presence of a side-branch resonator. The oral cavity forms the side-branch resonator in the case of a nasal murmur, while the nose should be considered a side-branch resonator in a nasalized vowel (because the amount of sound radiated through the nostrils is insignificant compared to the effect of the lowered velum on the formant structure of the sound output from the lips).

Nasalization of adjacent vowels is an important element in the synthesis of nasal consonants. The perceptually most important change associated with nasalization of a vowel is the reduction in amplitude of the first formant, brought on by the presence of a nearby low-frequency pole pair and zero pair. The first formant frequency also tends to shift slightly higher in most nasalized vowels.

Nasal murmurs and vowel nasalization are approximated by the insertion of an additional resonator RNP and anti-resonator RNZ into the cascade vocal tract model. The nasal pole frequency FNP can be set to a fixed value of about 270 Hz for all time. The nasal zero frequency FNZ should also be set to a value of about 270 Hz during non-nasalized sounds, but the frequency of the nasal zero must be increased during the production of nasals and nasalization. The RNP–RNZ pair is effectively removed from the cascade circuit during the synthesis of non-nasalized speech sounds if FNP = FNZ. Stragegies for controlling FNZ are given in Sec. V.

F. Parallel vocal tract model for frication sources

During frication excitation, the vocal tract transfer function contains both poles and zeros. The pole frequencies are temporally continuous with formant locations of adjacent phonetic segments because, by definition, the poles are the natural resonant frequencies of the entire vocal tract configuration, no matter where the source is located. Thus the use of vocalic formant frequency parameters to control the locations of frication maxima is theoretically well-motivated (and helpful in preventing the fricative noises from "dissociating" from the rest of the speech signal).

The zeros in the transfer function for fricatives are the frequencies for which the impedance looking back toward the larynx from the position of the frication source is infinite, since the series-connected pressure source of turbulence noise cannot produce any output volume velocity under these conditions. The effect of transfer-function zeros is twofold; they introduce notches in the spectrum and they modify the amplitudes of the formants. The perceptual importance of spectral notches is not great because masking effects of adjacent energy in format peaks limit the detectability of a spectral notch (Gauffin and Sundberg, 1974; Carlson, Granstrom, and Klatt, 1979). We have found that a satisfactory approximation to the vocal tract transfer function for frication excitation can be achieved with a parallel set of digital formant resonators having amplitude controls, and no antiresonators.

Formant amplitudes are set to provide frication excitation for selected formants, usually those associated with the cavity in front of the constriction (Stevens, 1972). The presence of any transfer function zeros is accounted for by appropriate settings of the formant amplitude controls. Relatively simple rules for determination of the formant amplitude settings (and bypass path amplitude values) as a function of place of articulation can be derived from a quantal theory of speech production (Stevens, 1972). The theory states that only formants associated with the cavity in front of the oral constriction are strongly excited. The theory is supported by the formant amplitude specifications for fricatives and plosive bursts in Sec. V. These amplitude control data were derived from trial-and-error attempts to match natural frication spectra.

There are six formant resonators in the parallel configuration of Fig. 6. A sixth formant has been added to the parallel branch specifically for the synthesis of very high frequency noise in [s, z]. The main energy concentration in these alveolar fricatives is centered on a frequency of about 6 kHz. This is above the highest frequency (5 kHz) that can be synthesized in a 10 000 sample/second simulation. However, in an [s], there is gradually increasing frication noise in the frequencies immediately below 5 kHz due to the low-frequency skirt of the 6-kHz formant resonance, and this noise spectrum can be approximated quite well by a resonator positioned at about 4900 Hz. We have found it better to include an extra resonator to simulate high-frequency noise than to move $F5$ up in frequency whenever a sibilant is to be synthesized because clicks and moving energy concentrations are thereby avoided.

Also included in the parallel vocal tract model is a bypass path. The bypass path with amplitude control AB is present because the transfer functions for [f, v, θ, δ, p, b] contain no prominent resonant peaks, and the synthesizer should include a means of bypassing all of the resonators to produce a flat transfer function.

During the production of a voiced fricative, there are two active sources of sound, one located at the glottis (voicing) and one at a constriction in the vocal tract (frication). The output of the quasi-sinusoidal voicing source is sent through the cascade vocal tract model, while the frication source excites the parallel branch to generate a voiced fricative.

G. Simualtion of the cascade configuration by the parallel configuration

The transfer function of the laryngeally excited vocal tract can also be approximated by five digital formant resonators connected in parallel. The same resonators that form the parallel branch for frication excitation can be used to synthesize any sonorant if suitable values are chosen for the formant amplitude controls.

The following rules summarize what happens to formant amplitudes in the transfer function $T(f)$ of a cascade model as the lowest five formant frequencies and bandwidths are changed. These relations follow directly from Eq. (6) under the assumption that each formant frequency $F(n)$ is at least 5 to 10 times as large as the formant bandwidth $BW(n)$:

1. The formant peaks in the transfer function are equal for the case that formant frequencies are set to 500, 1500, 2500, 3500, and 4500 Hz and formant bandwidths are set to be equal at 100 Hz. This corresponds to a vocal tract having a uniform cross-sectional area, a closed glottis, open lips (and a nonrealistic set of bandwidth values), as shown in part (a) of Fig. 11.

2. The amplitude of a formant peak is inversely proportional to its bandwidth. If a formant bandwidth is doubled, that formant peak is reduced in amplitude by 6 dB. If the bandwidth is halved, the peak is increased by 6 dB, as shown in part (b) of Fig. 11.

3. The amplitude of a formant peak is proportional to formant frequency. If a formant frequency is doubled, that formant peak is increased by 6 dB, as shown in part (c) of Fig. 11. [This is true of $T(f)$ but not of the resulting speech output spectrum since the glottal source spectrum falls off at about −12 dB/octave of frequency increase and the radiation characteristic imposes a +6 dB octave spectral tilt, resulting in a net change in formant amplitude of +6 −12 +6 = 0 dB.]

4. Changes to a formant frequency also affect the amplitudes of higher formant peaks by a factor proportional to frequency squared. For example, if a formant frequency is halved, amplitudes of all higher formants are decreased by 12 dB, i.e., one half squared, as shown in part (c) of Fig. 11.

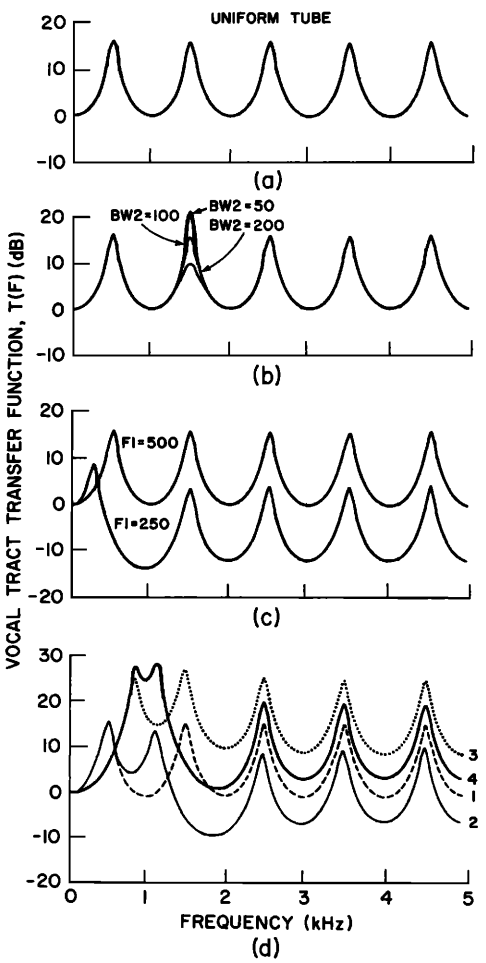

FIG. 11. Vocal tract transfer function Illustrating (a) the transfer function of a uniform vocal tract, (b) the influence of a change to a formant bandwidth, (c) changes to formant amplitudes caused by a shift in the frequency of a lower formant, and (d) the increase in formant amplitudes when two formants get close in frequency, see text.

5. The frequencies of two adjacent formants cannot come any closer than about 200 Hz because of coupling between the natural modes of the vocal tract. However, if two formants approach each other by about this amount, both formant peaks are increased by an additional 3 to 6 dB, as shown in part (d) of Fig. 11.

The amplitudes of the formant peaks generated by the parallel vocal tract model have been constrained such that, if A1 to A5 are all set to 60 dB, the transfer function will approximate that found in the cascade model. This is accomplished (1) by adjusting the gain of the higher frequency formants to take into account frequency changes in lower formants [since a higher formant rides on the skirts of the transfer function of all lower formants in a cascade model (Fant, 1960)], (2) by incorporation rules to cause formant amplitudes to increase whenever two formant frequencies come into proximity, and (3) by using a first difference calculation to remove low-frequency energy from the higher formants; this energy would otherwise distort the spectrum in the region if $F1$ during the synthesis of some vowels (Holmes, 1973).

The magnitude of the vocal tract transfer functions of the cascade and parallel vocal tract models are compared in Fig. 12 for several vowels. The match is quite good in the vicinity of formant peaks, but the parallel model introduces transfer function zeros (notches) in the spectrum between formant peaks. The notches are of relatively little perceptual importance because energy in the formant peak adjacent to the notch on the low frequency side tends to make the detectability of a spectral notch (Gauffin and Sundberg, 1974).

Many early parallel synthesizers were programmed to add together formant outputs without filtering out the energy at low frequencies from resonators other than $F1$. In other cases, formant outputs were combined in alternating sign. The deleterious effects of these choices are illustrated in Fig. 13. As can be seen, some vowel spectra are poorly modeled in both of these parallel methods of synthesis. The perceptual degradation is less in the alternating-sign case because spectral notches are less perceptible than energy fill in a spectral valley between two formants. Comparison of Fig. 12 and Fig. 13 indicates that our parallel configuration is better than either of those shown in Fig. 13.

A nasal formant resonator RNP appears in the parallel branch to assist in the approximation of nasal murmurs and vowel nasalization during parallel formant synthesis of vowels. Neither the parallel nasal formant nor the parallel first formant resonator are needed in the normal cascade/parallel synthesizer configuration ($SW = 0$), but they are required for the simulation of nasalization in the special-purpose all-parallel configuration ($SW = 0$).

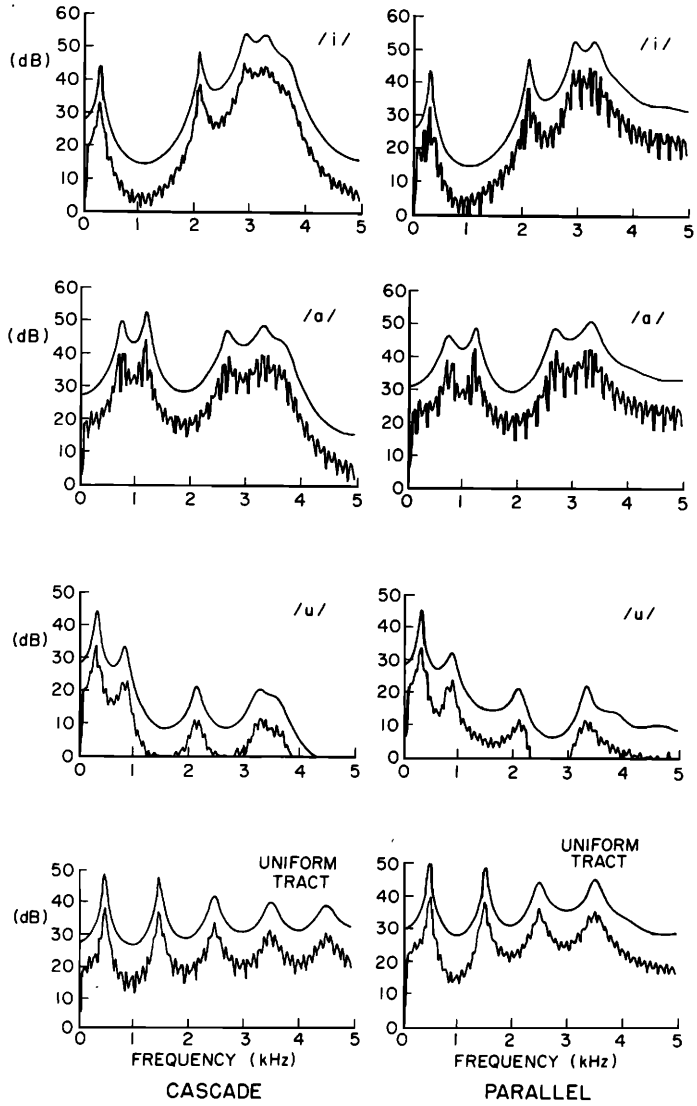

FIG. 12. Preemphasized output spectra (the DFT spectrum and a 14-pole LPC approximation) are shown for the vowels [i], [a], [u], and a uniform vocal tract when simulated by the theoretically correct cascade model and by the parallel approximation to the cascade model.

III. RADIATION CHARACTERISTIC

The box labeled radiation characteristic in Fig. 6 models the effect of directivity patterns of sound radiating from the head as function of frequency. The sound pressure measured directly in front of and about a meter from the lips is proportional to the temporal derivative of the lip-plus-nose volume velocity, and inversely proportional to r, the distance from the lips (Fant, 1960). The transformation is simulated in the synthesizer by taking the first difference of lip–nose volume velocity:

$$p(nT) = u(nT) - u(nT - T) \ . \tag{7}$$

The radiation characteristic adds a gradual rise in the overall spectrum, as shown in Fig. 14.

IV. HOST COMPUTER

The computer on which the software of Appendix B is installed must have digital-to-analog and analog-to-digital converters capable of transferring 10 000 12-bit waveform samples per second with precise control over the time between each sample. The computer should also have the appropriate audio equipment such as amplifiers, speaker, earphones, and tape recorder, as well as an external low-pass filter. The analog low-pass filter must have a sharp frequency cutoff near 5000 Hz, no appreciable ripple in the passband, and a nearly linear phase response in the passband. A filter that we constructed for this purpose is described in Appendix A.

A. Execution time

The synthesizer program is written in FORTRAN using floating point variables, so it is rather slow-running on some general-purpose digital computers (about 200 times real time on a PDP-11/40, but only about six times real time on a PDP-11/45 which has a faster floating-point multiply instruction). Even on a slower computer, the computational delay is not a serious handicap for most perceptual studies because stimuli can be generated and stored on disk for later presentation to subjects. Ultimately, it is hoped that the program will be implemented as a real-time digital device (Allen, 1977; Caldwell, 1979).

B. Graphical specification of variable control parameter data

A user could specify parametric data for an utterance to by synthesized in one of two general ways. One is to type in a sequence of ⟨time-value⟩ points for each variable control parameter and have the computer draw straight lines between them. However this is fairly time consuming and subject to error if no visual feedback in the form of a time plot of parameter values is provided.

In our laboratory, parameter specification can also be accomplished by selecting an individual parameter track for display. The computer is programmed to plot parameter values versus time as points on a refresh display, as shown in Fig. 15. Also displayed is the current ⟨time-value⟩ position of a cursor that is controlled by a hand-held pen. The pen can be moved over the surface of a graphical tablet (e.g., a Summagraphics model HW-1-11 data tablet digitizer or a Tektronix model 4953) to specify the parameter contour. A three-character symbol and a maximum y-axis value for the control parameter are also displayed.

It is relatively easy to watch the screen while moving the pen over the tablet with one hand. The other hand is positioned over three toggle switches which, when raised, have the following functions.

1. *Continuous input.* If the pen is swept horizontally past a 5 ms time point while toggle 1 is up, the stored value corresponding to this time is changed to the vertical value indicated by the pen at that moment.

2. *Straight line segment input.* A straight line is

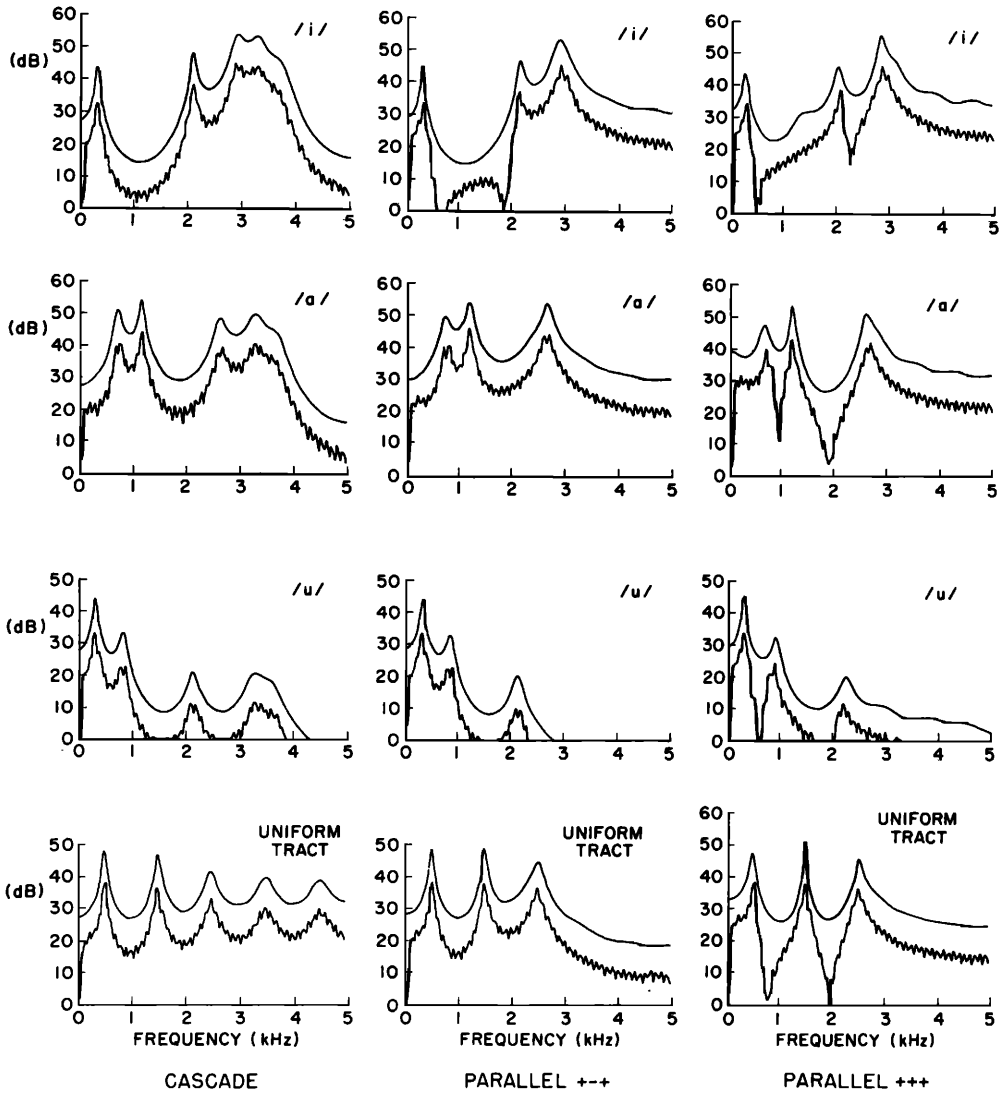

FIG. 13. Outputs from two common parallel synthesis configurations (in which resonator outputs are added with the same sign, or with alternating (+−+) signs, but without the first-difference preemphasis for $R2$ and $R3$, as in our model) are compared with the theoretically correct cascade output.

drawn from the previous ⟨time-value⟩ point to the point nearest the cursor at the moment the toggle is raised. The display is frozen until the toggle is returned to a down position. If no previous ⟨time-value⟩ point had been established, the action taken is to simply set the nearest point to the value of the cursor at the moment the toggle was raised.

3. *Set parameter track to a constant.* All time points are set to the value of the cursor at the moment the toggle is raised, and the displayed cursor position is frozen until the toggle is lowered.

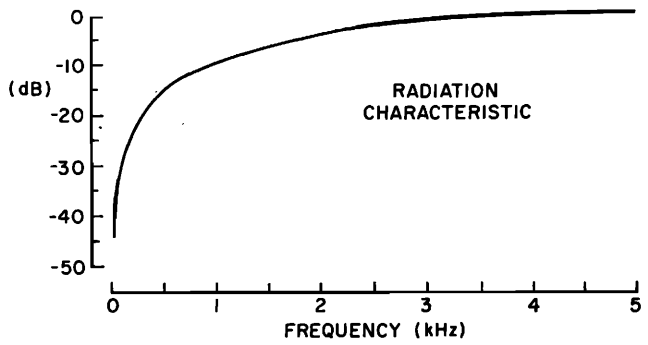

FIG. 14. Transfer function of the radiation characteristic.

In most experimental situations, we have found it advantageous to specify parameter contours in term of straight line segments, using toggle 2.

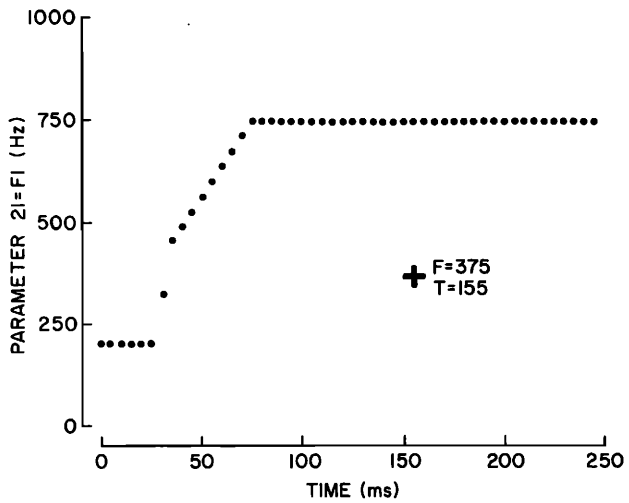

FIG. 15. A display of the first formant control parameter for the synthesis of the syllable (ba) is shown as it would appear on the computer scope. Formant frequency data are specified every 5 ms. The pen-controlled cursor position is indicated by a cross near the center of the plot.

C. Other software

The following executive commands are available for the manipulation of synthetic and natural waveforms:

1. Select a waveform buffer. At the beginning of a session, the user specifies the number and length of a set of disk waveform buffers. At any point in the session, one of these buffers is "current" and may be viewed or otherwise manipulated.

2. Synthesize a waveform. Take the control parameter files that have been drawn, and compute a synthetic waveform which is placed in the current waveform buffer.

3. Display the waveform. A 50-ms segment of the current waveform buffer is displayed. A variable knob is used to select the desired portion of the waveform for viewing.

4. Listen. The contents of the current waveform buffer are played out through the digital-to-analog converter.

5. Listen to all. The contents of each waveform buffer are played out in turn through the A/D converter with pauses of NPAUSE ms between waveforms.

6. Digitize. Begin to digitize a waveform obtained from a tape recorder or microphone and place it into the current waveform buffer. Stop digitizing when the duration of the buffer is exceeded.

7. Edit waveform beginning time. Chop off that portion of the current waveform before a centrally displayed waveform cursor, i.e., redefine the beginning of the waveform buffer. Use the knob to position the waveform relative to the fixed central waveform cursor before executing this command.

8. Edit waveform ending time. Chop off that portion of the current waveform after a centrally displayed cursor.

9. Generate DFT spectrum. Compute and display the magnitude of the discrete Fourier transform of a segment of waveform centered on the waveform display (see below for details).

10. Generate LP spectrum. Compute and display the magnitude of the linear prediction spectrum of a segment of waveform centered about the waveform display cursor (see below for details). Also list estimates of the five lowest formant frequencies.

11. Plot estimated fundamental frequency contour. Use the autocorrelation method to plot fundamental frequency versus time for the contents of the current waveform buffer.

12. Plot intensity contour. Compute rms intensity every 10 ms using a smooth time weighting window, and plot in decibels as a function of time for the contents of the current waveform buffer.

13. Generate an identification test. Randomize the set of waveform buffers to generate an identification test consisting of NTRIAL trials of each waveform buffer, with pauses of NPAUSE ms between trials.

14. Generate an AX *discrimination test.* Compare adjacent waveform buffers in a randomized order of NTRIAL trials with NPAUSE ms between trials and MPAUSE ms between members of an AX stimulus pair.

15. Generate an AXB *discrimination test.* Same as above, except the task is to detect whether $X = A$ or $X = B$.

Although not listed above, it is also desirable to have some means of saving parametric data and synthetic waveforms in digital form between computer sessions.

D. Spectral analysis

A spectral analysis capability is needed to verify that the output synthetic waveform has the desired spectral characteristics. Spectral analysis is particularly useful when attempting to mimic a digitized natural utterance. In our experience, if synthesis and natural spectral peaks are matched to within a couple of dB throughout the utterance, and the $F0$ contour and overall intensity contour are accurately duplicated, the synthetic utterance will be virtually indistinguishable from the original in both intelligibility and naturalness. Based on experience with several alternative spectral displays, we have found the following strategy to result in maximally useful spectral information.

A 25.6-ms segment of waveform is selected, the first difference of waveform samples is computed (to remove dc components and tilt the spectrum up slightly, somewhat analogous to the processing that takes place in the human peripheral auditory system), the differenced waveform is multiplied by a 25.6-ms Kaiser window with BETA = 7.0 (Kaiser, 1966) (so as to minimize the deleterious effects of having a nonintegral number of periods within the waveform segment), the discrete Fourier transform is computed, the magnitude of each spectral sample is converted to decibels and plotted as a set of lines connecting the ⟨dB, frequency⟩ samples, as shown in Fig. 16.

More useful in most cases is the linear prediction spectrum, which is obtained in the same way, except the discrete Fourier transform step is preceeded by a linear prediction analysis. The algorithm that we use is called the autocorrelation method (Makhoul, 1975) and uses 14 poles. The linear prediction coefficients are placed in the waveform buffer and padded with zeros before computing the discrete Fourier transform. Both DFT and linear prediction spectra can be superimposed, as in the figure, if there is some question as to the interpretation of the idealized spectra produced by the linear prediction algorithm.

Formant frequencies can often be estimated from the peaks observed in the linear prediction spectrum, such as is shown in Fig. 16, or by a root-solving technique (Markel and Gray, 1976). Plots of formant frequency trajectories can be obtained in this way, although the computations required make formant tracking of natural digitized waveforms a fairly lengthy process on most computers. One such plot is shown in Fig. 18 below. Extra points and missing points in this plot indicate that these quasi-formant tracks must be interpreted with care.

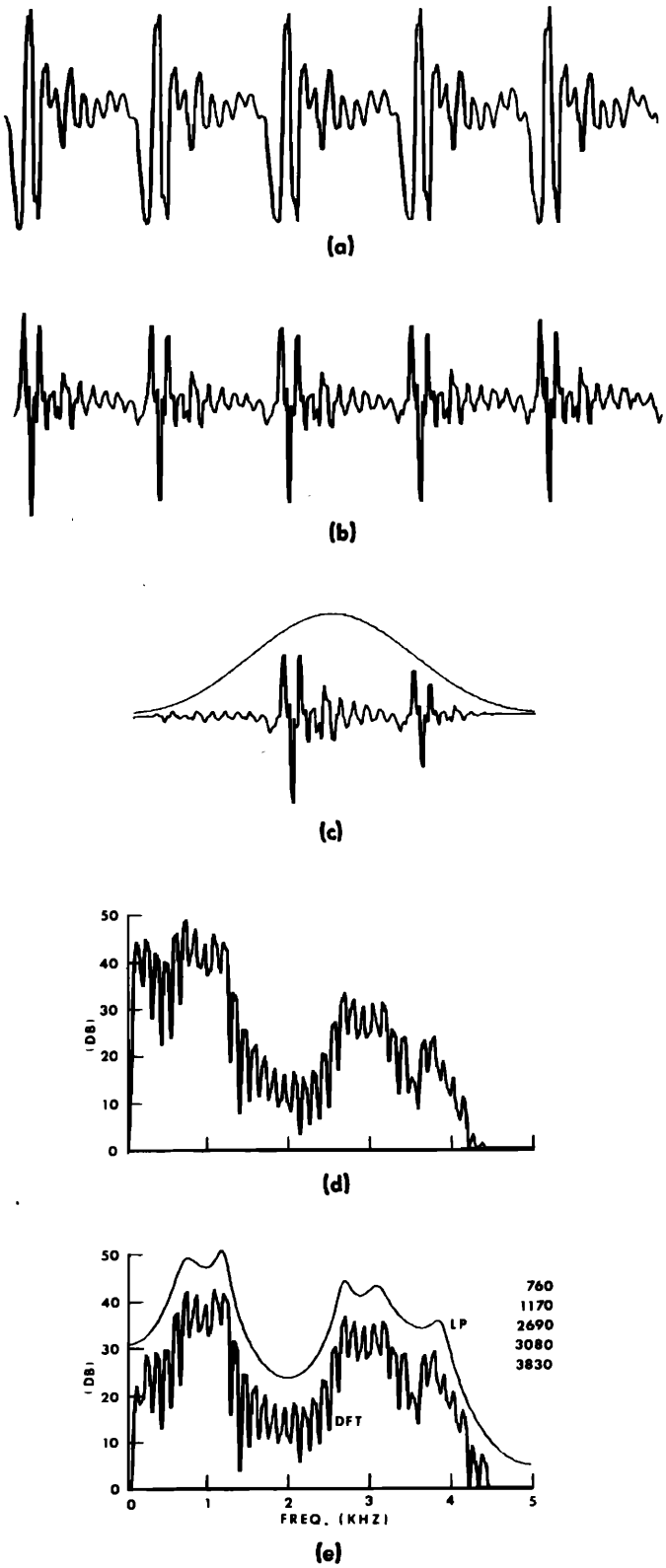

FIG. 16. The 25.6-ms (256 point) waveform segment extracted from a natural [a] with a fundamental frequency of 122 Hz, shown in (a) has been first differenced in (b), and multiplied by a Kaiser window in (c). The magnitude of the discrete Fourier transform of the non-preemphasized windowed waveform is shown in (d) and the DFT and magnitude of a linear prediction spectrum of the preemphasized windowed waveform is plotted in (e). Also listed are the frequencies of local maxima in the linear prediction spectrum; these maxima are usually good estimates of formant frequencies.

V. SYNTHESIS STRATEGY

General strategies for the synthesis of English syllables are beyond the scope of this paper, but the following paragraphs are intended to provide typical parameter values for a number of static speech sounds. The values presented below in Tables II and III should constitute a good starting point for the synthesis of an utterance if the procedures outlined below are adopted. The steps used to synthesize an utterance in our laboratory are described and a simple example is presented.

A. Synthesis of vowels

The control parameters that are usually varied to generate an isolated vowel are the amplitude of voicing AV, the fundamental frequency of vocal fold vibrations $F0$, the lowest three formant frequencies $F1$, $F2$, and $F3$, and bandwidths $B1$, $B2$, and $B3$. The fourth and fifth formant frequencies may be varied to simulate spectral details, but this is not essential for high intelligibility. To create a natural breathy vowel termination, the amplitude of aspiration AH and the amplitude of quasi-sinusoidal voicing AVS can be activated.

Table II includes suggested target values for variable

TABLE II. Parameter values for the synthesis of selected vowels. If two values are given, the vowel is diphthongized or has a schwa-like offglide in the speech of the author. The amplitude of voicing, AV, and fundamental frequency, $F0$, must also be given contours appropriate for an isolated vowel.

Vowel	$F1$	$F2$	$F3$	$B1$	$B2$	$B3$
[iʸ]	310	2020	2960	45	200	400
	290	2070	2960	60	200	400
[ɪᵊ]	400	1800	2570	50	100	140
	470	1600	2600	50	100	140
[eʸ]	480	1720	2520	70	100	200
	330	2020	2600	55	100	200
[ɛᵊ]	530	1680	2500	60	90	200
	620	1530	2530	60	90	200
[æᵊ]	620	1660	2430	70	150	320
	650	1490	2470	70	100	320
[ɑ]	700	1220	2600	130	70	160
[ɔᵊ]	600	990	2570	90	100	80
	630	1040	2600	90	100	80
[ʌ]	620	1220	2550	80	50	140
[oʷ]	540	1100	2300	80	70	70
	450	900	2300	80	70	70
[uᵊ]	450	1100	2350	80	100	80
	500	1180	2390	80	100	80
[uʷ]	350	1250	2200	65	110	140
	320	900	2200	65	110	140
[ɚ]	470	1270	1540	100	60	110
	420	1310	1540	100	60	110
[aʸ]	660	1200	2550	100	70	200
	400	1880	2500	70	100	200
[aʷ]	640	1230	2550	80	70	140
	420	940	2350	80	70	80
[oʸ]	550	960	2400	80	50	130
	360	1820	2450	60	50	160

control parameters that are used to differentiate among English vowels (Klatt, in preparation). Formant frequency and bandwidth targets were obtained by trial-and-error spectral matching to a large set of CV syllables spoken by the author. Bandwidth values are often larger than closed-glottis values obtained by Fujimura and Lindqvist (1971) because the bandwidths of Table II have been adjusted to take into account changes to observed formant amplitudes caused by factors such as glottal losses and irregularities in the voicing source spectrum.

The amplitude of the voicing source, AV, is set to about 60 dB for a stressed vowel, and falls gradually by a few dB near the end of the syllable. The fundamental frequency contour for an isolated vowel can be approximated by a linear falling $F0$, e.g., from 130 to 100 Hz.

B. Synthesis of consonants

If the vowel is to be preceded by a consonant, additional control parameters may have to be varied. Table III includes target values for variable control parameters that are used to synthesize portions of English consonants (frication spectra of fricatives, burst spectra of plosives, nasal murmurs for nasals, and steady portions of sonorants).

The sonorant consonants [w], [y], [r], and [l] are similar to vowels and require the same set of control parameters to be varied in order to differentiate among them. Formant values given in Table III for the prevocalic sonorants [r] and [l] depend somewhat on the following vowel. The source amplitude, AV, for a prevocalic sonorant should be about 10 dB less than in the vowel. The sonorant [h] (not shown in Table III) can be synthesized by taking formant frequency and bandwidth parameters from the following vowel, increasing the first formant bandwidth to about 300 Hz, and replacing voicing by aspiration.

The fricatives characterized in Table III include both voiceless fricatives (AF = 60, AV = 0, AVS = 0) and voiced fricatives (AF = 50, AV = 47, AVS = 47). Formants to be excited by the frication noise source are determined by the amplitude controls $A2$, $A3$, $A4$, $A5$, $A6$, and AB. Values presented in the table are appropriate only for consonants before front vowels.

The amplitude of the parallel second formant, $A2$, is zero for all of these consonants before front vowels, but the second formant is a front cavity resonance for velars before nonfront vowels and $A2$ is should be set to about 60 dB. The values given for $F2$ and $F3$ are not only valid during the fricative, but also can serve as "loci" for the characterization of the consonant–vowel formant transitions before front vowels (Klatt, in preparation). These are virtual loci in that formant frequency values observed at the onset of voicing are somewhere between the locus and the vowel target fre-

TABLE III. Parameter values for the synthesis of selected components of English consonants before front vowels (see text for source amplitude values).

Sonor	$F1$	$F2$	$F3$	$B1$	$B2$	$B3$						
[w]	290	610	2150	50	80	60						
[y]	260	2070	3020	40	250	500						
[r]	310	1060	1380	70	100	120						
[l]	310	1050	2880	50	100	280						
Fric.	$F1$	$F2$	$F3$	$B1$	$B2$	$B3$	$A2$	$A3$	$A4$	$A5$	$A6$	AB
[f]	340	1100	2080	200	120	150	0	0	0	0	0	57
[v]	220	1100	2080	60	90	120	0	0	0	0	0	57
[θ]	320	1290	2540	200	90	200	0	0	0		28	48
[ð]	270	1290	2540	60	80	170	0	0	0	0	28	48
[s]	320	1390	2530	200	80	200	0	0	0	0	52	0
[z]	240	1390	2530	70	60	180	0	0	0	0	52	0
[š]	300	1840	2750	200	100	300	0	57	48	48	46	0
Affricate												
[č]	350	1800	2820	200	90	300	0	44	60	53	53	0
[ǰ]	260	1800	2820	60	80	270	0	44	60	53	53	0
Plosive												
[p]	400	1100	2150	300	150	220	0	0	0	0	0	63
[b]	200	1100	2150	60	110	130	0	0	0	0	0	63
[t]	400	1600	2600	300	120	250	0	30	45	57	63	0
[d]	200	1600	2600	60	100	170	0	47	60	62	60	0
[k]	300	1990	2850	250	160	330	0	53	43	45	45	0
[g]	200	1990	2850	60	150	280	0	53	43	45	45	0

Nasal	FNP	FNZ	$F1$	$F2$	$F3$	$B1$	$B2$	$B3$
[m]	270	450	480	1270	2130	40	200	200
[n]	270	450	480	1340	2470	40	300	300

quency—the amount of virtual transition being dependent on formant-cavity affiliations. The specification of frication spectra in the table is accurate only before front vowels in the speech of the author. Before back and rounded vowels, systematic changes are observed to the fricative spectra because of anticipatory coarticulation. Specification of control parameter values for consonants in any phonetic environment is beyond the scope of the present paper, but appropriate values are easily found by trial-and-error matching to natural speech.

The affricate parameters in Table III refer to the fricative portion of the affricate. Similarly, the plosive parameters in Table III refer to the brief burst of frication noise generated at plosive release. Formant frequency values again serve as loci for predicting formant positions at voicing onset. In addition to differences in source amplitudes, voiced and voiceless consonants differ in that $F1$ is higher and $B1$ is larger when the glottis is open.

The parameters that are used to generate a nasal murmur include the nasal pole and zero frequencies FNP and FNZ. The nasal pole and zero are used primarily to approximate vowel nasalization at nasal release by splitting $F1$ into a pole-zero-pole complex. The details of nasal murmurs that have been described by Fujimura (1962) are approximated by formant bandwidth adjustments rather than by the theoretically correct method of pole-zero insertion. The reason is that it is not possible to simulate both the higher-frequency pole-zero details of nasal murmurs and vowel nasalization simultaneously without moving the frequency of the nasal pole and zero very fast at release, which would generate an objectionable click in the output, and vowel nasalization has been found to be perceptually more important. A nasalized vowel is generated by increasing $F1$ by about 100 Hz, and by setting the frequency of the nasal zero to be the average of this new $F1$ value and 270 Hz (the frequency of the fixed nasal pole).

Not included in Tables II and III are unstressed allophones, postvocalic allophones, flaps, glottal stops, voicebars, and consonant clusters. Characterization of even the static properties of these phonetic segments is beyond the scope of the present paper, but it it hoped that the information contained in the tables can be combined with the synthesis strategy described below for the rapid synthesis of an arbitrary utterance.

C. Synthesis of a novel utterance

The first step in the preparation of a new utterance is to obtain a natural model. The availability of a naturally spoken utterance is important because experience has shown that not all of the synthesis control parameter values can be deduced from theoretical considerations, and an unnatural, marginally intelligible synthetic utterance often results if one relies entirely on available theory.

A broadband spectrogram of the spoken word is then produced in order to visualize general acoustic characteristics of the utterance and determine the approximate duration of its component acoustic events. Computer analyses described below can provide much of the same information, but it is easier to visualize the time–frequency–intensity relations in the recording if a spectrogram is available.

The utterance is then 5-kHz low-pass filtered, digitized at 10 000 sample/s, and saved as a disk file for subsequent direct comparisons with the synthetic imitation that is to be created. The utterance "string", as spoken by a female talker, will be used as an illustration of the steps required to achieve a close acoustic match to a natural model of a word. A spectrogram of the recorded word is shown at the left in Fig. 17.

The intensity-versus-time plot shown in Fig. 18(a) was obtained by computing rms energy in dB every 10 ms, using a 25.6-ms Kaiser weighting window centered on the time of each displayed point. Ultimately, the synthetic utterance should have a matching intensity pattern, although it is not eady to deduce values for the various amplitude controls that will result in a close match to this contour on the first try.

The voiced portion of the word "string" was further analyzed by computer routines that extract an estimate of voicing fundamental frequency, as shown in Fig. 18(b), and formant frequency trajectories, as shown in Fig. 18(c). Observed formant motions can be used directly to specify formant frequency control parameters $F1$ through $F5$ during voiced portions of the utterance (only four formants are seen below 5 kHz for many female speakers, in which case the control parameter NFC that controls the number of formants in the cascade branch of the synthesizer would be set to 4). The third formant is invisible in the formant track of Fig. 18, but its position can be deduced by examination of the spectra in Fig. 19.

Linear prediction spectra sampled at various times in the natural utterance are plotted in Fig. 19. The spectra were all obtained in the manner described previously, except that the average spectrum of the [s] frication noise that is shown in Fig. 19(a) was obtained using a time weighting window having a longer effective duration of 40 ms. The longer duration window provides a better estimate of stationary spectra.

The general procedure for synthesis of voiced sonorants (the [ɹɪŋ] of "string") is to use information such as appears in Figs. 17–19 to (1) set the number of formants in the cascade vocal tract model, (2) adjust the fundamental frequency contour to within 2 Hz and formant frequencies to within about 5% by matching dig-

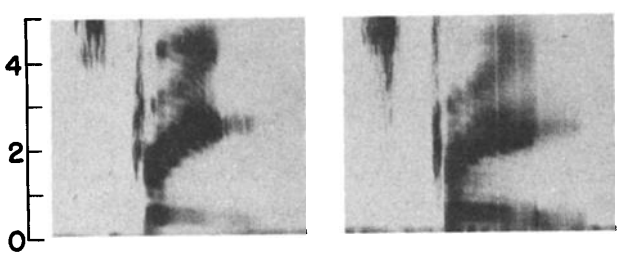

FIG. 17. Broadband spectrograms are compared of a natural and synthetic word, "string," spoken by a female talker.

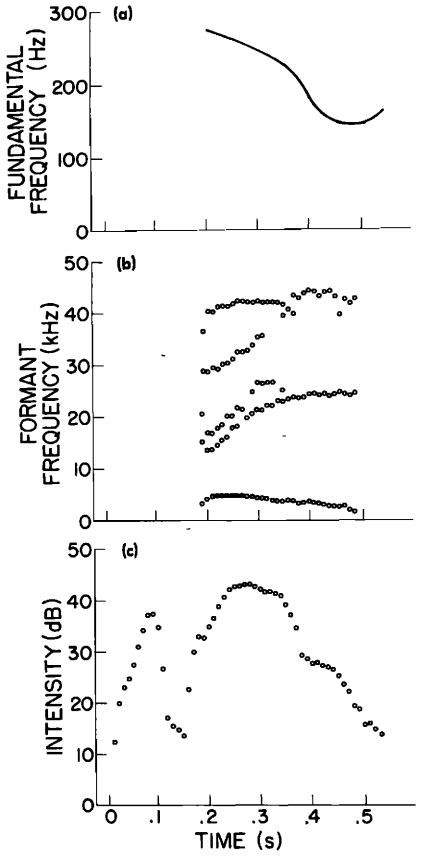

FIG. 18. The natural utterance "string" was analyzed by computer programs that produce plots of (a) fundamental frequency, (b) formant frequencies, and (c) intensity versus time.

itized waveform and spectra every 10 ms during transitions and every 50 ms during quasi-stationary intervals, (3) modify the source spectrum and/or formant bandwidths in order to set relative formant amplitudes to within about 2 dB, and (4) use the voicing source amplitude control, AV, to set the overall intensity contour to within about 2 dB. An intelligible utterance and a convincing imitation of the speaker are likely to result from satisfying these criteria (Holmes, 1961).

Detailed examination of the Fourier spectra of sonorant intervals indicates the presence of some aspiration noise during voicing for this speaker, i.e., the spectrum is less than perfectly harmonic at higher frequencies. Therefore, AH was adjusted to follow the same contour as AV, but with a value about 3 dB less. This produces a somewhat breathy voice quality that is typical of many female talkers. It also (fortunately) alleviates a fundamental problem of digital synthesis having to do with the perfect periodicity of portions of the synthesis when control parameter values do not change very much. Voiced sounds are often observed to become unpleasantly mechanical sounding as fundamental frequency is increased for female and children's voices, but a little aspiration noise breaks up the regularity of the harmonic structure.

The general procedure for synthesis of fricatives, affricates, and plosive bursts is to use the information in Figs. 17–19 to (1) select which formants are excited on the basis of continuity between the spectral peaks of the noise and formant peaks in adjacent voiced intervals, (2) set formant frequencies, (3) use formant bandwidth controls and parallel-branch formant amplitude controls to adjust the amplitude and width of peaks in the frication spectra, and (4) use AF, AV, and AVS to adjust the overall intensity contour and the mixture of periodic voicing to aperiodic noise (the discrete Fourier transform is useful for this purpose since it

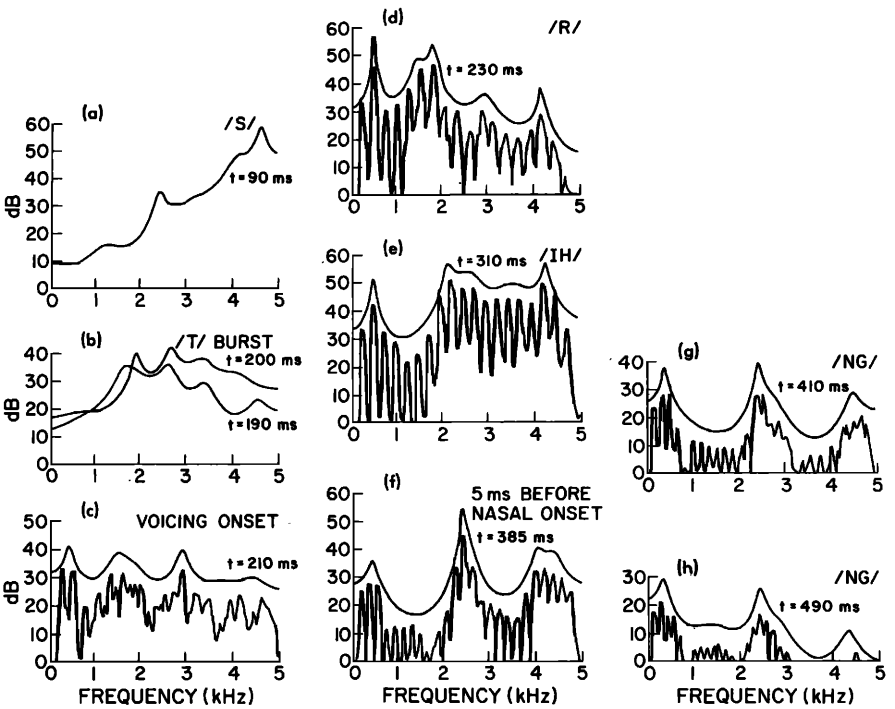

FIG. 19. Linear prediction spectra are plotted (a) during [s] noise, (b) centered on the [t] burst, (c) at voicing onset for the [r], (d) during the sonorant-vowel transition, (e) at the midpoint of the vowel, (f) just prior to closure, and (g)–(h) during the nasal murmur.

indicates individual harmonic amplitudes). For those formants that are invisible during frication generation, formant frequencies are filled in from continuity constraints and theoretical considerations.

For example, in order to synthesize the observed [s] spectrum, it is necessary to produce a strong spectral peak near 5 kHz. To do this, $F6$ can be positioned to 4.9 kHz and turned on at the appropriate time. Default values are chosen for the other formant frequencies and formant bandwidths are adjusted on a trial-and-error basis in order to match the [s] spectrum of Fig. 19.

VI. CONCLUSIONS

We have described a flexible software synthesizer that can run on any laboratory computer having sufficient core, peripheral equipment, and a Fortran compiler. The software is included as an appendix. The synthesizer and an associated control program can be used by a novice with minimal training and thus can serve as a research tool for those whose primary interest is not speech synthesis per se, but, e.g., the perception of speech and the relative perceptual importance of different acoustic cues to phonetic contrasts. It should be easier to replicate the results of perceptual experiments performed using this synthesizer because the synthesizer is fully documented here, and hopefully the control parameter values for the stimuli of an experiment will be carefully specified in the publication. The synthesizer may also find application in programs for speech synthesis by rule (Klatt, 1976a), for computer audio response, and in a reading machine for the blind (Allen, 1977).

Experience over the past few years suggest that the synthesizer is sufficiently flexible to generate good imitations to most if not all male and female voices. It also appears possible to synthesize any phonetic sequence of English with excellent intelligibility if the steps outlined in the previous section are followed. A consonant–vowel synthesis cookbook that is based on this synthesizer is in preparation (Klatt, in preparation). Intelligibility tests of 337 different CV syllables produced by the rules contained in the cookbook indicate that better than 98% of the vowels and 95% of the consonants are identified correctly by trained phoneticians who are unfamiliar with synthetic speech.

ACKNOWLEDGMENT

This research was supported by an NIH grant.

APPENDIX A: EXTERNAL 5000 Hz LOW-PASS FILTER

An external low-pass filter is required to convert the staircase waveform that comes from the D/A converter into an analog signal containing energy only below 5000 Hz. A suitable filter for this purpose has been designed and built by Dr. Joe Teirney of the M. I. T. Lincoln Laboratories. It is a seventh-order passive eliptic low-pass filter with component values that are indicated in Fig. A1.

The steps required to fabricate this filter are (1) wind the coils on something like Ferroxcube Corp. 3622A-600387 to plus-or-minus 0.5% (coils should have Q's of over 100 at the frequencies to which they are tuned), (2) use trim capacitors to tune the three L-C combinations along the top of the ladder network to have resonant frequencies in isolation of 7765, 5030, and 5427 Hz, respectively, (3) trim the remaining capacitors of the circuit to 0.5%, and (4) use 1% resistors.

The magnitude of the filter transfer function is shown at the bottom of Fig. A1. All frequency components in the input signal above 5 kHz are attenuated by at least 40 dB, while frequency components below 4780 Hz are within 0.6 dB of no attenuation at all.

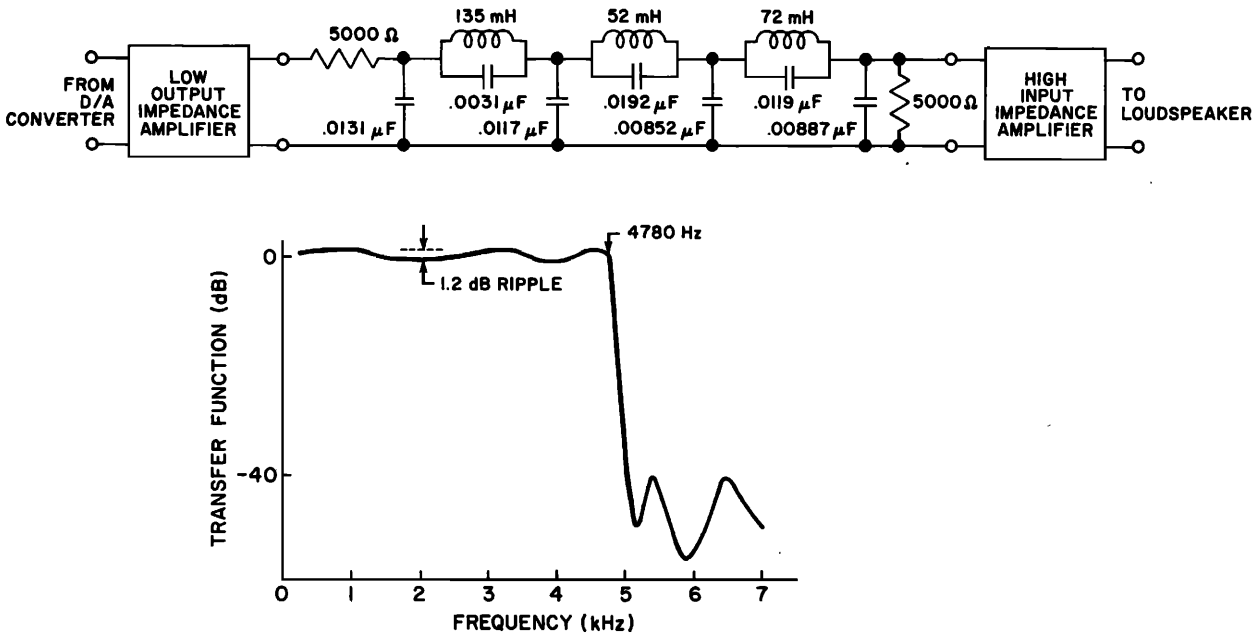

FIG. A1. The passive eliptic low-pass filter shown at the top has a transfer function with a very sharp cuttoff near 5000 Hz, as shown in the lower panel.