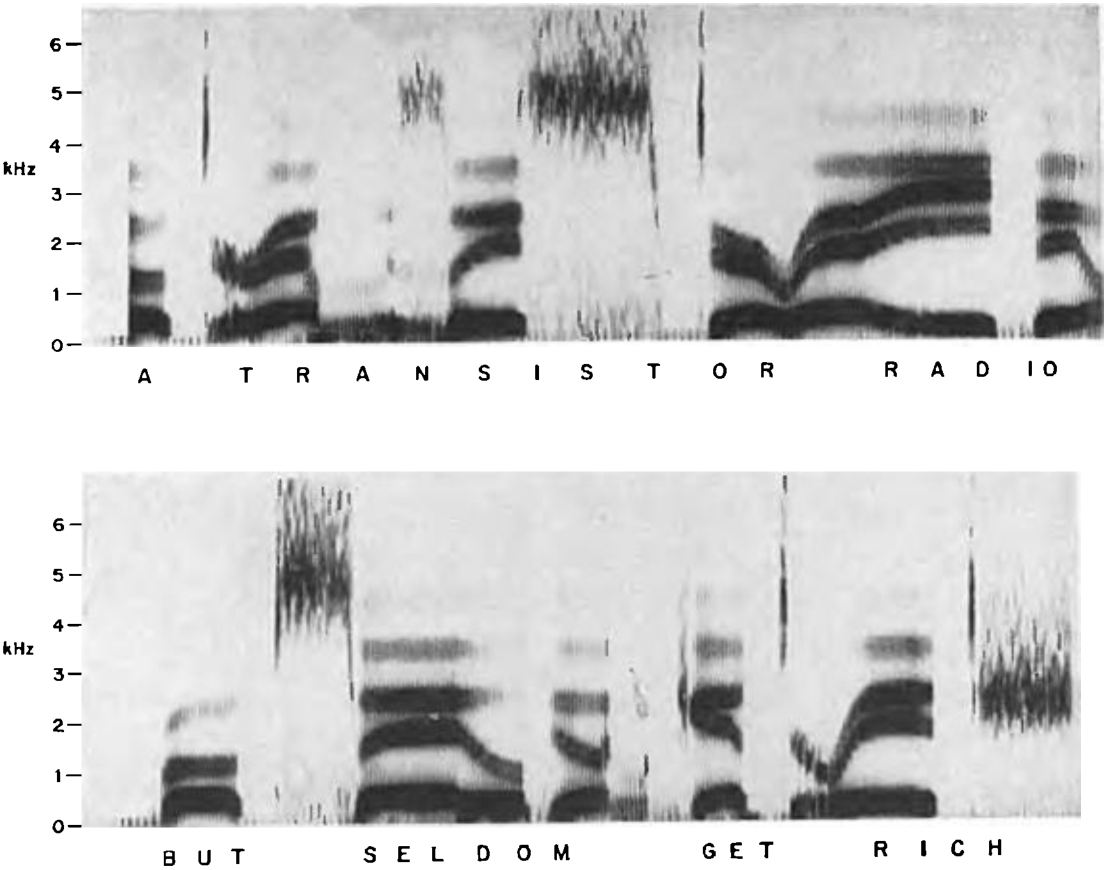

FIG. 6. Spectrograms of synthetic speech.

the bandwidths of the additional poles are greater than the bandwidths of the similar poles for a 10-kHz synthesizer. The bandwidths used are seen in Table I. The values chosen were determined from data by Dunn.[18]

IV. SYNTHESIS CHARACTERISTICS OF THE SPEECH SOUNDS

In order to produce the various classes of speech sounds, only certain portions of the synthesizer need generally be used at one time. Vowels, semivowels, liquids, and glides are produced using only the upper branch of the synthesizer. The pitch-impulse generator is the source of excitation for these sounds.

Voiced stop consonants are produced using both the upper and middle branches of the synthesizer. The voice bar is produced by gating the output of the pitch-impulse generator to the shaping network and then multiplying the output of this network by an external amplitude control (AVB). This provides a low-frequency low-energy waveform whose characteristics are similar to voice bars observed in real speech. The transitional segment of voiced stops is produced by exciting the upper branch of the synthesizer by the pitch impulses.

Voiceless stop consonants are produced using the upper and lower branches of the synthesizer. The time interval following the period of closure can be broken down into three segments. During the first segment, there is a short burst of frication noise. This noise is produced using the lower branch of the synthesizer. The output of the frication generator is multiplied by an external amplitude control (AFRIC) and then multiplied by a fixed constant. The output of the second multiplier is fed into the fricative pole and zero network, and then into the shaping network.

The second segment of a voiceless stop is the aspiration phase. Aspiration is produced by gating the output of the frication generator to the upper branch of the synthesizer. The third segment of the voiceless stop, the voiced transitional phase, is produced in exactly the same manner as for voiced stops.

The voiceless fricatives are produced in the same way as the frication noise for voiceless stop consonants. As shown above, a voiced fricative is synthesized as two separate components. The voiced component is produced in a manner similar to the voiced components of other speech sounds—i.e., by exciting the upper branch

[18] H. K. Dunn, "Methods of Measuring Vowel Formant Bandwidths," J. Acoust. Soc. Am. 33, 1737–1746 (1961).

of the synthesizer by pitch impulses. The unvoiced component is produced by modulating the frication generator output by a pitch-synchronous waveform—the output of the voiced-fricative excitation network. The modulated waveform is then passed through the lower branch of the synthesizer and added to the voiced component to produce the voiced fricative. An /h/ sound and whispered speech are produced in a manner similar to that for aspiration.

A. Spectrographic Examples

In order to illustrate the capabilities of the synthesizer, we show, in Fig. 6, spectrograms of two synthetic utterances. The control signals used to drive the synthesizer were determined by rule using an auxiliary program. Therefore, there are no original speech samples to present for comparison purposes.

These spectrograms exhibit the various classes of speech sounds. The upper spectrogram shows both voiced sounds and unvoiced sounds. Of most interest in this spectrogram are the phonemes in the word *transistor.* During the /t/, both the frication noise burst and aspiration are visible. The effects of the nasal pole and zero for /n/ are noticeable in the region of 2 kHz, where there is little energy present. Finally, a comparison is available between the voiced fricative /z/ and its voiceless counterpart /s/. The effects of the pitch-synchronous modulation are easily identified by the vertical striations through the noise of /z/. Furthermore, the lack of a low-frequency voiced spectrum for /s/ is easily observed.

The lower spectrogram exhibits two interesting features. The voice bars of /d/ in "seldom" and /g/ in "get" appear as low-frequency vertical striations during their respective stop gaps. The voice bars appear as light lines indicating that they are of low energy. Finally, the affricate /tʃ/ of "rich" appears as a stop gap followed by a burst and ending in a period of frication energy. The frication energy is similar to the energy of the /ʃ/ phoneme. An affricate is synthesized as a stoplike sound followed by a fricativelike sound on this synthesizer.

B. Parallel versus Serial Realization

The discussion in this paper has been concerned primarily with a serial terminal-analog synthesizer. A parallel synthesizer is generally realized by gating either of two sources (a noise or pitch-impulse generator) to a series of parallel branches, each containing a resonator and a gain control, and adding the output of each branch to produce the speech. Figure 7 shows a simple realization of a parallel synthesizer[12,19] (a more sophisticated parallel synthesizer would account for formant phase[20]).

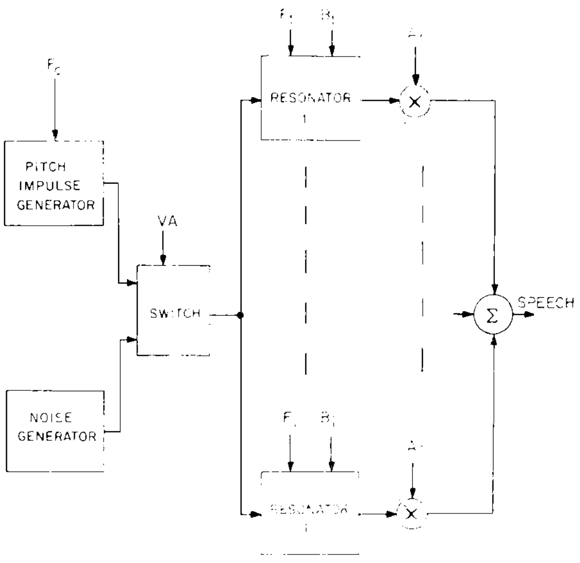

FIG. 7. Parallel terminal-analog synthesizer.

The serial synthesizer presents several advantages over the synthesizer of Fig. 7. Individual amplitudes of each of the resonances do not have to be determined for a serial synthesizer. For research on speech synthesis by rule, this is a significant advantage, as it reduces the complexity of the rules. The vowel spectra from the parallel synthesizer of Fig. 7 contain extraneous zeros, whereas a serial synthesizer produces spectra containing only poles. These zeros generally fall at frequencies between the resonances and may be perceptible, and hence a corrupting factor. Because of the presence of zeros in the spectra, parallel synthesizers generally have no higher-pole correction networks, relying instead on the zeros to provide the low-frequency emphasis. This is an unreliable technique because the positions of the zeros move as the resonances are varied.

The parallel synthesizer has two advantages over a serial synthesizer. Noise generated in the parallel synthesizer propagates additively rather than multiplicatively, as in a serial synthesizer. For simulations where small register lengths are required, this advantage is particularly valuable. For a given signal-to-noise ratio, signal sizes in a parallel synthesizer are smaller than in a serial synthesizer. The second advantage is the ability of a parallel synthesizer to reproduce consonant spectra accurately through independent control of formant amplitudes. Hence, for synthesis from spectrograms or in a formant vocoder, this degree of freedom can be of great value.

[19] J. Holmes, I. Mattingly, and J. Shearme, "Speech Synthesis by Rule," Language and Speech 7, 127–143 (1964).

[20] J. L. Flanagan, "Recent Studies in Speech Research at Bell Telephone Laboratories (II)," Proc Intern. Congr. Acoust., 5th Liège (1965), Paper A22.

The choice of a parallel or serial synthesizer is determined primarily by its application. For research in speech synthesis by rule, the serial synthesizer has been very useful.

V. SUMMARY

A new design for a digital terminal-analog synthesizer has been presented and discussed. The new features include a voiced fricative excitation network, a method of producing a voice bar, and a simple higher-pole correction network based on the properties of sampled data systems. Spectrographic examples of synthetic speech have been included to demonstrate the capabilities of the synthesizer.

ACKNOWLEDGMENTS

I am indebted to Dr. Cecil Coker of Bell Telephone Laboratories and Professor Osamu Fujimura of the Research Institute of Logopedics and Phoniatrics for many valuable discussions on the design of a speech synthesizer; to Dr. James Flanagan, Professor Kenneth Stevens, and Professor William Henke for suggestions and ideas about speech synthesis; and to Dr. Harry Levitt for valuable criticisms of the manuscript.