Software for a cascade/parallel formant synthesizer

Dennis H. Klatt

Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

(Received 1 July 1979; accepted for publication 12 October 1979)

A software formant synthesizer is described that can generate synthetic speech using a laboratory digital computer. A flexible synthesizer configuration permits the synthesis of sonorants by either a cascade or parallel connection of digital resonators, but frication spectra must be synthesized by a set of resonators connected in parallel. A control program lets the user specify variable control parameter data, such as formant frequencies as a function of time, as a sequence of ⟨time, value⟩ points. The synthesizer design is described and motivated in Secs. I–III, and FORTRAN listings for the synthesizer and control program are provided in an appendix. Computer requirements and necessary support software are described in Sec. IV. Strategies for the imitation of any speech utterance are described in Sec. V, and suggested values of control parameters for the synthesis of many English sounds are presented in tabular form.

PACS numbers: 43.70.Jt, 43.70.Qa

INTRODUCTION

A need exists in psychology and the speech sciences for a flexible research tool in order to study speech perception through the synthesis of speech and speech-like sounds. Since most perceptual experiments are now performed under control of a general purpose digital computer, there are advantages to the use of the same laboratory computer for generating speech-like stimuli (rather than the purchase of special-purpose hardware).

The cascade/parallel formant synthesizer to be described can be simulated on a general-purpose digital computer in the manner depicted in Fig. 1. Synthesizer control parameter data such as the frequency motions of the first formant as a function of time are specified by the experimenter, using the synthesizer control program HANDSY.FOR. As many as 20 control parameters may be varied as a function of time to serve as input to the waveform generating synthesizer subroutines PARCOE.FOR and COEWAV.FOR. These three FORTRAN programs appear in Appendix B. Output waveform samples are computed in nonreal time and stored on a disk for subsequent playback through a digital-to-analog converter, analog low-pass filter, and loudspeaker.

Software simulation versus hardware construction

The advantages of a software implementation over the construction of analog hardware are substantial. The synthesizer does not need repeated calibration, it is stable, and the signal-to-noise ratio (quantization noise in the case of a digital simulation) can be made as large as desired. The configuration can easily be changed as new ideas are proposed. For example, the voices of women and children can be synthesized with appropriate modifications to the voicing source and cascade vocal tract configuration. Display terminals are usually available in a computer facility and can be programmed to view control parameter data or selected portions of the output speech waveform (see Sec. IV). Short-time spectra can also be computed and displayed in order to make detailed spectral comparisons between natural and synthetic waveforms (see Sec. V).

Formant synthesis versus articulatory synthesis

Speech synthesizers fall into two broad categories: (1) articulatory synthesizers that attempt to model faithfully the mechanical motions of the articulators and the resulting distributions of volume velocity and sound pressure in the lungs, larynx, and vocal and nasal tracts (Flanagan, Ishizaka, and Shipley, 1975), and (2) formant synthesizers which derive an approximation to

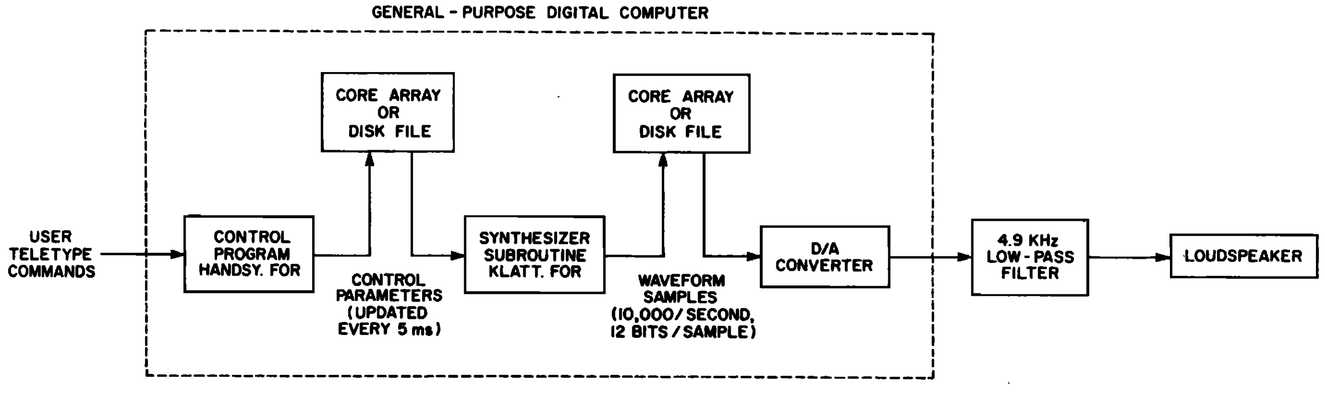
GENERAL - PURPOSE DIGITAL COMPUTER
USER TELETYPE COMMANDS
CONTROL PROGRAM HANDSY. FOR
CORE ARRAY OR DISK FILE
CONTROL PARAMETERS (UPDATED EVERY 5 ms)
SYNTHESIZER SUBROUTINE KLATT. FOR
CORE ARRAY OR DISK FILE
WAVEFORM SAMPLES (10,000/SECOND, 12 BITS/SAMPLE)
D/A CONVERTER
4.9 KHz LOW-PASS FILTER
LOUDSPEAKER

FIG. 1. Relation of the software synthesizer to the hardware and supporting software of a small general-purpose digital computer.

0001-4966/80/030971-25$00.80

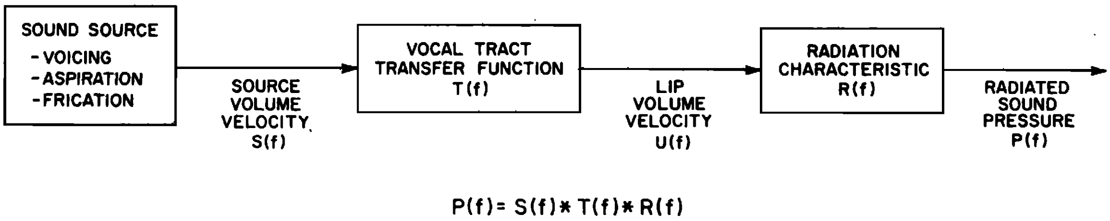
SOUND SOURCE
- VOICING
- ASPIRATION
- FRICATION
SOURCE VOLUME VELOCITY S(f)
VOCAL TRACT TRANSFER FUNCTION T(f)
LIP VOLUME VELOCITY U(f)
RADIATION CHARACTERISTIC R(f)
RADIATED SOUND PRESSURE P(f)
P(f)= S(f)* T(f)* R(f)

FIG. 2. The output spectrum of a speech sound, $P(f)$, can be represented in the frequency domain as a product of a source spectrum $S(f)$, a vocal tract transfer function, $T(f)$, and a radiation characteristic, $R(f)$.

a speech waveform by a simpler set of rules formulated in the acoustic domain. The present paper is concerned only with formant models of speech generation since current articulartory models require several orders of magnitude more computation, and the resultant speech output cannot be specified with sufficient precision for psychophysical experimentation.

The synthesizer design is based on an acoustic theory of speech production presented in Fant (1960), and is summarized in Fig. 2. According to this view, one or more sources of sound energy are activated by the buildup of lung pressure. Treating each sound source separately, we may characterize it in the frequency domain by a source spectrum, $S(f)$, where f is frequency in Hz. Each sound source excites the vocal tract which acts as a resonating system analogous to an organ pipe. Since the vocal tract is a linear system, it can be characterized in the frequency domain by a linear transfer function, $T(f)$, which is a ratio of lip-plus-nose volume velocity, $U(f)$, to source input, $S(f)$. Finally, the spectrum of the sound pressure that would be recorded some distance from the lips of the talker, $P(f)$, is related to lip-plus-nose volume velocity, $U(f)$, by a radiation characteristic, $R(f)$, that describes the effects of directional sound propagation from the head.

Each of the above relations can also be recast in the time (waveform) domain. This is actually how a waveform is generated in the computer. The synthesizer includes components to simulate the generation of several different kinds of sound sources (described in Sec. I), components to simulate the vocal tract transfer function (Sec. II), and a component to simulate sound radiation from the head (Sec. III).

Cascade versus parallel

A number of hardware and software speech synthesizers have been described (Dudley, Riesz, and Watkins, 1939; Cooper, Liberman, and Borst, 1951; Lawrence, 1953; Stevens, Bastide, and Smith, 1955; Fant, 1959; Fant and Martony, 1962; Flanagan, Coker, and Bird, 1962; Holmes, Mattingly, and Shearme, 1964; Epstein, 1965; Tomlinson, 1966; Scott, Glace, and Mattingly, 1966; Liljencrants, 1968; Rabiner *et al.*, 1971; Klatt, 1972; Holmes, 1973). They employ different configurations to achieve what is hopefully the same result: high-quality approximation to human speech. A few of the synthesizers have stability and calibration problems, and a few have design deficiencies that make it impossible to synthesize a good voiced fricative, but many others have an excellent design. Of the best synthesizers that have been proposed, two general configurations are common.

In one type of configuration, called a parallel formant synthesizer (see, e.g., Lawrence, 1953; Holmes, 1973), the formant resonators that simulate the transfer function of the vocal tract are connected in parallel, as shown in the lower portion of Fig. 3. Each formant resonator is preceded by an amplitude control that determines the relative amplitude of a spectral peak (formant) in the output spectrum for both voiced and voiceless speech sounds. In the second type of configuration, called a cascade formant synthesizer (see, e.g., Fant, 1959; Klatt, 1972), sonorants are synthesized using a set of formant resonators connected in cascade, as shown in the upper part of Fig. 3.

The advantage of the cascade connection is that the relative amplitudes of formant peaks for vowels come out just right (Fant, 1956) without the need for individual amplitude controls for each formant. The disadvantage is that one still needs a parallel formant configuration for the generation of fricatives and plosive bursts (the vocal tract transfer function cannot be modeled adequately by five cascaded resonators when the sound source is above the larynx) so that cascade synthesizers are generally more complex in overall structure.

A second advantage of the cascade configuration is that it is a more accurate model of the vocal tract transfer

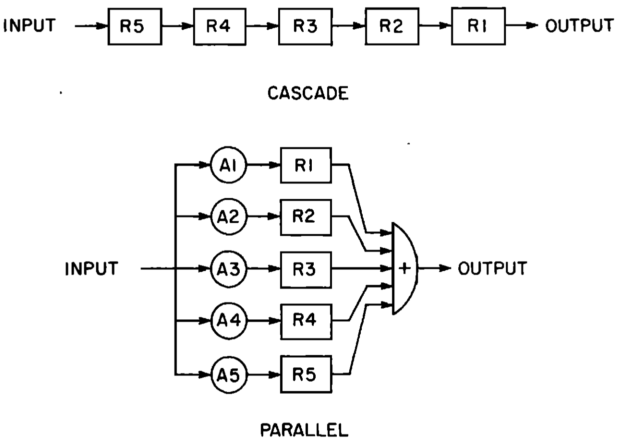

FIG. 3. The transfer function of the vocal tract may be simulated by a set of digital formant resonators R connected in cascade (the output of one feeding into the input of the next), or by a set of resonators connected in parallel (where each resonator must be preceded by an amplitude control A).

function during the production of non-nasal sonorants (Flanagan, 1957). As will be shown, the transfer functions of certain vowels are difficult to match using a parallel formant synthesizer. Although not optimal, a parallel synthesizer is particularly useful for generating stimuli that violate the normal amplitude relations between formants or if one wishes to generate, e.g., single-formant patterns.

The software simulation to be described has been programmed for normal use as a hybrid cascade/parallel synthesizer [Fig. 4(a)] or alternatively for special-purpose use as a strictly parallel synthesizer [Fig. 4(b)]. The experimenter must decide beforehand which configuration is to be employed. The change in configuration depends on the state of a single switch, and the program is smart enough to avoid performing unnecessary computations for resonators that are not used. To the extent that it is possible, the synthesizer has been adjusted so as to generate about the same output waveform whether the cascade/parallel configuration or the all-parallel configuration is selected.

Waveform sampling rate

Most of the sound energy of speech is contained in frequencies between about 80 and 8000 Hz (Dunn and White, 1940). However, intelligibility tests of bandpass filtered speech indicate that intelligibility is not measurably changed if the energy in frequencies above about 5000 Hz is removed (French and Steinberg, 1947). Speech low-pass filtered in this way sounds perfectly natural. Thus we have selected 5000 Hz (10000 samples per second) as the digital sampling rate of the synthesizer.

Parameter update rate

Control parameter values are updated every 5 ms. This is frequent enough to mimic even the most rapid of formant transitions and brief plosive bursts. If desired, the program can be modified to update parameter values only every 10 ms with relatively little decrement in output quality.

Digital resonators

The basic building block of the synthesizer is a digital resonator having the properties illustrated in Fig. 5. Two parameters are used to specify the input–output characteristics of a resonator, the resonant (formant) frequency F and the resonance bandwidth BW. Samples of the output of a digital resonator, $y(nT)$, are computed from the input sequence, $x(nT)$, by the equation

$$y(nT) = Ax(nT) + By(nT - T) + Cy(nT - 2T), \qquad (1)$$

where $y(nT - T)$ and $y(nT - 2T)$ are the previous two sample values of the output sequence $y(nT)$. The constants A, B, and C are related to the resonant frequency F and the bandwidth BW of a resonator by the impulse-invariant transformation (Gold and Rabiner, 1968)

$$\begin{aligned} C &= -\exp(-2PI\,BW\,T), \\ B &= 2\exp(-PI\,BW\,T)\cos(2PI\,F\,T), \qquad (2) \\ A &= 1 - B - C, \end{aligned}$$

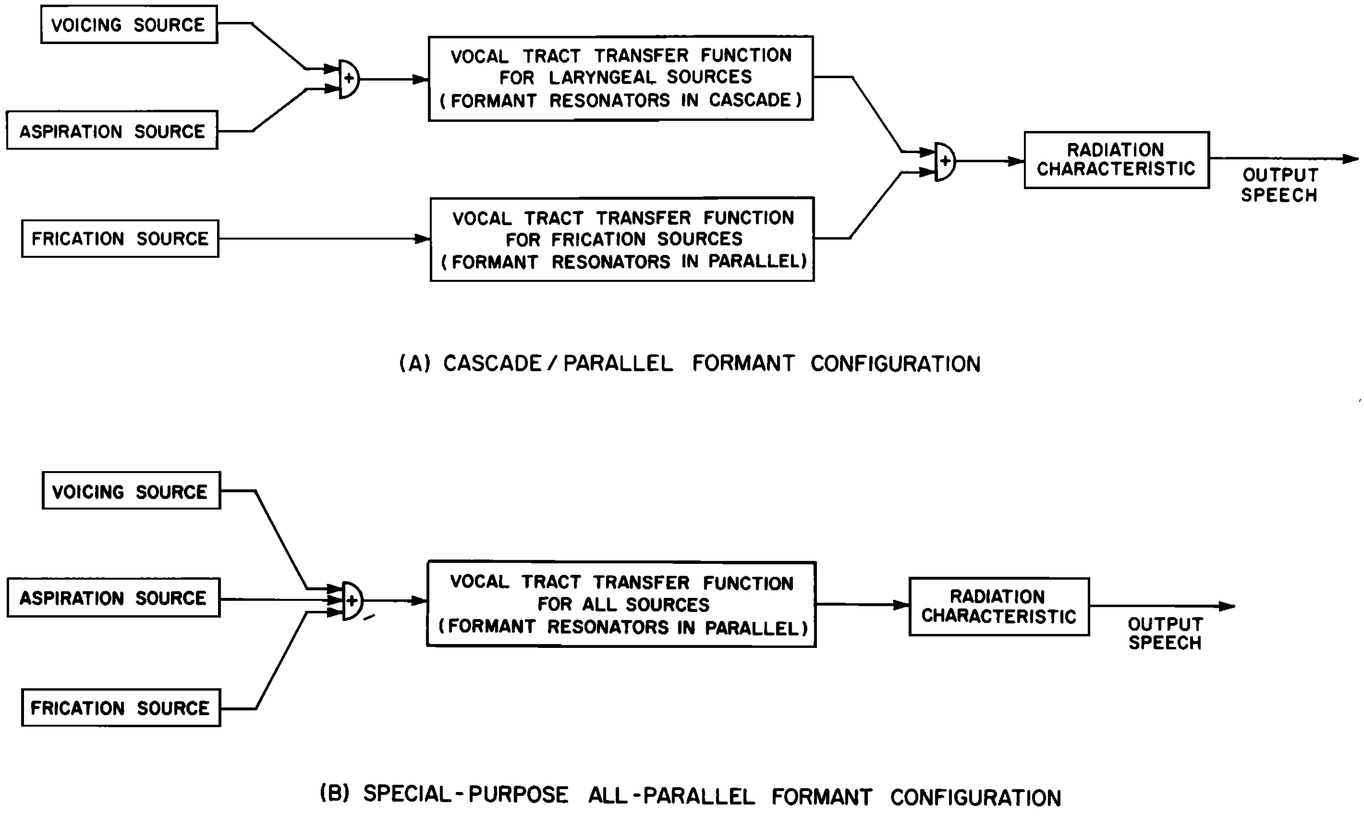
VOICING SOURCE
ASPIRATION SOURCE
FRICATION SOURCE
+
VOCAL TRACT TRANSFER FUNCTION FOR LARYNGEAL SOURCES (FORMANT RESONATORS IN CASCADE)
VOCAL TRACT TRANSFER FUNCTION FOR FRICATION SOURCES (FORMANT RESONATORS IN PARALLEL)
RADIATION CHARACTERISTIC
OUTPUT SPEECH
(A) CASCADE / PARALLEL FORMANT CONFIGURATION
VOICING SOURCE
ASPIRATION SOURCE
FRICATION SOURCE
+
VOCAL TRACT TRANSFER FUNCTION FOR ALL SOURCES (FORMANT RESONATORS IN PARALLEL)
RADIATION CHARACTERISTIC
OUTPUT SPEECH
(B) SPECIAL-PURPOSE ALL-PARALLEL FORMANT CONFIGURATION

FIG. 4. The synthesizer is normally used in a cascade/parallel configuration shown at the top, but may be used in an all parallel version shown at the bottom if one wishes to exercise independent control over formant amplitudes for vowels.

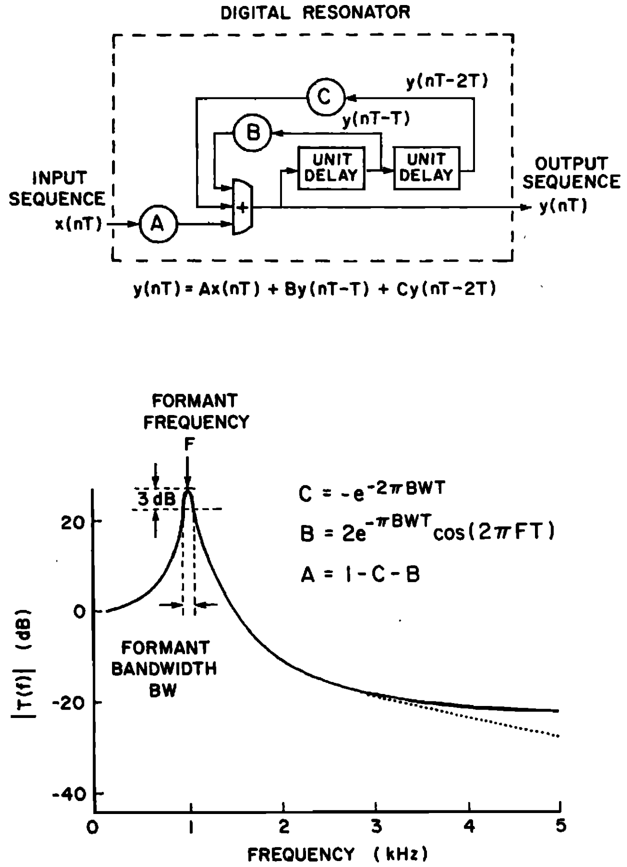

FIG. 5. The digital resonator shown in the form of a block diagram in the upper part of the figure has a transfer function (magnitude of the ratio of output to input in the frequency domain) as shown. In this example, $F=1000$ Hz and $BW=50$ Hz. The transfer function of a corresponding analog resonator is shown by the dotted line.

where *PI* is the familiar ratio of the circumference of a circle to its diameter. The constant *T* is one over the sampling rate and equals 0.0001 s in the present 5-kHz simulation.

The values of the resonator control parameters *F* and *BW* are updated every 5 ms, causing the difference equation constants to change discretely in small steps every 5 ms as an utterance is synthesized. Large sudden changes to these constants may introduce clicks and burps in the synthesizer output. Fortunately, acoustic theory indicates that formant frequencies must always change slowly and continuously relative to the 5-ms undate interval for control parameters.

A digital resonator is a second-order difference equation. The transfer function of a digital resonator has a sampled frequency response given by

$$T(f)=\frac{A}{1.-Bz^{-1}-Cz^{-2}}\ , \qquad (3)$$

where $z=\exp(j\,2PI\,f\,T)$, j is an imaginary number corresponding to the square root of -1, and f is frequency in Hz and ranges from 0 to 5 kHz. The transfer function has a (sampled) impulse response identical to a corresponding analog resonator circuit at sample times nT (Gold and Rabiner, 1968), but the frequency responses of an analog and digital resonator are not exactly the same, as can be seen in Fig. 5.

Digital antiresonator

An antiresonance (also called an antiformant or transfer-function zero pair) can be realized by slight modifications to these equations. The frequency response of an antiresonator is the mirror image of the response plotted in Fig. 5(i.e., replace dB by −dB). An antiresonator is used in the synthesizer to shape the spectrum of the voicing source and another is used to simulate the effects of nasalization in the cascade model of the vocal tract transfer function.

The output of an antiformant resonator, $y(nT)$, is related to the input $x(nT)$ by the equation

$$y(nT)=A'x(nT)+B'x(nT-T)+C'x(nT-2T), \qquad (4)$$

where $x(nT-T)$ and $x(nT-2T)$ are the previous two samples of the input $x(nT)$, the constants A', B', and C' are defined by the equations:

$$A'=1.0/A,\quad B'=-B/A,\quad C'=-CA\ , \qquad (5)$$

and where A, B, and C are obtained by inserting the antiresonance center frequency F and bandwidth BW into Eqs. (2).

Low-pass resonator

As a special case, the frequency F of a digital resonator can be set to zero, producing, in effect, a low-pass filter which has a nominal attentuation skirt of −12 dB per octave of frequency increase and a 3 dB down break frequency equal to $BW/2$. The voicing source contains a digital resonator RGP used as a low-pass filter that transforms a glottal impulse into a pulse having a waveform and spectrum similar to normal voicing. A second digital resonator RGS is used to low-pass filter the normal voicing waveform to produce the quasi-sinusoidal glottal waveform seen during the closure interval for an intervocalic voiced plosive.

Synthesizer block diagram

A block diagram of the synthesizer is shown in Fig. 6. There are 39 control parameters that determine the characteristics of the output. The name and range of values for each parameter are given in Table I. As can be seen from the table, one might wish to vary as many as 20 of the 39 parameters to achieve optimum matches to an arbitrary English utterance. The constant parameters in Table I have been given values appropriate for a particular male voice, and would have to adjusted slightly to approximate the speech of other male or female talkers. The list of variable control parameters is long compared with some synthesizers, but the emphasis here is on defining strategies for the synthesis of high quality speech. We are not concerned with searching for compromises that would minimize the information content in the control parameter specification.

I. SOURCES OF SOUND

There are two kinds of sound sources that may be activated during speech production (Stevens and Klatt,

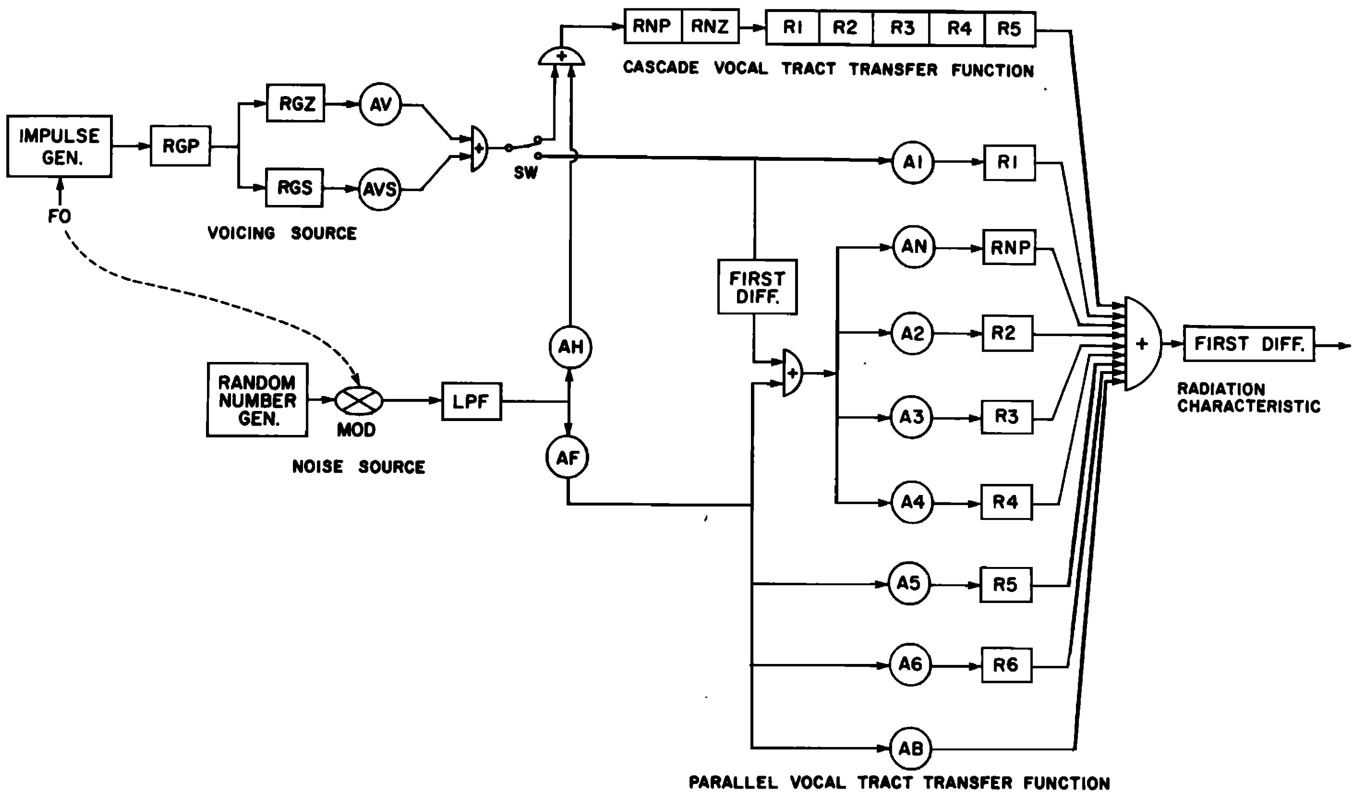

FIG. 6. Block diagram of the cascade/parallel formant synthesizer. Digital resonators are indicated by the prefix R and amplitude controls by the prefix A. Each resonator Rn has an associated resonant frequency control parameter Fn and a resonance bandwidth control parameter Bn.

1972). One involves quasi-periodic vibrations of some structure, usually the vocal folds. Vibration of the vocal folds is called voicing. (Other structures such as the lips, tongue tip, or uvula may be cuased to vibrate in sound types of some languages, but not in English.)

The second kind of sound source involves the generation of turbulence noise by the rapid flow of air past a narrow constriction. The resulting noise is called aspiration if the constriction is located at the level of the vocal folds, as for example during the production of the sound [*h*]. If the constriction is located above the larynx, as for example during the production of sounds such as [s], the resulting noise is called frication noise. The explosion of a plosive release also consists primarily of frication noise.

When voicing and turbulence noise generation co-exist, as in a voiced fricative such as [z] or a voiced [h], the noise is amplitude modulated periodically by the vibrations of the vocal folds. In addition, the vocal folds may vibrate without meeting in the midline. In this type of voicing, the amplitude of higher frequency harmonics of the voicing source spectrum is significantly reduced and the waveform looks nearly sinusoidal. Therefore the synthesizer should be capable of generating at least two types of voicing waveforms (normal voicing and quasi-sinusoidal voicing), two types of frication waveforms (normal frication and amplitude-modulated frication), and two types of aspiration (normal aspiration and amplitude-modulated aspiration). These are the only kinds of sound sources required for English, although trills and clicks of other languages may call for the addition of other source controls to the synthesizer in the future.

A. Voicing source

The structure of the voicing source is shown at the top left in Fig. 6. Variable control parameters are used to specify the fundamental frequency of voicing ($F0$), the amplitude of normal voicing (AV), and the amplitude of quasi-sinusoidal voicing (AVS).

An impulse train corresponding to normal voicing is generated whenever $F0$ is greater than zero. The amplitude of each impulse is determined by AV, the amplitude of normal voicing in dB. AV ranges from about 60 dB in a strong vowel to 0 dB when the voicing source is turned off. Fundamental frequency is specified in Hz; a value of $F0 = 100$ would produce a 100-Hz impulse train. The number of samples between impulses, $T0$, is determined by $SR/F0$, e.g., for a sampling rate of 10 000 and a fundamental frequency of 200 Hz, an impulse is generated every 50th sample.

Under some circumstances, the quantization of the fundamental period to be an integral number of samples might be perceived in a slow prolonged fundamental frequency transition as a sort of staircase of mechanical sounds (similar to the rather unnatural speech one gets by setting $F0$ to a constant value in a synthetic utterance), but the problem is not sufficiently serious to merit running the source model of the synthesizer at a higher sampling rate. If desired, some aspiration noise can be added to the normal voicing waveform to

TABLE I. List of control parameters for the software formant synthesizer. The second column indicates whether the parameter is normally constant (*C*) or variable (*V*) during the synthesis of English sentences. Also listed are the permitted range of values for each parameter, and a typical constant value.

N	*V/C*	Sym	Name	Min	Max	Typ
1	*V*	AV	Amplitude of voicing (dB)	0	80	0
2	*V*	AF	Amplitude of frication (dB)	0	80	0
3	*V*	AH	Amplitude of aspiration (dB)	0	80	0
4	*V*	AVS	Amplitude of sinusoidal voicing (dB)	0	80	0
5	*V*	F0	Fundamental freq. of voicing (Hz)	0	500	0
6	*V*	F1	First formant frequency (Hz)	150	900	450
7	*V*	F2	Second formant frequency (Hz)	500	2500	1450
8	*V*	F3	Third formant frequency (Hz)	1300	3500	2450
9	*V*	F4	Fourth formant frequency (Hz)	2500	4500	3300
10	*V*	FNZ	Nasal zero frequency (Hz)	200	700	250
11	*C*	AN	Nasal formant amplitude (dB)	0	80	0
12	*C*	A1	First formant amplitude (dB)	0	80	0
13	*V*	A2	Second formant amplitude (dB)	0	80	0
14	*V*	A3	Third formant amplitude (dB)	0	80	0
15	*V*	A4	Fourth formant amplitude (dB)	0	80	0
16	*V*	A5	Fifth formant amplitude (dB)	0	80	0
17	*V*	A6	Sixth formant amplitude (dB)	0	80	0
18	*V*	AB	Bypass path amplitude (dB)	0	80	0
19	*V*	B1	First formant bandwidth (Hz)	40	500	50
20	*V*	B2	Second formant bandwidth (Hz)	40	500	70
21	*V*	B3	Third formant bandwidth (Hz)	40	500	110
22	*C*	SW	Cascade/parallel switch	0(CASC)	1(PARA)	0
23	*C*	FGP	Glottal resonator 1 frequency (Hz)	0	600	0
24	*C*	BGP	Glottal resonator 1 bandwidth (Hz)	100	2000	100
25	*C*	FGZ	Glottal zero frequency (Hz)	0	5000	1500
26	*C*	BGZ	Glottal zero bandwidth (Hz)	100	9000	6000
27	*C*	B4	Fourth formant bandwidth (Hz)	100	500	250
28	*V*	F5	Fifth formant frequency (Hz)	3500	4900	3750
29	*C*	B5	Fifth formant bandwidth (Hz)	150	700	200
30	*C*	F6	Sixth formant frequency (Hz)	4000	4999	4900
31	*C*	B6	Sixth formant bandwidth (Hz)	200	2000	1000
32	*C*	FNP	Nasal pole frequency (Hz)	200	500	250
33	*C*	BNP	Nasal pole bandwidth (Hz)	50	500	100
34	*C*	BNZ	Nasal zero bandwidth (Hz)	50	500	100
35	*C*	BGS	Glottal resonator 2 bandwidth	100	1000	200
36	*C*	SR	Sampling rate	5000	20 000	10 000
37	*C*	NWS	Number of waveform samples per chunk	1	200	50
38	*C*	G0	Overall gain control (dB)	0	80	47
39	*C*	NFC	Number of cascaded formants	4	6	5

partially alleviate the problem and create a somewhat breathy voice quality.

B. Normal voicing

Ignoring for the moment the effects of RGZ, we see that the train of impulses is sent through a low-pass filter, RGP, to produce a smooth waveform that resembles a typical glottal volume velocity waveform (Flanagan, 1958). The resonator frequency FGP is set to 0 Hz and BGP to 100 Hz. The filtered impuses thus have a spectrum that falls off smoothly at approximately −12 dB per octave above 50 Hz. The waveform thus generated does not have the same phase spectrum as a typical glottal pulse, nor does it contain spectral zeros of the kind that often appear in natural voicing. These differences may be of some perceptual importance (for, e.g., naturalness of voice quality), in which case future versions of the program should provide for a more flexible voicing source specification.

The antiresonator RGZ is used to modify the detailed shape of the spectrum of the voicing source for particular individuals with greater precision that would be possible using only a single low-pass filter. The values chosen for FGZ and BGZ in Table I are such as to tilt the general voicing spectrum up somewhat to match the vocal characteristics of the author. The waveform and spectral envelope of normal voicing that is produced by sending an impulse train through RGP and RGZ are shown in Fig. 7(a).

C. Quasi-sinusoidal voicing

The amplitude control parameter AVS determines the amount of smoothed voicing generated during voiced fricatives, voiced aspirates, and the voicebars present in intervocalic voiced plosives. An appropriate wave shape for quasi-sinusoidal voicing is obtained by low-pass filtering an impulse by low-pass digital resonators RGP and RGS. The frequency control of RGS is set to zero to produce a low-pass filter, and BGS = 200 determines the cutoff frequency beyond which harmonics are

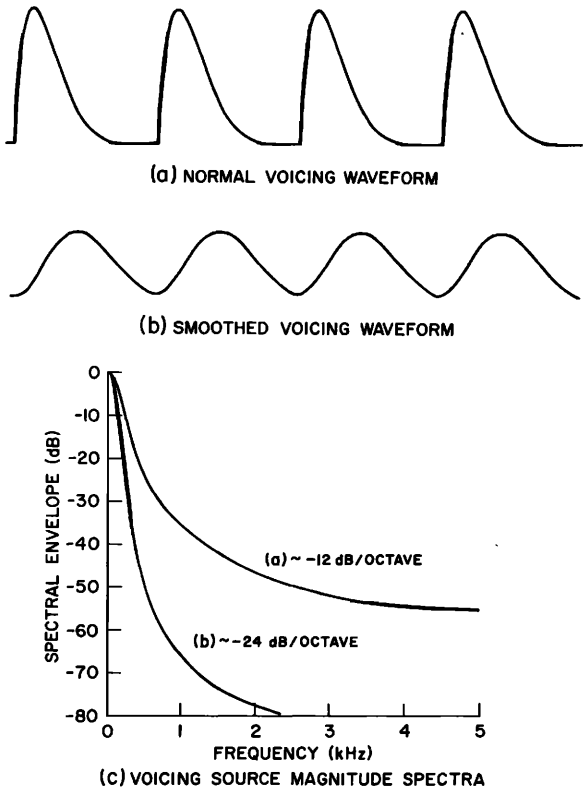

FIG. 7. Four periods from the synthetic waveforms of (a) normal voicing and (b) quasi-sinusoidal voicing are shown at the top, and the envelopes of the two resulting line spectra are shown in (c).

strongly attenuated.

The waveform and spectral envelope of quasi-sinusoidal voicing are shown in Fig. 7(b). After the effects of the vocal tract transfer function and radiation characteristic are imposed on the source spectrum, the output waveform of quasi-sinusoidal voicing contains significant energy only at the first and second harmonics of the fundamental frequency. AVS ranges from about 60 dB in a voicebar or strongly voiced fricative to 0 dB if no quasi-sinusoidal voicing is present. Some degree of quasi-sinusoidal voicing can be added to the normal voicing source (in combination with aspiration noise) to produce a breathy voice quality (e.g., AH = AV − 3, AVS = AV − 6).

D. Frication source

A turbulent noise source is simulated in the synthesizer by a pseudo-random number generator, a modulator, an amplitude control AF, and a −6 dB/octave low-pass digital filter LPF, as shown previously in Fig. 6. The spectrum of the frication source should be approximately flat (Stevens, 1971), and the amplitude distribution should be Gaussian. Signals produced by the random number generator have a flat spectrum, but they have a uniform amplitude distribution between limits determined by the value of the amplitude control parameter AF. A pseudo-Gaussian amplitude distribution is obtained in the synthesizer by summing 16 of the numbers produced by the random number generator.

In theory, the noise source is an ideal pressure source. The volume velocity of the frication noise depends on the impedance seen by the noise source. Since the vocal tract transfer function $T(f)$ relates source volume velocity to lip volume velocity, one must estimate noise volume velocity to determine lip output. In the general case, this is a complex calculation, but we will assume that source volume velocity is proportional to the integral of source pressure (an excellent approximation for a frication source at the lips because the radiation impedance is largely inductive, but only an approximation for other source locations). The integral is approximated by a first-order low-pass digital filter LPF that is shown in Fig. 6. Output samples from this filter, $y(nT)$, are related to the input sequence, $x(nT)$, by the equation

$$y(nT) = x(nT) + y(nT - T) \ .$$

As will be seen later, the radiation characteristic is a digital high-pass filter that exactly cancels out the effects of LPF. (For computational efficiency, the radiation characteristic can be moved into both the voicing source and the noise source; then the combination of radiation characteristic and the low-pass filter LPF can be removed from the noise source.)

An example of synthetic frication noise volume velocity that was generated in this way is shown in Fig. 8. The spectrum of this sample of noise fluctuates randomly about the expected long-term average noise spectrum (dashed line). Short samples of noise vary in their spectral properties due to the nature of random processes.

The output of the random number generator is amplitude modulated by the component labeled "MOD" in Fig. 6 whenever the fundamental frequency $F0$ and the amplitude of voicing AV are both greater than zero. Voiceless sounds (AV = 0) are not amplitude modulated because the vocal folds are spread and stiffened, and do not vibrate to modulate the airflow. The degree of

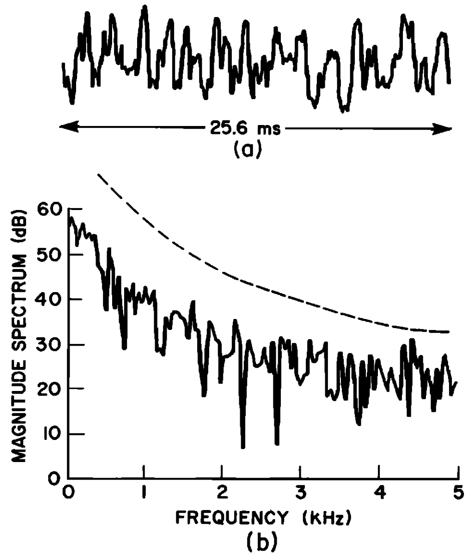

FIG. 8. A waveform segment and magnitude spectrum are shown of a 25.6-ms sample of frication noise. The expected long-term average spectrum of the output of the frication source is shown by the dashed curve. The dashed curve has been shifted up by 10 dB for clarity.

amplitude modulation is fixed at 50% in the synthesizer. The modulation envelope is a square wave with a period equal to the fundamental period. Experience has shown that it is not necessary to vary the degree of amplitude modulation over the course of a sentence, but only to ensure that it is present in voiced fricatives and voiced aspirated sounds.

The amplitude of the frication noise is determined by AF, which is given in dB. A value of 60 will generate a strong frication noise, while a value of zero effectively turns off the frication source.

E. Aspiration source

Aspiration noise is essentially the same as frication noise, except that it is generated in the larnyx. In a strictly parallel vocal tract model, AF can be used to generate both frication and aspiration noise. However, in the cascade synthesizer configuration, aspiration noise is sent through the cascade vocal tract model (since the cascade configuration is specially designed to model vocal tract characteristics for laryngeal sound sources), while fricatives require a parallel vocal tract configuration. Therefore separate amplitude controls are needed for frication and aspiration in a cascade/parallel configuration. The amplitude of aspiration noise sent to the cascade vocal tract model is determined by AH, which is given in dB. A value of 60 will generate strong aspiration, while a value of zero effectively turns off the aspiration source. Since frication and aspiration are generated by an identical process in the synthesizer, Fig. 8 describes the characteristics of the aspiration source as well.

F. Control of source amplitudes

Parameter values specifying source amplitudes AV, AVS, AF, and AH are adjusted by the user to new values every 5 ms. However, AV and AVS only have an effect on the synthetic waveform when a glottal impulse is issued. The reason for adjusting voicing amplitudes discontinuously at the onset of each glottal period is to prevent the creation of pops and clicks due to waveform discontinuities introduced by the sudden change in an amplitude control in the middle of a voicing period.

The noise amplitudes AF and AH are used to interpolate the intensity of the noise sources linearly over the 5-ms (50-sample) interval. (Thus there is a 5-ms delay in the attainment of a new amplitude value for a noise source.) Interpolation permits a more gradual onset for a fricative or [h] than would otherwise be possible. There is, however, one exception to this internal control strategy. A plosive burst involves a more-rapid source onset than can be achieved by 5-ms linear interpolation. Therefore, if AF increases by more than 50 dB from its value specified in the previous 5-ms segment, AF is (automatically) changed instantaneously to its new target value. We are presently evaluating the desirability of also injecting a step excitation of the vocal tract transfer function at this plosive release time so as to simulate the acoustic effect of a sudden release of the oral pressure behind the plosive occlusion.

G. Control of fundamental frequency

At times it is desired to specify precisely the timing of the first glottal pulse (voicing onset) relative to a plosive burst. For example, in the syllable [pa], it might be desired to produce a 5-ms burst of frication noise, 40 ms of aspiration noise, and voicing onset exactly 45 ms from the onset of the burst. Usually, a glottal pulse is issued in the synthesizer at a time specified by one over the value of the fundamental frequency control parameter value extant when the last glottal pulse was issued. However, if either AV or F0 is set to zero, no glottal pulse is issued during this 5-ms time interval; in fact no glottal pulses are issued until precisely the moment that both the AV and $F0$ control parameters become nonzero. In the case of the [pa] example above, both AV and $F0$ would normally be set to zero during the closure interval, burst, and aspiration phase, and AV would be set to about 60 dB and $F0$ to about 130 Hz at exactly 45 ms after the synthetic burst onset.

Since the update interval in the synthesizer is set to 5 ms, voice onset time can be specified exactly in 5-ms steps. If greater precision is needed, it would be necessary to change the parameter update interval from 5 ms (NWS = 50) to say 2 ms (NWS = 20).

H. Control of noise samples in a stimulus continuum

A pseudo-random number generator is used to generate both burst and aspiration for a plosive such as [pa]. The spectrum and intensity of a long sample of noise produced by the pseudo-random number generator can be expected to have the desired amplitude and spectral characteristics, but short samples of noise will vary considerably due to the random nature of pseudo-random numbers (recall Fig. 8). A particular brief noise sequence may have greater or lesser total intensity, or a peculiar spectral peak or valley not shared by other samples of noise that are used to generate a set of stimuli varying in voice onset time or burst frequency.

These random fluctuations in noise characteristics can cause some stimuli in a supposed continuum to stand out as different. When performing psychological experiments involving stimuli generated by the pseudo-random number generator, there are two ways to get around this problem. One is to use the same random number sequence in the generation of each member of the continuum by reinitializing the random number function, as is done by default if the synthesizer program is reloaded each time that a new stimulus is to be generated. The other way to minimize response fluctuations due to random noise is to generate several tokens of each stimulus type with different initial values given to the random number generator, and average listener responses over each type to try to wash out token variations.

II. VOCAL TRACT TRANSFER FUNCTIONS

The acoustic characteristics of the vocal tract are determined by its cross-sectional area as a function of distance from the larynx to the lips. The vocal tract forms a nonuniform transmission line whose behavior can be determined for frequencies below about 5 kHz by solving a one-dimensional wave equation (Fant, 1960). (Above 5 kHz, three-dimensional resonance modes would have to be considered.) Solutions to the wave equation result in a transfer function that relates samples of the glottal source volume velocity to output volume velocity at the lips.

The synthesizer configuration in Fig. 6 includes components to realize two different types of vocal tract transfer function. The first, a cascade configuration of digital resonators, models the resonant properties of the vocal tract whenever the source of sound is within the larynx. The second, a parallel configuration of digital resonators and amplitude controls, models the resonant properties of the vocal tract during the production of frication noise. The parallel configuration can also be used to model vocal tract characteristics for laryngeal sound sources, although the approximation is not quite as good as in the cascade model, see below.

A. Cascade vocal tract model

Assuming that the one-dimensional wave equation is a valid approximation below 5 kHz, the vocal tract transfer function can be represented in the frequency domain by a product of poles and zeros. Furthermore, the transfer function contains only about five complex pole pairs and no zeros in the frequency range of interest, as long as the articulation is non-nasalized and the sound source is at the larynx (Fant, 1960). The transfer function conforms to an all-pole model because there are no side-branch resonators or multiple sound paths. (The glottis is partially open during the production of aspiration so that the poles and zeros of the subglottal system are often seen in aspiration spectra; the only way to approximate their effects in the synthesizer is to increase the first formant bandwidth to about 300 Hz. The perceptual importance of the remaining spectral distortions caused by the poles and zeros of the subglottal system is probably minimal).

Five resonators are appropriate for simulating a vocal tract with a length of about 17 cm, the length of a typical male vocal tract, because the average spacing between formants is equal to the velocity of sound divided by half the wavelength, which works out to be 1000 Hz. A typical female vocal tract is 15 to 20% shorter, suggesting that only four formant resonators be used to represent a female voice in a 5 kHz simulation (or that the simulation should be extended to about 6 kHz). It is suggested that the voices of women and children be approximated by setting the control parameter NFC to 4, thus removing the fifth formant from the cascade branch of the block diagram shown in Fig. 6. For a male talker with a very long vocal tract, it may be necessary to add a sixth resonator to the cascade branch. As currently programmed, NFC can be set to four, five, or six formants in the cascade branch. Any change to NFC implies a change in the effective length of the vocal tract. NFC should not be used simply to remove a formant already present below 5 kHz because the spectrum of the resulting sound is tilted down in an in appropriate way: stimuli with a reduced number of formants must be generated using the all-parallel synthesizer configuration. (So such changes must be made with care.)

Ignoring for the moment the nasal pole resonator RNP and the nasal zero anti-resonator RNZ, the cascade model of Fig. 6, consisting of five formant resonators, has a volume velocity transfer function that can be represented in the frequency domain as a product of transfer functions identical to Eq. (3) (Gold and Rabiner, 1968):

$$T(f)=\prod_{n=1}^{5}\frac{A(n)}{1.-B(n)z^{-1}-C(n)z^{-2}}\ , \qquad (6)$$

where the constants $A(n)$, $B(n)$, and $C(n)$ are determined by the values of the nth formant frequency $F(n)$ and nth formant bandwidth $BW(n)$ by the relations given earlier in Eq. (2). The constants $A(n)$ in the numerator of Eq. (6) insure that the transfer function has a value of unity at zero frequency, i.e., the dc airflow is unimpeded. The magnitude of $T(f)$ is plotted in Fig. 9 for several values of formant frequencies and formant bandwidths

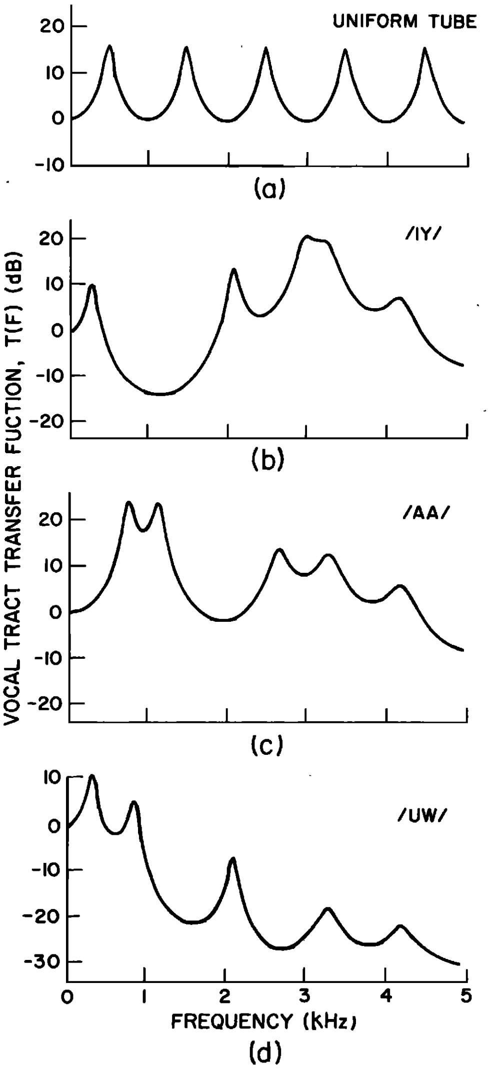

FIG. 9. The magnitude of the vocal tract transfer function is plotted for an ideal uniform vocal tract, and for the vowels [i], [a], and [u].

B. Relation to analog models of the vocal tract

The transfer function of the vocal tract can also be expressed in the continuous world of differential equations. Equation (6) is then rewritten as an infinite product of poles in the Laplace transform s plane:

$$T(f) = \prod_{n=1}^{\infty} \frac{s(n)\, s^*(n)}{[s+s(n)][s+s^*(n)]} \,, \tag{6a}$$

where $s=j\ 2\,PI\ f$, and the constants $s(n)$ and $s^*(n)$ are determined by the values of the nth formant frequency $F(n)$ and the nth formant bandwidth $BW(n)$ by the relations

$$s(n) = PI\ BW(n) + j\,2PI\,F(n),$$

$$s^*(n) = PI\,BW(n) - j\,2PI\,F(n)\ .$$

The two formulations (6) and (6a) are exactly equivalent representation of the transfer function for an ideal vocal tract configuration corresponding to a uniform tube closed at the glottis and having all formant bandwidths equal to, e.g., 100 Hz. The two formulations are indistinguishable at representing vocal tract transfer functions below 5 kHz. However, in a practical synthesizer, the infinite product of poles can only be approximated [e.g., by building five electronic resonators and a higher-pole correction network (Fant, 1959)].

C. Formant frequencies

Each formant resonator introduces a peak in the magnitude spectra shown in Fig. 9. The frequency of formant peak "n" is determined by the formant frequency control parameter Fn. (The amplitude of a formant peak depends not only on Fn and the formant bandwidth control parameter BWn, but also on the frequencies of the other formants, as will be discussed below.)

Formant frequency values are determined by the detailed shape of the vocal tract. Formant frequency values associated with different phonetic segments in the speech of the author will be presented in Sec. V. The frequencies of the lowest three formants vary substantially with changes to articulation (e.g., the observed range of $F1$ is from about 180 to 750 Hz, of $F2$ is 600 to 2300 Hz, and of $F3$ is 1300 to 3100 Hz for a typical male talker). The frequencies and bandwidths of the 4th and 5th formant resonators do not vary as much and could be held constant with little decrement in output sound quality. These higher frequency resonators help to shape the overall spectrum, but otherwise contribute little to intelligibility for vowels. The particular values chosen for the fourth and fifth formant frequencies (Table I) produce an energy concentration around 3 to 3.5 kHz and a rapid falloff in spectral energy above about 4 kHz, which is a pattern typical of many talkers.

D. Formant bandwadths

Formant bandwidths are a function of energy losses due to heat conduction, viscosity, cavity-wall motions, radiation of sound from the lips, and the real part of the glottal source impedance. Bandwidths are difficult to deduce from analyses of natural speech because of irregularities in the glottal source spectrum. Bandwidths have been estimated by other techniques such as using a sinusoidal swept-tone sound source (Fujimura and Lindqvist, 1971). Results indicate that bandwidths vary by a factor of 2 or more as a function of the particular phonetic segment being spoken. The primary perceptual effect of a bandwidth change is an increase or decrease in the effective intensity of a formant energy concentration [see Fig. 11(b) below] because formant bandwidths are narrower than a critical band (Carlson, Granstrom, and Klatt, 1979). Bandwidth variation is small enough that all formant bandwidths might be held constant in some applications, in which case only $F1$, $F2$ and $F3$ would be varied to simulate the vocal tract transfer functions for non-nasalized vowels and sonorant consonants.

E. Nasals and nasalization of vowels

It is not possible to approximate nasal murmurs and the nasalization of vowels that are adjacent to nasals with a cascade system of five resonators alone. More than five formants are often present in these sounds and formant amplitudes do not conform to the relations inherent in a cascade configuration because of the presence of transfer function zeros (Fujimura, 1961; 1962). Typical transfer functions for a nasal murmur and for a nasalized [Ĩ] are shown in Fig. 10. Nasalization introduces additional poles and zeros into the transfer function of the vocal–nasal tract due to the

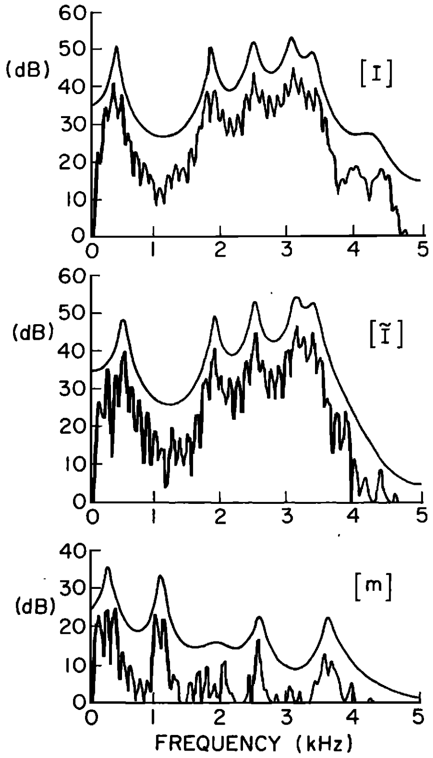

FIG. 10. Spectra are compared of a the vowel [I], the same vowel when nasalized, and a nasal murmur [m], all obtained from the recorded syllable "dim". The nasal murmur and the nasalized [Ĩ] have an extra transfer function pole pair and zero pair near $F1$. The extra peak and valley are not appearant in the linear prediction spectrum, but can be discerned in the pattern of harmonic amplitudes near $F1$ in the discrete Fourier transform spectrum.