9.3, 9.4 Received 27 December 1966

Machine-Aided Formant Determination for Speech Synthesis

WILLIAM J. STRONG*

Air Force Cambridge Research Laboratories, L. G. Hanscom Field, Bedford, Massachusetts 01731

A semi-automatic analysis–synthesis scheme that can be viewed as a "manual formant vocoder" is described. A human operator makes decisions about formant positions on processed speech data. The parameters which result from the operator decisions are used to control a four-pole parallel synthesizer. Speech processed by the system had an error rate of 4.2% for vowels and 16.9% for consonants.

INTRODUCTION

THIS paper describes a method for speech analysis that is an outgrowth of an attempt to do speech synthesis by rule using a terminal analog synthesizer. In speech synthesis by rule, the rules accept an input string of phonemes and, based on the input string, generate control parameters that can then be used to control a speech synthesizer. The synthetic speech produced by the synthesizer can be subjected to various measures to determine its validity, but the ear must be the final criterion. However, the ear does not indicate in an explicit way wherein the control parameters may be improved. We feel, therefore, that being able to compare control data generated by rules with those extracted from the real speech of some talker can be a useful guide.

Our approach is similar to that of Holmes, Mattingly, and Shearme,[1] who started with a set rules and then modified these rules, guided by spectrographic analysis and listening. However, a basic difference between our approach and that of Holmes *et al.* is that, in their case, one knows the general bounds of the control parameters and tries to write and modify rules that will generate satisfactory control parameters within these bounds; whereas in our case, one uses analysis to derive detailed control parameters and tries to write rules that will generate these parameters. (The latter scheme has the characteristic of being closely related to the particular speaker whose speech is analyzed.) To implement our scheme, it was necessary to perform some extensive analysis on natural speech in order to determine detailed control parameters with which to compare rule-generated control parameters. We attempted, therefore, to determine control parameters (e.g., formant frequencies, formant amplitudes, and voicing frequency) sufficient to generate intelligible speech. We have not yet attempted to write rules that will give rise to similar control parameters.

We chose a terminal analog synthesizer because of the comparative ease of obtaining and modifying its control parameters. Of the two alternative configurations for terminal analog synthesizers (a cascade or a parallel combination of simple resonators), we chose the parallel combination of resonators for the following reasons: (1) The cascade combination of resonators requires some additional circuitry or some special configuration of the cascade circuitry for the synthesis of certain consonants; whereas the parallel combination of resonators can, at least in principle, handle consonants in the same manner as it handles vowels. (2) For modeling filters on a digital computer, there is no cumulative overflow problem with the parallel combination of resonators as there is with the cascade combination. (3) When a pole is moved by a discrete amount, noise resulting from this discrete change will propagate into the skirts of the pole. In a parallel synthesizer, this noise will tend to be masked by adjacent poles; whereas in the cascade synthesizer this noise will tend to be enhanced by each succeeding pole in the cascade.

In a sense, the parallel arrangement assumes that the speech pressure waveform can be encoded by specifying the frequency positions and amplitudes of the major peaks in the pressure spectrum. Whether the spectrum peaks result primarily from the vocal-tract configura-

* Present address: Dept. Phys., Brigham Young Univ., Provo, Utah 84601.

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

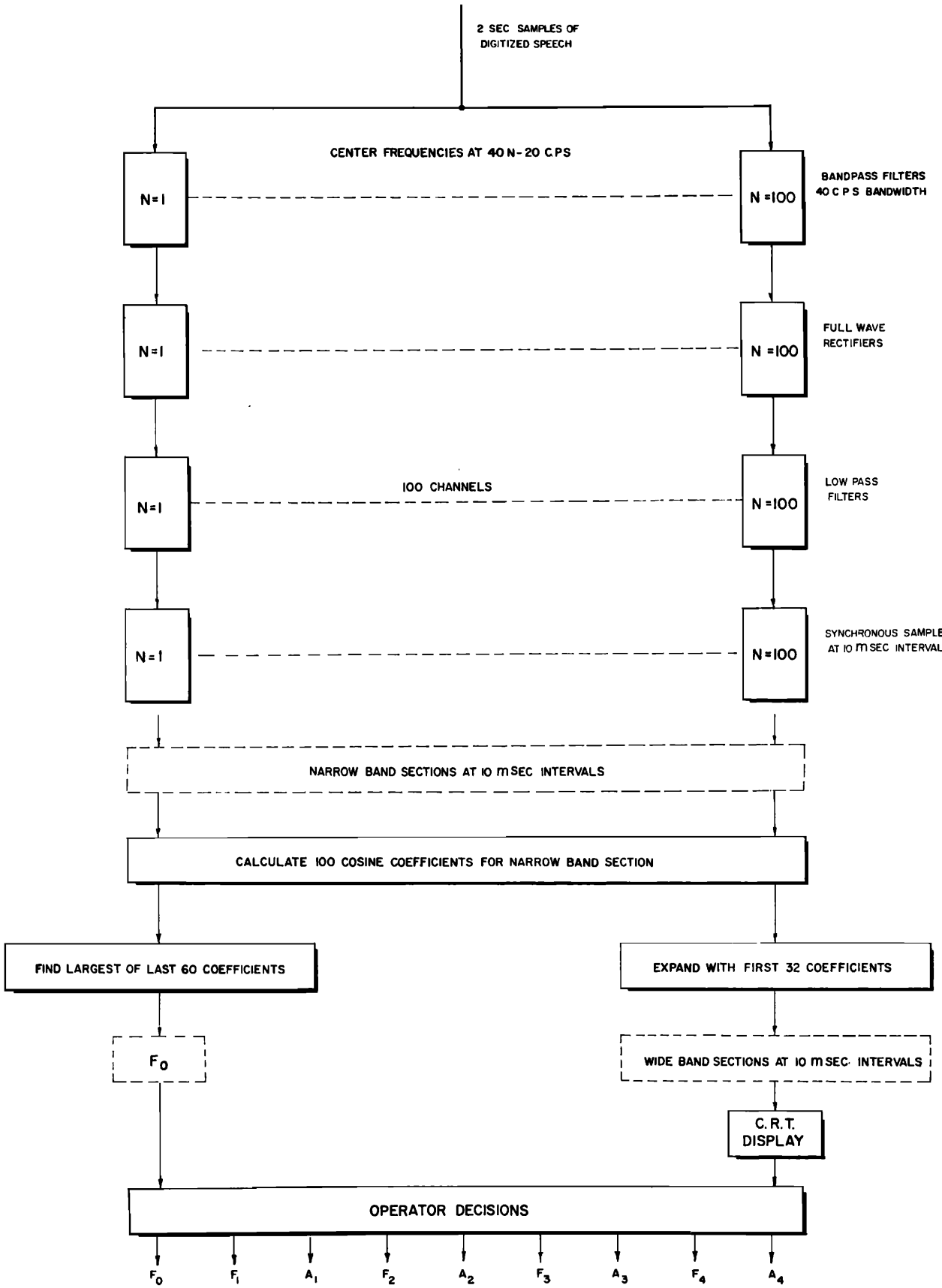

FIG. 1. Basic elements of the analysis.

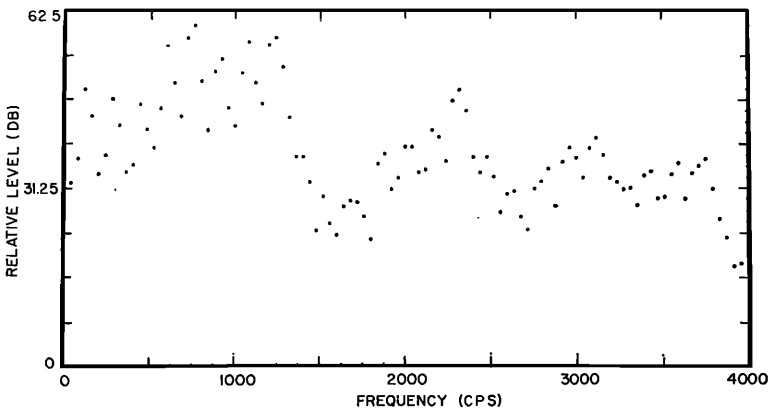

FIG. 2. CRT display of a "narrow-band section."

tion (as is the case with vowel sounds, etc.) or from the excitation, or from a combination of the two, is to some extent immaterial, since the spectrum peaks are to be tracked independently of their cause. The scheme can be viewed as a channel-type encoding device in which the number of channels is fixed (at four, in the present case), but in which the frequency position of each channel, along with its amplitude, is variable. In the present model, the transfer characteristics of any channel are those of a simple resonator and are not bandpass in nature as are those of a conventional channel vocoder. It should be noted, however, that the scheme implemented here does not claim to give a completely satisfactory match to pressure spectra but represents a possible step in that direction.

A particular semi-automatic method of analysis for determining formant information from natural speech is described. An important feature of the system is that it is a single-pass system—i.e., no iterative procedures like those in analysis-by-synthesis are used to aid in the determination of formant positions. On the other hand, the method has many arbitrary aspects and claims to be neither unique nor optimum. The scheme (including both analysis and synthesis) can be viewed as a "manual formant vocoder," in which decisions about formant positions are made by a human operator on processed speech data. It is felt that this is a useful approach because the human operator can presumably make more plausible decisions than can any automatic method presently available. Furthermore, it gives the human operator an opportunity to study, at least qualitatively, the kinds of decisions that an automatic procedure would be required to make and may therefore give rise to better means for automatic formant extraction. The adequacy of the scheme was tested by subjecting the synthetic speech (which results from controlling a synthesizer with the control parameters from the analysis) to an intelligibility test.

EXPERIMENTAL PROCEDURE

The basic elements of the analysis appear in Fig. 1. A 2-sec sample of digitized speech (low-pass filtered at 5000 cps and sampled 10 000 times sec) is processed through each of 100 computer-simulated channels. Each channel consists of a bandpass filter, followed by a full-wave rectifier and "low-pass filter." Each bandpass filter is a two-pole filter with a bandwidth of 40 cps at the 3-dB points. These were simulated in the computer by using the method of z transforms. The 40-cps bandwidth was chosen so that the analysis results would correspond roughly to a "narrow-band" spectrogram and so that the voicing harmonics for male speech could be resolved. The center frequencies of the bandpass filters are set to [40N-20] cps, where N is the channel number ($N=1, 2, \cdots, 100$). The low-pass filters for all 100 channels are synchronously sampled at 10-msec intervals (where each low-pass filter simply determines the maximum value of the wave coming from the full-wave rectifier during the sampling interval), and the 100 amplitudes are converted to decibels (in $\frac{1}{4}$-dB steps) and stored. The 100 values can be thought of as specifying a "narrow-band section" at 40-cps intervals in the frequency domain, an example of which is shown in Fig. 2. The time sequence of these "narrow-band sections" at 10-msec intervals then gives some description of the speech wave.

A technique that has characteristics of techniques described by Noll[2] and by Schroeder and Noll[3] is used to determine the fundamental frequency and a "wide-band section" from each narrow-band section. The technique involves the determination of 100 coefficients for a cosine series expansion of the narrow-band section. The lowest 32 of these coefficients are used to construct a smoothed section (the wide-band section).

The wide-band sections are displayed in time sequence on a CRT as shown in Fig. 3. (The lowest wide-band section in Fig. 3 is derived from the narrow-band

[2] A. M. Noll, "Short-Time Spectrum and 'Cepstrum' Techniques for Vocal Pitch Detection," J. Acoust. Soc. Am. 36, 296–302 (1964).

[3] M. R. Shroeder and A. M. Noll, "Recent Studies in Speech Research at Bell Telephone Laboratories (I)," Paper A21 in *Proceedings of the Fifth International Congress of Acoustics, Liège, 1965*, D. E. Commins, Ed. (Imprimerie Georges Thone, Liège, 1965).

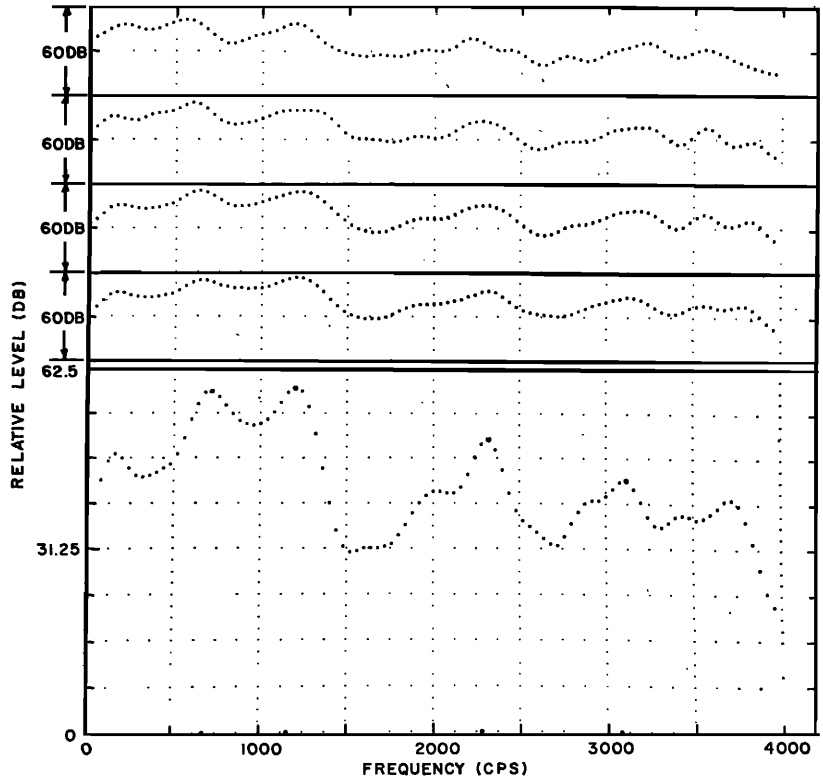

FIG. 3. CRT display of "wide-band sections."

section in Fig. 2.) By means of voltage-knob inputs to the computer, the operator specifies the positions of four poles (shown by the extra large points) on the lowest wide-band section in the display. The four upper wide-band sections (compressed in amplitude by a factor of 4 to 1) are the next ones in the time sequence and give the operator information about where the poles should be positioned when there is ambiguity in the lowest wide-band section. The four points at the bottom of the display show the pole positions for the preceding sample. In addition, the operator has a knowledge of the utterance being analyzed. There are, however, no iterative procedures to assist the operator in his decision making. When the pole positions have been determined to the satisfaction of the operator, the computer, under sense-switch control, stores the four pole frequencies and the four pole amplitudes, and the sequence is advanced to the next section. The procedure is repeated a total of 200 times for each 2-sec sample of speech.

Calculation of the fundamental frequency of voicing for a particular narrow-band section is based on which of the highest 60 cosine-series coefficients is the largest. (The procedure amounts to a version of the "cepstrum"

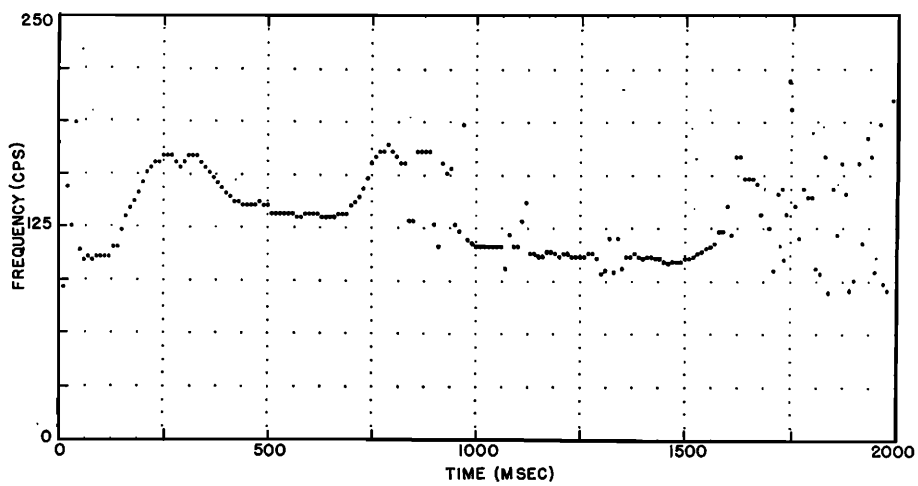

FIG. 4. Unmodified fundamental-frequency curve.

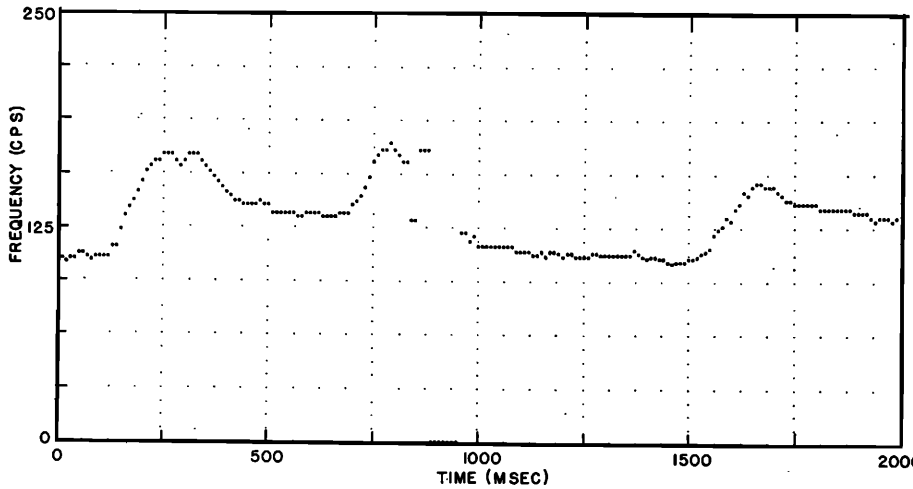

FIG. 5. Modified fundamental-frequency curve.

technique.[2]) An example of the fundamental frequency obtained by this method for the 2-sec utterance, *Robby will like you daddy-oh*, appears in Fig. 4. The determination is quite stable when the voicing is relatively strong but is unstable when the voicing is weak. The difficulty is by-passed for present purposes by manually filling in the fundamental frequency curves in the unstable portions by use of a light pen. Furthermore, it is possible for the operator to specify noise excitation (to be synchronous with the particular section under consideration) by means of a sense-switch input to the computer; this determination is made by the operator from a qualitative evaluation of the relative amounts of energy in different parts of the spectrum. The fundamental frequency curve from Fig. 4, with the unstable portions modified and with a short noise burst (specified by setting F0 equal to minus zero) for the /k/ in "like," is portrayed in Fig. 5.

After processing 200 sections, all nine parameters (four pole frequencies, four pole amplitudes, and the fundamental frequency) are displayed on a CRT, and modifications can be made with a light pen if desired. However, for the results of this paper, the formant frequencies and amplitudes were used as extracted from the wide-band sections without any light-pen modifications. The unmodified control parameters (excluding F0), used to synthesize "Robby will like you daddy-oh," appear in Fig. 6. The parameters are punched out on paper tape for permanent storage and for use with the synthesizer.

The synthesizer is shown in Fig. 7. The only unusual features about the synthesizer are the amplitude modu-

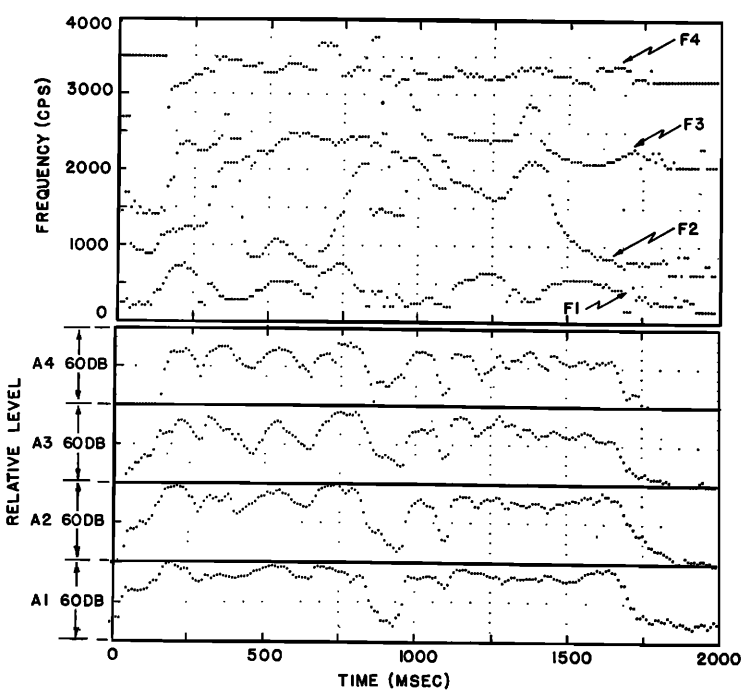

FIG. 6. Unmodified synthesizer control parameters.

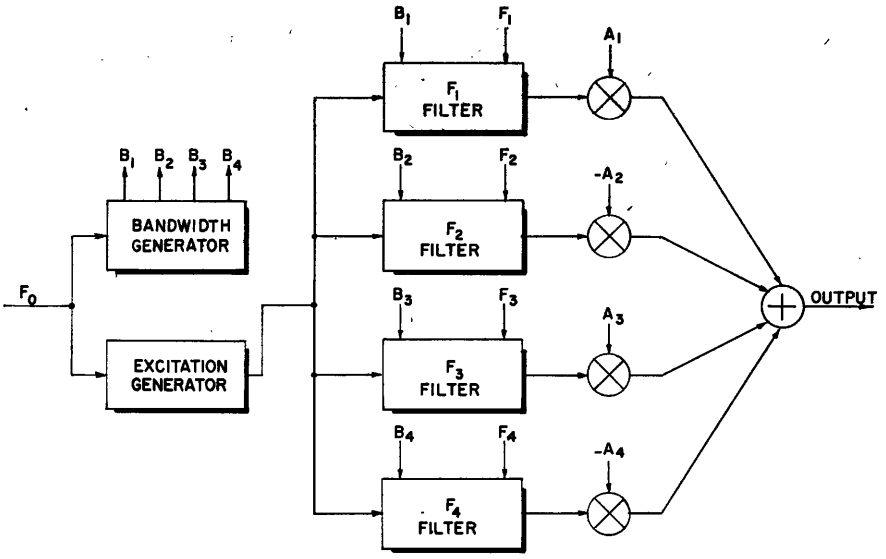

FIG. 7. Basic elements of the synthesizer.

lations following, instead of preceding, the poles (as is the more common procedure) and the phase reversing of the output from the even-numbered poles in an attempt to remove the zeros between consecutive poles. There is no specific provision made for introducing zeros into the synthesizer, even though they have some perceptual significance. The poles have a frequency characteristic given by

$$P(s)=\{2as/[(s+a)^2+\omega^2]\}\cdot[\omega/(s+\omega)],$$

where a is the half-bandwidth, ω is the center frequency, and s is the frequency of interest.

Nine signals are used to control the synthesizer: four formant frequencies (F1, F2, F3, and F4), four formant amplitudes (A1, A2, A3, and A4), and the voicing frequency (F0). These signals are specified at 10-msec intervals in the input, but are linearly interpolated at 1-msec intervals for the actual control of the synthesizer. When F0 is positive, B1, B2, B3, and B4 (the bandwidths of the poles) are set equal to 70, 80, 100, and 140 cps, respectively. When F0 is minus zero (which signals noise excitation), B1, B2, B3, and B4 are set equal to 100, 150, 200, and 250 cps, respectively. The bandwidth values are also linearly interpolated at 1-msec intervals.

With F0 positive, a pulse train having a flat spectrum is generated by the excitation source, while noise having a flat spectrum is generated when F0 is minus zero.

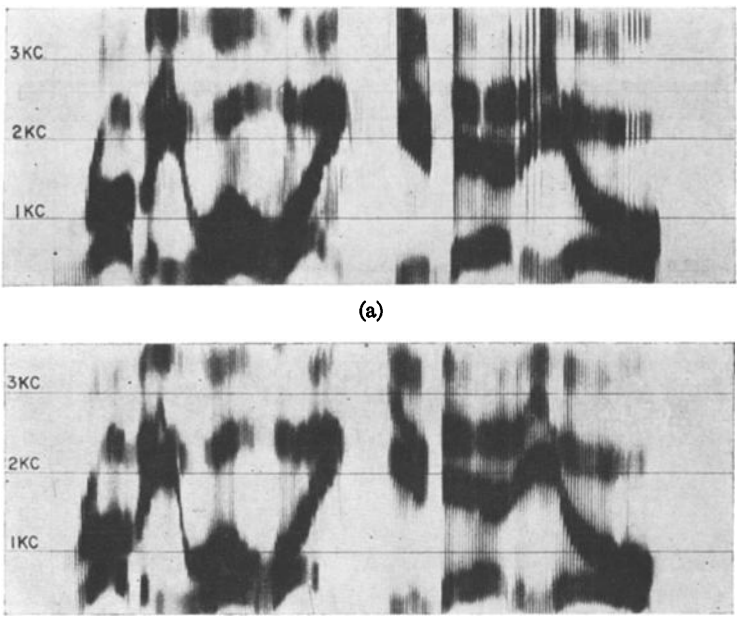

FIG. 8. Spectrograms of "Robby will like you daddy-oh." *Upper:* Natural. *Lower:* Synthetic.

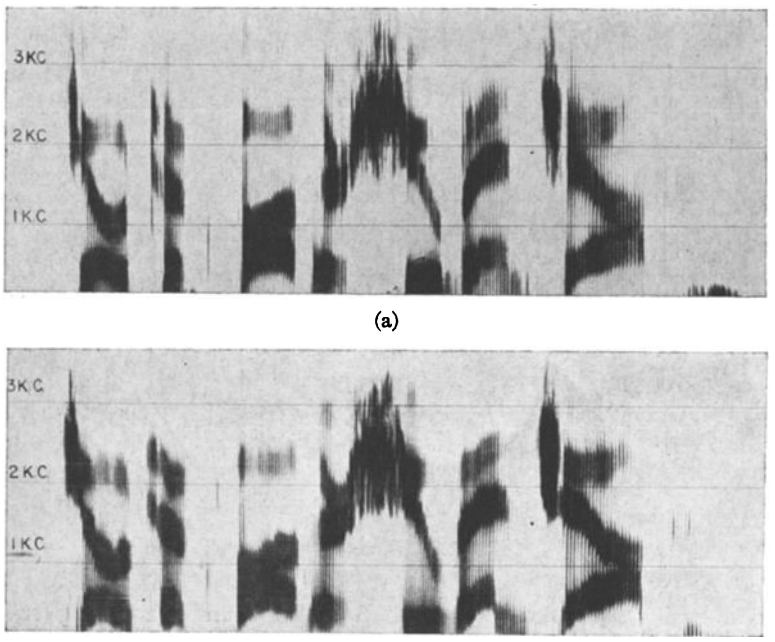

FIG. 9. Spectrograms of "Joe took father's shoe bench out." *Upper:* Natural. *Lower:* Synthetic.

Samples of speech 2 sec in duration are synthesized in the computer and stored on a drum; when the synthesis of a particular sample is complete, it can be listened to repeatedly under sense-switch control of the computer.

Several sentences were processed with the system. Spectrograms of the natural and synthetic versions of the utterance *Robby will like you daddy-oh*, are shown in Fig. 8. Figure 9 shows spectrograms of "Joe took father's shoe bench out." For both utterances, there is substantial correlation between the spectrograms of the natural and synthetic utterances.

II. RESULTS AND DISCUSSION

The vowel list and consonant lists *A* and *B* of Griffiths' modified rhyme test[4] were recorded for one speaker (WKW in the original paper). The speech was processed by the analysis-synthesis system and the resulting synthetic speech was presented binaurally to 10 auditors via Telephonics TDH-39 headphones. A total of 150 words was presented to each auditor (50 words in each of the three lists), and each auditor made one judgment on each word (for a total of 1500 judgments for all auditors). Each time a word was presented, the auditor was required to identify it as one of five words on a list that he had before him. The words were presented at the rate of 1 every 5 sec, with a 10-sec silence after each tenth word.

Of the 50 words in the vowel list, 39 were identified by all 10 auditors as the words intended, seven were identified incorrectly by only one auditor, and four were identified incorrectly by two or more of the 10 auditors. Table I summarizes the auditor responses to the four words for which two or more errors were made. The "Spoken" column in the table lists the sounds that were uttered by the speaker. The "Responded" column lists all the sounds that were believed heard when a particular spoken sound was presented; the number in parentheses is the number of times (out of 10 possible) that the particular sound was believed heard. (When a "1" appears in parentheses, the response is not considered significant.) The "Environment" column shows the phonetic environment in which the intended sound was presented. The "Other Contrasts" column lists the other sounds that were possible responses (there being only five permissible responses for each utterance). There was a total of 21 errors out of a possible 500 (50 words and 10 auditors) for the vowels.

Of the 50 words from Lists A and B that contrast initial consonants, 24 were identified by all 10 auditors as the words spoken, seven were identified incorrectly by one auditor, and 19 were identified incorrectly by two or more auditors. The auditor responses for the latter 19 words are presented in Table II. Line 1 of Table II (where /z/ always goes to /l/) can be regarded

TABLE I. Auditor responses to processed vowels.

Line	Spoken	Responded (out of 10)	Environment	Other contrasts
1	ɔ	ɔ(4), ɑ(6)	n–t	o, u, æ
2	ɑ	ɑ(7), ɔ(3)	n–t	o, u, æ
3	o	o(7), u(3)	n–t	ɑ, ɔ, æ
4	e	e(8), i(1), ɛ(1)	w–n	ʌ, I

[4] J. D. Griffiths, "Further Rhyme-Test Modification for Diagnostic Articulation Testing," J. Acoust. Soc. Am. **40**, 1256 (A) (1966).

as due to computer-operator error, since no noise excitation was specified in the control signals. Lines involving fewer than three errors (Lines 4, 8, 15, 17, 18, and 19) are not discussed further because no trend can be established with two or fewer errors. Also, individual entries involving only one error are considered insignificant and are not taken into account in the discussion. Lines 5, 6, and 9 show confusions between labiodentals and dentals for both the voiced and unvoiced versions, which same confusions also seem to occur for natural speech.[4] Lines 2, 3, 10, and 11 show fricatives going into unvoiced stops (errors in manner and place). Line 7 shows a voiced fricative going to nasal and lateral (manner error). Lines 12, 13, and 14 show voiced stops going into voiced fricatives, nasals, and other voiced stops (manner and place). Line 16 shows errors in place for an unvoiced stop.

In summary then, most of the errors are in manner, with several place errors, and no significant errors in voicing for initial consonants (where Lines 1, 5, 6, and 9 are not considered for the reasons noted above). A total of 99 errors out of a possible 500 occurred for initial consonants. Within the significant errors from lines 2, 3, 7, 10, 11, 12, 13, 14, and 16 of Table II (which are considered the important errors so far as the analysis–synthesis system is concerned), there were 40 errors in manner and 16 in place. Most of the manner errors (32 out of 40) are for fricatives going to something other than fricatives (mostly stops). Most of the place errors (13 out of 16) are for back going to either middle or front.

Of the 50 words from Lists A and B that contrast final consonants, 26 were identified by all 10 auditors as the words spoken, seven were identified incorrectly by one auditor, and 17 were identified incorrectly by two or more auditors. The results for the latter 17 words

TABLE II. Auditor responses to processed initial consonants.

Line	Spoken	Responded (out of 10)	Environment	Other contrasts
1	z	z(0), l(10)	−Ip	š, n, ǰ
2	h	h(1), k(7), t(2)	−Il	w, b
3	h	h(2), t(8)	−Ip	l, d, r
4	h	h(8), k(1), p(1)	−ɑp	š, t
5	v	v(2), ð(8)	−æt	m, f, r
6	v	v(4), ð(6)	−ɑi	θ, h, f
7	ð	ð(2), n(5), l(3)	−i	z, d
8	ð	ð(8), v(1), θ(1)	−ɑi	h, f
9	f	f(4), θ(6)	−In	k, š, t
10	f	f(6), p(4)	−il	h, k, #
11	θ	θ(7), t(3)	−In	ǰ, š, č
12	d	d(4), ð(3), n(2), l(1)	−i	z
13	g	g(6), b(3), m(1)	−el	p, t
14	b	b(6), v(3), n(1)	−ɛst	r, w
15	b	b(8), m(1), p(1)	−ɑrk	l, d
16	p	p(6), t(3), f(1)	−In	w, s
17	k	k(8), p(1), č(1)	−k	s, θ
18	w	w(8), m(1), ð(1)	−e	n, g
19	m	m(8), n(2)	−e	w, g, ð

is the null element.

TABLE III. Auditor responses to processed final consonants.

Line	Spoken	Responded (out of 10)	Environment	Other contrasts
1	θ	θ(1), k(9)	pæ−	s, d, t
2	θ	θ(8), f(2)	hæ−	s, š, v
3	f	f(7), k(3)	pʌ−	b, p, s
4	f	f(8), p(2)	kʌ−	b, d, t
5	v	v(4), ð(5), θ(1)	ši−	f, n
6	v	v(7), d(3)	li−	č, ǰ, š
7	ð	ð(5), k(4), g(1)	wI−	t, č
8	ɲ	ɲ(4), g(6)	lɑ−	ǰ, b, z
9	n	n(5), m(5)	dI−	l, d, g
10	n	n(8), ɲ(1), d(1)	dʌ−	b, g
11	n	n(8), m(1), ɲ(1)	tæ−	b, p
12	p	p(5), k(4), t(1)	sI−	n, ɲ
13	p	p(8), k(2)	pʌ	b, f, s
14	d	d(6), b(2), ɲ(2)	sʌ−	m, n
15	d	d(7), g(2), k(1)	sæ−	t, p
16	d	d(8), n(2)	mæ−	θ, s, t
17	d	d(8), g(1), n(1)	dʌ−	b, ɲ

are shown in Table III. Lines 2, 4, 10, 11, 13, 16, and 17 of Table III are not discussed further because they have fewer than three errors. Line 5 shows a voiced labiodental going to a voiced dental, which confusion also occurs for natural speech.[4] Lines 1 and 3 show unvoiced fricatives going to unvoiced stops (manner and place). Lines 6 and 7 show voiced fricatives going to voiced and unvoiced stops (manner, place, and voicing). Line 8 shows a nasal going to a voiced stop (manner). Line 9 shows a nasal going to another nasal (place). Line 12 shows an unvoiced stop going to another unvoiced stop (place). Lines 14 and 15 show a voiced stop going into other voiced stops (place) and a nasal (place and manner).

In summary, place and manner errors are about equally frequent with an instance of voicing error for final consonants. There was a total of 70 errors out of a possible 500 for final consonants. Within the significant errors from Lines 1, 3, 6, 7, 8, 9, 12, 14, and 15 of Table III (which are considered the important errors so far as the analysis–synthesis system is concerned), there were 34 errors in place and 27 in manner. Most of the place errors (27 out of 34) are for middle going to back or front going to either middle or back. The majority of the manner errors (19 out of 27) are for fricatives going to stops.

Of the 50 significant place errors (for both initial and final consonants), 30 move in one direction ("front to middle or back" and "middle to back") and 20 move in the other direction ("back to middle or front" and "middle to front"). Ten of the group of 20 are for initial /h/ going into something other than /h/. The bulk of the place errors are probably due to second-formant transitions that are incorrect (i.e., that start from either too high or too low a frequency position), although the difficulty is not completely obvious.

The manner errors (which were mainly "fricative to stop" for both initial and final consonants) may be due to amplitude transitions that are too abrupt, incorrect

formant transitions, or they may reflect errors in the voicing decisions, even though there were very few voicing errors as such.

III. CRITIQUE

The analysis–synthesis scheme described herein has an error rate of 4.2% for vowel sounds (21 errors in 500 presentations) and an error rate of 16.9% for consonant sounds (169 errors in 1000 presentations). A subjective evaluation of the synthetic speech, along with the intelligibility results, leads to the conclusion that the synthetic speech could be improved by more attention to details in the analyzer and synthesizer. In other words, the basic form of the analysis–synthesis technique has not been fully exploited.

One comment has been that the synthesizer has too little low-frequency energy in voiced sounds. This might be remedied, in part, by removing the first-formant zero at the origin. In the configuration reported here, all four formants had zeros at the origin.

The response time of the analyzing filters may obscure some of the more transient features of certain consonant sounds, though there is no very obvious correlation between the errors in this experiment and the response time of the filters. However, the response time is slow, as may be noted by comparing the spectrograms for the natural and synthetic consonants /ĵ, t, k, č/ in Fig. 9. If necessary, this response time could be reduced by using wider band filters or by using a time-limited Fourier transform instead of filters.

A more reliable means for making the voiced–unvoiced decision needs to be incorporated. Very few errors in voicing perception were made in the intelligibility test, but voicing errors in the synthesizer control signals may reduce the intelligibility even when they are not apparent as voicing errors to the listener. Note that the synthetic stops /t, k/ in Fig. 9 exhibit voicing errors.

It seems clear, from an examination of many wideband sections, that an additional pole should be included in the synthesizer to accomodate nasals more realistically. A discrepancy between the natural and synthetic versions of /n/ in "bench" can be seen in Fig. 9. The lowest-frequency pole in the synthesizer acts as the first formant of the vowel and then jumps abruptly to accomodate the nasal, whereas in reality, the pole in the vowel should be slowly reduced in amplitude while a pole for the nasal is slowly increased in amplitude.

And finally, although not a part of this experiment, the development of an automatic scheme for extracting formant data is of interest. In principle, one should be able to write algorithms that will extract formant data. These algorithms may involve consideration of such things as spectral peaks, the frequency region within which a particular formant is allowed, and so on. We have taken a step in this direction by partially implementing an automatic method, but no intelligibility tests have been run to date. We feel that additional insights into the problem of automatic extraction can be gained from further refinement and application of the semi-automatic method described herein.

ACKNOWLEDGMENTS

I am indebted to many colleagues for stimulating discussions and for assistance in processing the data, running the intelligibility tests, and preparing the manuscript. Ben Gold and Charles Rader provided some crucial assistance with digital filter techniques, and Cecil Coker pointed out certain advantages of the parallel synthesizer and some improvements for the system.