THE ROLE OF ACOUSTIC PROCESSING IN SPEECH UNDERSTANDING SYSTEMS

A. S. Hoffman

RAND Corporation

Prepared for:

Advanced Research Projects Agency

October 1973

AD-777876

DOCUMENT CONTROL DATA

1. ORIGINATING ACTIVITY The Rand Corporation	2a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED 2b. GROUP

3. REPORT TITLE
THE ROLE OF ACOUSTIC PROCESSING IN SPEECH UNDERSTANDING SYSTEMS

4. AUTHOR(S) (Last name, first name, initial)
Hoffman, A. S.

5. REPORT DATE	6a. TOTAL NO. OF PAGES	6b. NO. OF REFS.
October 1973	38	--

7. CONTRACT OR GRANT NO.	8. ORIGINATOR'S REPORT NO.
DAHC15-73-C-0181	R-1356-ARPA

9a. AVAILABILITY/LIMITATION NOTICES	9b. SPONSORING AGENCY
Approved for Public Release Distribution Unlimited	Defense Advanced Research Projects Agency

10 ABSTRACT

Discusses methods and problems of acoustic signal processing for systems to enable machines to understand spoken communication. Emphasis is on research outside of the ARPA-sponsored SUR (Speech Understanding Research) study. This acoustic level processing includes three steps, not necessarily distinct: (1) preprocessing the original analog signal or its digitized form by basic techniques such as amplitude compression; (2) analysis of the preprocessed signals using fast Fourier transformations, digital filtering, etc; and (3) parameterizing the results in phoneme-sized chunks by formats, autocorrelation techniques, etc. Problems include (1) environmental noise, (2) transducer limitations, (3) determining an appropriate parameterization technique, and (4) coping with wide phonetic, syntactic, and semantic variability of speech. Choice of the voice coding technique depends on the characteristics of the speech understanding system to be used. Many SUR workers are using linear predictive coding and formant tracking. Progress is being made in segmentation and in use of prosodic features. Uncooperative speakers remain a problem.

11. KEY WORDS
SPEECH
ACOUSTICS
SIGNAL PROCESSING

ARPA ORDER NO.: 189-1
3P10 Distributed Information Systems

R-1356-ARPA
October 1973

The Role of Acoustic Processing in Speech Understanding Systems

A. S. Hoffman

A Report prepared for

DEFENSE ADVANCED RESEARCH PROJECTS AGENCY

Rand
SANTA MONICA, CA. 90406

PREFACE

This report is part of a Rand study of "Voice Data Processing Capabilities Applied to Defense Requirements." The project is designed to augment the current speech understanding research (SUR) of other Defense Advanced Research Projects Agency contractors by investigating applications of this research to military systems. Among the various components of the project are:

- Analysis of the nature of speech as a man/computer interface.
- Identification of military man/computer interfaces where the use of speech is operationally attractive.
- Study of the acoustic signal processing aspects of speech understanding systems.
- Study of the natural language and linguistics aspects of speech understanding systems.

This report focuses on the acoustic signal processing component and, more specifically, considers various speech and signal processing techniques that affect the eventual configurations of speech understanding systems for various military applications. The specific applications themselves will be the topic of a subsequent report.

It is assumed that the readers of this report are already familiar with the bases of SUR.

The Information Processing Techniques branch of ARPA should find this report especially useful in its larger study of computer speech understanding.

SUMMARY

The purpose of this report is to explore the role of acoustic level signal processing in speech understanding research (SUR), particularly as practiced outside of the SUR community. The acoustic level processing in speech understanding systems (SUSs) might be considered to consist of three not necessarily distinct steps: (1) the initial preprocessing of the original analog signal or its digitized form using the basic techniques, such as amplitude compression, preemphasis, simple frequency domain filtering, etc.; (2) analysis of preprocessed signals using fast Fourier transformations (FFTs), digital filtering, etc., and (3) parameterizing the analyzed signal in terms of phoneme-sized chunks using such techniques as formants, distinctive features, autocorrelation coefficients, etc.

Problems in acoustic processing include environmental noise (i.e., any type of signal in the amplitude and frequency range of the speaker's voice, including other speakers) and transducer limitations. These are included in the discussion of Sec. V. The problems having to do with the parameterization of the raw acoustic signal arise both (1) in determining an appropriate parameterization technique with a reasonably low data rate and reasonably good representation of the information content of the utterance, and (2) in properly correcting for the wide variability in that representation for the effects of phonetic, syntactic, and semantic contexts.

The voice analysis/synthesis systems discussed in this report are generally low data-rate digital systems with a high coding level. The choice of the appropriate voice coding technique for SUSs or, for that matter, any application, obviously depends on many things. It depends on whether the system is to be all digital or whether a mixture of analog and digital information is to be transmitted and processed. It depends on whether the system is to be single channel or multichannel or whether the processing is parallel or serial, etc. Digital systems are convenient because they lend themselves to digital regeneration, wireline transmission, and forward error control techniques. On the other hand, modems for wireline and forward error control involve additional cost, while analog transmission systems are cost-effective and economical of space, weight, and power. Furthermore, analog systems do not exhibit the thresholding effect peculiar to digital systems, and their reliability is inherently higher because of their basic simplicity of design. Limitations imposed on analog systems include their incompatability with the higher order processing of SUSs and their inability to regenerate.

In the SUS application, the question is not which voice coding technique is most appropriate for the synthesis of high quality speech, but how to extract meaningful

parameters from the acoustic signal. It is well known that the concept of acoustically invariant parameters of speech is not an easy one with which to work, but some sort of parameterization must be implemented even though the higher level processing might make that specific choice a noncritical one. The acoustic analysis technique must also be compatible with a reasonable amount of digital processing. A wide dynamic range and a high sampling rate (with a corresponding minimized amount of coding) might make the determination of acoustic parameters easier, but probably at the cost of an intolerable amount of acoustic processing and, more importantly, with no improvement in the basic SUS task.

The various SUS workers seem to be using linear-predictive-coding (LPC) techniques and formant tracking at the SUS front end. Some of the speech workers outside the community seem to be following this trend (e.g., National Security Agency, Bell Telephone Laboratories) while others seem to be taking other approaches (e.g., Texas Instruments, various Japanese workers).

It would seem that many of the problems at the acoustic front end of the SUS are under control. The transducer, while still a concern, will eventually be fairly flexible (e.g., Texas Instruments' use of the telephone in its recognition scheme). Noise will always cause some problems (e.g., crosstalk on a telephone line), but most of these can be handled with a mixture of sophisticated acoustic processing and further higher level processing. Progress is being made in segmentation, but this will remain a prime source of difficulty. The use of prosodic features is widening in SUSs as witnessed by the advances in that area being made by some of the ARPA contractors. Speaker variability and speaker-to-speaker variations do remain a problem. The processing, memory, and cost constraints of the acoustic processing and, indeed, the SUS, will be more fully addressed in subsequent reports.

CONTENTS

I. INTRODUCTION

The concept of speech analysis used in its broadest sense includes such things as high fidelity speech digitization, speaker verification, speech understanding, speech transcription, etc. This report deals with only two of these components of the broad area of speech analysis. The first might be called acoustic signal analysis and the second, speech recognition.

ACOUSTIC SIGNAL ANALYSIS

Acoustic signal analysis includes bandwidth compression techniques such as vocoders, amplitude limiters, frequency limiters, etc., as well as those techniques that extract various acoustic parameters for further analysis. These are not recognition techniques but only allow for the transmission of sufficient verbal clues to permit a human listener to piece together the linguistic content of the utterance. They depend exclusively on the acoustic properties of speech and make no use whatever of the linguistic properties of the utterance. Speech bandwidth compression is of interest in this study for two reasons—speech bandwidth compression systems form the basis of the speech understanding system's (SUS's) front end and much of the early work in acoustic signal processing was concerned with speech bandwidth compression systems of one sort or another.

SPEECH RECOGNITION SYSTEMS

The essential goal of computer-based speech recognition systems, and more particularly SUS, is that of avoiding the necessity of involving a perceptive human in the loop. This often requires a computer analysis of both the acoustic content as well as some of the higher level linguistic information of the utterance.

Through the mid-fifties, the efforts to build speech recognition machines dealt almost exclusively with the acoustic properties of input signals and corresponding signal analyses. Use was made of formant tracking and spectral pattern comparison as well as rudimentary binary decision methods. From the mid-fifties to the mid-sixties, attempts were also made to tackle the segmentation problem. Subsequently,

the linguistic aspect of the problem received more attention. Research since the mid-sixties has emphasized both the incorporation of linguistic information into the speech recognition matrix and the continued concern with acoustic signal processing. There has been a continued interest in the articulatory and the perception processes as well as an interest in the acoustic signal processing techniques.

The purpose of this report is, basically, to explore the role of acoustic level signal processing in speech understanding research (SUR), particularly as practiced outside of the SUR community. Section II discusses in general terms the reasons behind the interest in speech signal processing. Section III considers various approaches to the parameterization of the speech analysis/synthesis process in some detail with a strong leaning toward the highly coded systems. Section IV considers speech synthesis approaches and speaking machines, but only insofar as they affect the speech understanding problem. The coordination of speech signal processing with higher level processing in SUSs is the subject of Sec. V.

II. WHY THE INTEREST IN SPEECH SIGNAL PROCESSING?

There are many reasons why people are interested in speech signal processing. The physiologist and the psychologist are interested in the mechanisms of the vocal and the auditory systems, the biological signal processing in speech formulation, and auditory perception. The clinical physician is interested in the pathological clues that speech signal analysis promises to provide. The linguist needs to understand the building blocks of human communication. The communication engineer wants to process and transmit speech efficiently and inexpensively. Additionally, the intelligence engineer is concerned with the security of speech communications. The military engineer wants to use speech where it can do a given job better or allow a new kind of job to be done. Finally, the commercial engineer would just like to use speech where it makes economic sense. Thus, speech signal processing is multidisciplined in outlook and multi-goal oriented in application. ARPA's interest in speech understanding that is oriented toward the accomplishment of a task reflects the multi-disciplined approach.

Identifying the properties of the acoustic signal and processing it are necessarily a part of any SUS. A large part of the knowledge in acoustic signal processing is rooted in the analysis/synthesis schemes developed for speech communication systems. Section III considers this analysis/synthesis area in some detail and tries to relate it to acoustic parameterization in the SUS. Before looking at some of the specific methods of speech analysis/synthesis, it would be instructive for the sake of perspective to outline the hierarchy of coding schemes for speech. The lowest level coding is listed first with progression through various digital schemes [1] to the most highly coded technique. The progression of coding sophistication is as follows:

- Band-limited speech—sampling schemes such as pulse amplitude modulation (PAM).
- Dynamic-range-limited speech—amplitude quantization such as pulse code modulation (PCM) and delta modulation.
- Speech where quantization noise depends on step size—logarithmic signal compression such as nonlinear PCM.
- Speech in which pauses in the speech waveform are not transmitted.
- Speech where use is made of the fact that the information parameters of the speech signal vary much more slowly than the signal itself—predictive coding at various levels from adaptive delta modulation to more sophisticated predictive coding.

Through the techniques listed so far, the required transmission channel capacity can be reduced by 75 to 90 percent from an uncoded signal. In practice, narrow-band voice communication has been constrained by a maximum data rate of 2.4 kbit/s because of the capability of available modems. The following low bit-rate techniques are relatively complex in implementation, and they generally provide only marginal voice quality:

- Description of the signal by a set of orthogonal functions (such as Walsh functions or Laguerre polynomials). These are used in various kinds of vocoders.
- Normalization and classification of the vocoder signals—these are pattern-matching or formant vocoders.
- Derivation of a phonetic transcription deleting a speaker description.

In an actual speech communication system, the choice of coding technique obviously depends on the specific application, the quality of speech desired, the transmission requirements, the user preference, and the hardware requirements. In an SUS, the acoustic processing characteristics of interest include, for example, the background noise, bandwidth, and dynamic range. The segmentation method and prosodic feature considerations are also of considerable importance.

The acoustic level processing in an SUS might be considered to consist of three not necessarily distinct steps: (1) the initial preprocessing of the original analog signal or its digitized form using basic techniques such as amplitude compression, pre-emphasis, simple frequency domain filtering, etc.; (2) analysis of preprocessed signals using FFTs, digital filtering, etc., and (3) parameterizing the analyzed signal in terms of phoneme-sized chunks using such techniques as measuring formants, distinctive features, autocorrelation coefficients, etc.

Problems in acoustic processing include environmental noise (i.e., any type of signal in the amplitude and frequency range of the speaker's voice including other speakers) and transducer limitations. These are included in the discussion of Sec. V. The problems having to do with the parameterization of the raw acoustic signal arise both (1) in determining an appropriate parameterization technique with a reasonably low data rate and reasonably good representation of the informational content of the utterance, and (2) in properly correcting for the wide variability in that representation for the effects of phonetic, syntactic, and semantic contexts. This brings us directly to Sec. III, which, in part, discusses some of the parameterization techniques.

III. SPEECH ANALYSIS/SYNTHESIS APPROACHES

The parameterization process is an information-reducing transformation of the raw speech signal. The objectives of this transformation are, first, to reduce the information rate of the input signal and, secondly, to facilitate the recognition of phonetic features and to decode the phonological rules of speech production. A number of parameterizations are currently in use by the ARPA contractors for SUS purposes and others for speech recognition purposes. Still other parameterizations that have been developed in part for vocoder use are not being pursued in speech understanding or speech recognition work, but conceivably might be. In this section, consideration will be given to some of the parameterization techniques. The problems associated with the parametric variability from segmental, syntactic, and semantic context; prosodic features; and speaker characteristics are not covered in this section, but are discussed in Sec. V.

Some of the coding techniques used in the acoustic analysis and synthesis of the speech waveform rely on a very basic model of speech production that considers the vocal tract as a variable-shaped tube terminated at one end by the vocal chords and at the other end by the nose and mouth. The driving force for this system is a sustained air pressure wave generated by the lungs. Voiced sounds are produced by exciting the vocal tract with quasi-periodic pulses of air pressure generated by the lungs and modified by the vibrations of the vocal chords. Schematically (see Fig. 1), the production of voiced sounds is represented by an impulse-train generator at the pitch period. The vocal tract, which changes shape throughout as well as constricting severely at various points, acts as a time-varying filter for the vocal tract excitation function. For the production of unvoiced sounds, the vocal tract filter is excited by a uniform random number generator instead of the impulse-train generator. A number of the coding techniques described below make use of this simple model as representative of the speech process by simply estimating the parameters of the model from the speech waveform.

CHANNEL VOCODER

One of the oldest methods* for speech analysis/synthesis uses the short-time spectrum of the acoustic signal, recognizing that speech perception is, to a large

* Although the conventional channel vocoder acoustic analysis scheme is not used in the front end of an SUS, the short-time spectrum is important in acoustic analysis. The shortcomings and various

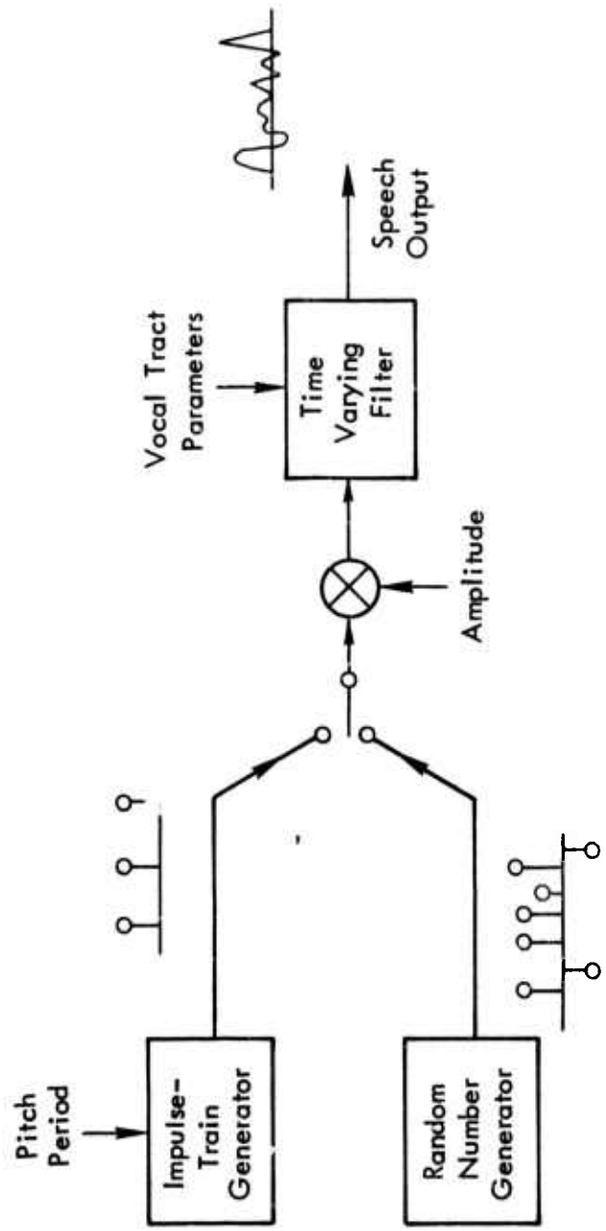

Fig. 1—Model for speech production (adapted from Ref. 2)

degree, dependent on preservation of the shape of the short-time amplitude spectrum. The short-time spectrum is a particular parameterization of the time-varying filter representation of the vocal tract. The spectral envelope at any given time is typically represented by 5 to 20 samples along the frequency axis. The excitation information is contained in parameters that measure the absence or presence of voicing and, if present, its pitch. A block diagram of the basic channel vocoder configuration, both analyzer and synthesizer, is shown in Fig. 2. The channel vocoder thus describes speech by two parameters. The first, called excitation, characterizes the excitation of the vocal tract as being voiced or unvoiced and, if voiced, measures its pitch. The second parameter is concerned with the characterization, in particular, with specification of the short-time amplitude spectrum, of the structure of the vocal tract. The channel vocoder is a fairly efficient data-compression device. For example, at a sampling rate of 40 samples per second with 3 bits per sample on each of 14 spectrum channels and 6 bits per sample for the pitch signal, the analyzer output corresponds to a bit rate of $40(14 \cdot 3 + 6) = 1920$ bit/s. Transmitting a digitized version of the original speech signal at, say, 7 bits per sample with 7000 samples per second corresponds to a 49,000-bit/s rate—a 25 to 1 reduction. The voice quality of such a channel vocoder, however, is poor with the reconstituted speech sounding unnatural.

Over the past 30 years, the channel vocoder has been repeatedly modified: The number of filters, their spacing, their selectivity, their overlap, their bandwidths, etc., have all been varied, and various techniques for the voiced-unvoiced detection and pitch extraction have been tried. In the more successful versions of the channel vocoder, speech intelligibility is fairly high, but there is invariably a degradation of speech naturalness and quality. Several factors seem to be responsible for this. One is that voiced-unvoiced discriminations are not always made accurately. This is partly because the actual concept of voicing is not as simple as the model makes it appear, partly because the electronic measure of voicing does not always reflect physiological voicing, and partly because the synthesizing pulse source waveform and phase do not reflect the details of the physiological voicing. Another factor behind the channel vocoder's inadequacy is the measurement of pitch: The temporal variation of pitch may not be followed or tracked accurately enough, or the pitch may often be measured incorrectly. Other sources of error include inadequate resolution of the spectral analysis and loss of phase information in determining the spectral power density. Finally, the amplitude dynamic range of the acoustic signal itself is often quite large so that practical rectifiers and amplifiers might provide inadequate coverage.

Many of these shortcomings have been overcome to some extent in practical channel vocoders. The excitation problem, however, is inherent in the very nature of the channel vocoder so that attempts to overcome that limitation have been via other types of vocoder techniques (e.g., cepstral techniques, voice-excitation, etc.). The problems of the time-varying filter parameterizations of the channel vocoder configuration have been attacked via normalization of channel signals, the dynamic range problem via higher resolution in particularly sensitive spectral regions, etc.

The channel vocoder is characterized by reasonably high intelligibility, but displays poor speaker recognition and relatively poor quality. The low quality and

---

improvements do have enough direct applicability to SUSs to warrant a fairly complete discussion of the channel vocoder and a few of its successors.

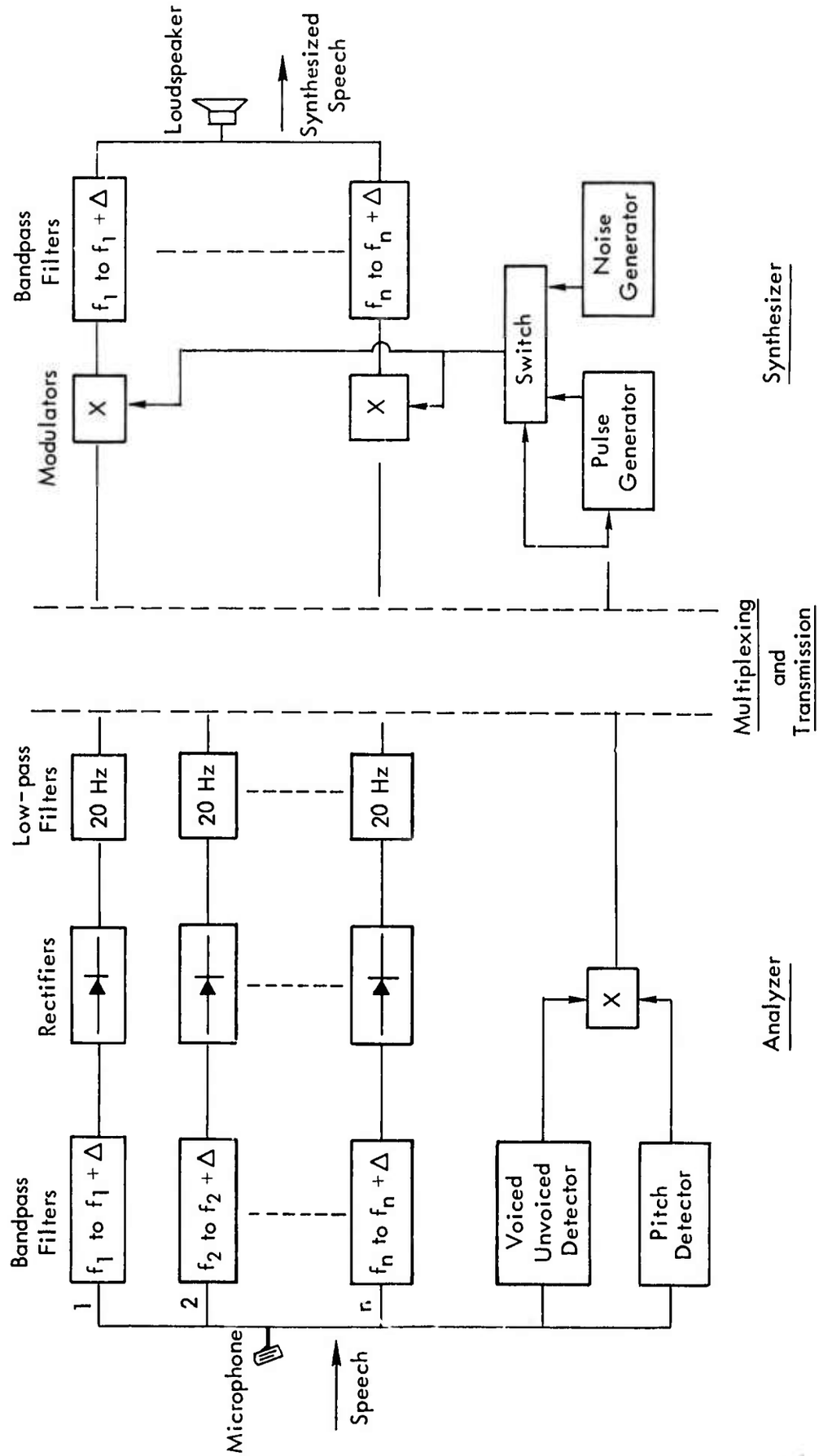

Fig. 2—Basic channel vocoder configuration (adapted from Ref. 3)

poor speaker recognition of the channel vocoder have limited its utility. At 2400 bit/s, the channel vocoder can interface with a large number of readily available wireline modems and can be operated over conventional telephone channels. Current versions of the digitized channel vocoder occupy about 700 $in.^3$, weigh approximately 30 lb, and draw 25 to 30 W.

Improvements of the conventional channel vocoder have been accomplished by all-digital versions. One from Lincoln Laboratory [4] uses the discrete Fourier transform technique and incorporates an improved pitch extractor. The improvement in this version's quality can be attributed directly to the use of an improved pitch extraction technique that is a modification of the pitch extraction approach developed by Gold of Lincoln Laboratory [5]. The key to high quality, low data-rate vocoders seems to be the accuracy with which pitch can be extracted. Accurate pitch extraction, however, can result in a costly implementation. The Lincoln Laboratory's modified pitch extractor is economic and also performs as well for female speakers or for male pitch frequencies exceeding the normal limits (as might occur under a high stress condition) as for males with normal pitch frequencies. To achieve a data output rate of 2400 bit/s (for modem compatibility), some data compression is needed if, say, 32 vocoder channels, each of which is sampled 50 times a second, are to be transmitted within the 2400-bit/s design data rate. The reduction in data rate is brought about, in part, by making use of the fact that the spectral measurements between adjacent filters are highly correlated and, in part, by the fact that at the higher frequencies less discrimination is required by the listener. Furthermore, the ear responds approximately logarithmically to sound intensity. Based on these characteristics, grouping the higher frequency channels and using the Hadamard matrix transformation of the lower 16 frequency channels permits the required data reduction. The net result of these encoding techniques permits the spectrum to be described by a 32-bit word that, when added to an 8-bit description for pitch and an 8-bit description for other excitation parameters, forms a 48-bit word transmitted every 20 ms producing a 2400-bit/s data rate.

VOICE-EXCITED VOCODER

The voice-excited vocoder [6] offers a basic improvement over the channel vocoder. Like the channel vocoder, the voice-excited vocoder performs a spectrum analysis of the speech by means of a contiguous filter bank. Unlike the channel vocoder, the voice-excited vocoder does not extract the pitch or a waveform that is proportional to pitch. Rather, in a representative trial device, a selected baseband between 100 and 700 Hz is A/D converted in the analyzer with a bit description equivalent to 4-bit PCM and transmitted to the synthesizer along with the spectrum channel information and the voiced-unvoiced decision. The reconstructed voice baseband is used in the synthesizer to excite the synthesis filter bank. As a result of the voice excitation requirement, data rates in the range of 7.2 to 9.6 kbit/s are necessary to transmit the digitized voice, and as a result, the voice-excited vocoder requires wireline modems operating at rates up to 9.6 kbit/s, which are expensive and require conditioned data lines.

Bell Telephone Laboratories [7] successfully demonstrated the feasibility of a digitized voice-excited vocoder that operates at 4.8 kbit/s. This relatively low data

rate was obtained through processing of the excitation band by one of two means—bandpass sampling or harmonic compression. In the first approach, a straightforward bandpass sampling technique minimized the number of bits required to describe the excitation band—a band from 187 to 562 Hz was used. This bandpass sampling approach resulted in an overall 3000-bit/s description for the excitation band and 1800 bit/s for the spectral channels. Bandpass sampling allows for minimization of the number of samples per second required to describe a bandpass process and eliminates the need for translating prior to sampling. In the second approach, harmonic compression was used to first reduce the bandwidth of the voice excitation prior to digital encoding. This technique also resulted in a 3000-bit/s description of the voice excitation band, but it required a multitude of narrow-band filters followed by harmonic compression to bring about the bandwidth reduction prior to sampling. For this reason, the harmonic compression technique was prohibitively complex, and the bandpass sampling approach was thought to show more promise. Using either bandpass sampling or harmonic compression, vocoder performance at 4.8 kbit/s was of a quality rivaling that of a voice-excited vocoder operating at 7.2 kbit/s. The 4.8-kHz data rate of the voice-excited vocoder is compatible with conventional wireline modems for conditioned lines, although the modem costs are relatively expensive.

PHASE VOCODER

Bell Telphone Laboratories [8] has developed another type of vocoder that is capable of high quality speech synthesis. This vocoder is the phase vocoder. Unlike the voice excited vocoder, which relies on channel vocoder technology coupled with voice excitation to effect a high quality voice system, the phase vocoder utilizes amplitude and phase derivative information from the short-time spectrum to synthesize speech. The system offers a basic bandwidth reduction of approximately two to one with the analysis portion of the vocoder being relatively complex compared to the synthesis portion. If the number of analysis channels is large, the phase signals contain mostly information about the excitation. The amplitude signals, on the other hand, depend on the vocal tract transmission properties and the source spectrum.

In addition to its improved transmission efficiency, the phase vocoder offers some flexibility in manipulating the basic speech parameters. The frequency range of the signal can be expanded or compressed without affecting the time scale. The time scale as well can be compressed or expanded without affecting the frequency range.

Carlson has estimated [9] that an all-digital phase vocoder can be implemented with a data rate in the range 7.2 to 9.6 kbit/s. The major simplifications suggested to lower the high cost of the phase vocoder were the use of nonuniform filters in a reduced configuration size (30 to 8 filters).

With the completion of the discussion of the conventional channel vocoder and two of its derivative techniques, the discussion can turn to vocoder and parameterization techniques that are currently more in vogue for SUS and speech recognition front ends.

FORMANT TRACKING VOCODERS

Adjacent values of the short-time amplitude spectrum are fairly well correlated. A coding of the vocal tract response in terms of the complex poles and zeroes of the vocal mode pattern is a fairly efficient specification of the amplitude spectrum. The spectral envelopes of many speech sounds are characterized by several prominent maxima that represent the resonances of the vocal tract. There are often three so-called "formants" below 3000 Hz in much of adult speech. The practicality of formant vocoders depends on how well these formants can be derived, how well their temporal variation can be followed or tracked, and how well the associated tuned resonant synthesizer circuits collectively reproduce the vocal mode pattern. In addition, just as in the channel vocoder, excitation information must also be provided. A block diagram of the basic formant tracking vocoder is shown in Fig. 3. The determination of the formant frequencies, their relative strengths (F_j and A_j, respectively, in Fig. 3), and their temporal variation has been a major problem in the development of formant vocoders. The most direct measurement procedure, bandpass filtering the spectral signal for each of the formants, has not worked well because of the considerable spectral overlap of the various formants and because of their often rapid changes.

An "analysis by synthesis" approach [10] has shown limited success. The speech signal spectrum is matched by iteration techniques to artificially generated spectra. The formant frequencies and bandwidths of the spectrum generator are taken as a representation of the speech signal. Other techniques for formant tracking have also shown limited success (e.g., progressively subtracting formant envelopes [11] and enhancing the formant resolution by narrow-band analysis along a contour passing close to the poles [12]). Of recent interest to both formant vocoders and to parameterization for the SUS is the use of linear predictive coding for formant tracking. This will be discussed later in this section.

Several formant vocoders have been implemented. Representative of these, although not typical of all the others, is the one built by Philco-Ford Corporation [6]. While the intent of the design of this vocoder is to provide an analog description of the voice signal compatible with 3-kHz bandwidth transmission, it is readily extendable to digital implementation. This particular vocoder extracts the formant amplitude and frequency information associated with the formants F_2 and F_3 and transmits the raw voice signal occupying a bandwidth between 300 and 700 Hz. This is unlike most formant vocoders that extract both amplitude and frequency information of the first three formants. It detects pitch and makes a voiced-unvoiced decision in the usual manner. In the synthesizer, formants F_2 and F_3 are excited using the pitch information obtained in the analyzer. However, the raw baseband signal occupying a spectrum from 300 to 700 Hz is allowed to pass through to the output of the synthesizer where it is finally mixed with the pitch excited formant outputs F_2 and F_3. This vocoder does not use the voice excitation band to excite formants F_2 and F_3 for it was found through experimentation that improved results could be obtained by standard pitch excitation of the upper two formants. The required transmission rate is about 4.8 kbit/s. In terms of size, weight, power, and complexity, as well as quality, formant vocoders are somewhat better than the channel vocoder and its digital equivalent.

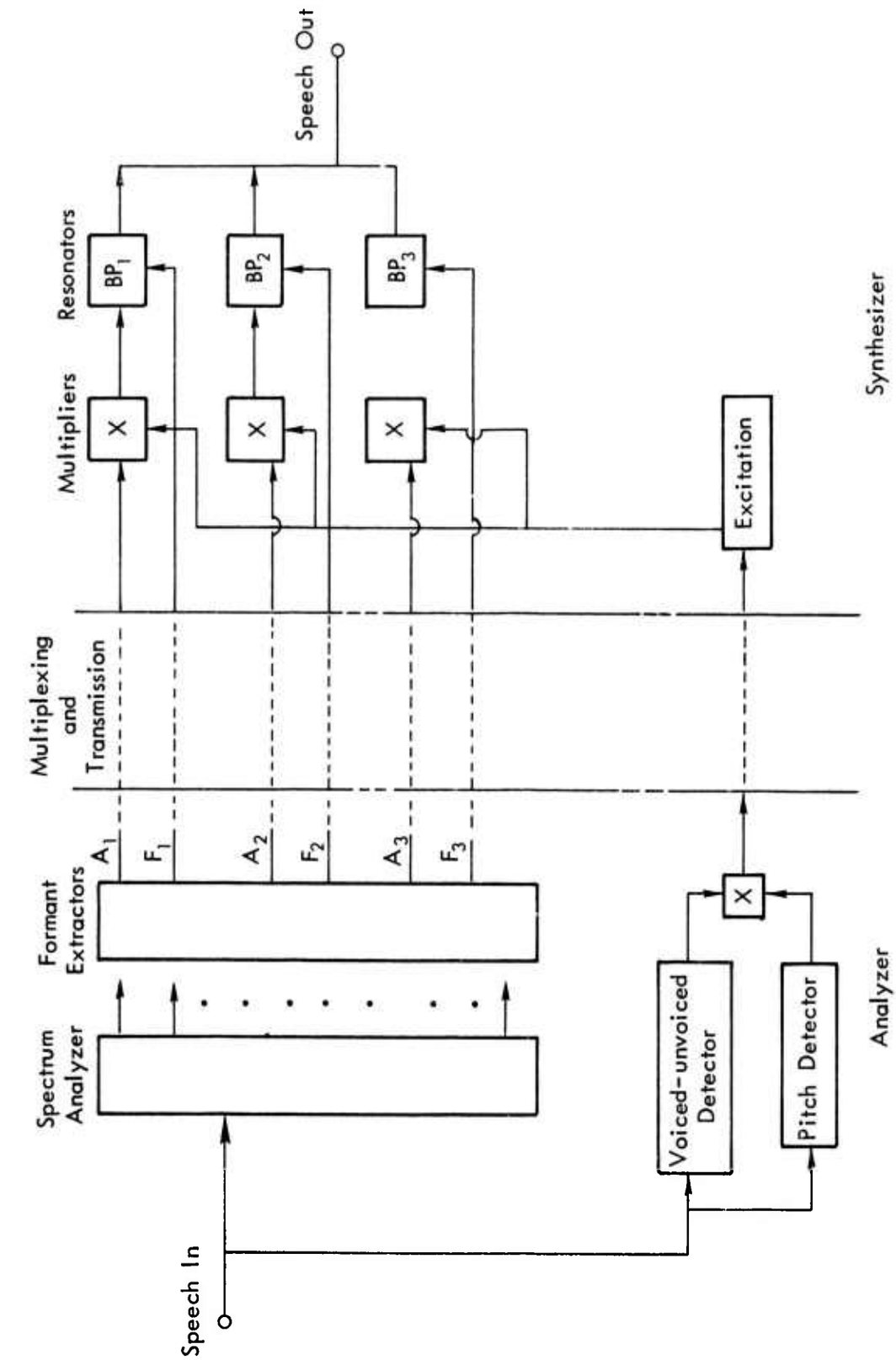

Fig. 3—Formant tracking vocoder configuration

CEPSTRUM VOCODER

The cepstrum technique is based on a particular deconvolution of the speech waveform that separates the speech excitation function from the vocal-tract impulse responses [13,14]. In the following paragraphs, reference is made to Fig. 4, which shows a block diagram of a general cepstrum vocoder configuration, and to Fig. 5, which shows appropriate waveforms for both voiced and unvoiced speech samples. In the cepstrum approach, the Fourier transform of a segment of speech is computed first. Next, the logarithm of the magnitude of the Fourier transform is computed. From this logarithm vs. frequency plot, two components are visible—a rapidly varying periodic component associated with the vocal tract excitation and a slowly varying component associated with the vocal tract transmission. To separate these components, a standard technique is employed that filters the log spectrum function by taking its Fourier transform. This Fourier transform of the log of the spectrum (which is now back in the time domain) is called the cepstrum. In the cepstrum plot (as a function of time) the rapidly varying periodic portion of the log spectrum shows up as a peak at the high end of the plot. The slowly varying component of the log spectrum shows up at the lower end of the time axis. To separate the excitation function from the vocal tract transmission function, the following procedure is followed. A simple truncation of the cepstrum (low-pass filtering), so as not to include the cepstral peak, allows one to obtain the log spectrum of the vocal tract function (unpolluted by the excitation function) by simply taking the inverse transform of the truncated cepstrum. If this sounds complex, it is; but a study of Figs. 4 and 5 should serve to clarify it.

One of the advantages to this approach is that voicing and pitch information is easily and accurately obtained from the cepstrum plot. A disadvantage is the large amount of processing—two Fourier transforms and a logarithmic calculation are required for each speech sample. Voicing is present if there is a cepstral peak at high values on the time axis. The corresponding pitch frequency can then be obtained from the location of the peak in the cepstral plot. As a bonus, the pitch fundamental will show up in the cepstral plot even if only the harmonics and not the fundamental are visible in the original spectrum plot. Moreover, a good estimate of the formants are obtained by determining the peaks in the smoothed spectrum; these are the vocal tract resonances.

A digital speech analyzer/synthesizer that uses such a cepstrum deconvolution to bring about effective speech coding has been described [15]. It was estimated that this system can operate in the region 7.2 to 9.6 kbit/s. The computer-simulated voice synthesized from this technique is of high quality. Estimates of the hardware required for an operational system based on that particular cepstrum and deconvolution approach appears to be rather costly. High-speed digital computation would be required to bring about a real-time system with the costs of the real-time digital processor being quite high compared to other voice coding techniques.

LTV Electronics Systems [6] has developed an all-digital time domain vocoder with quality speech production capabilities that operates at 2400 bit/s. Cepstrum analysis is used in this analyzer to derive both the power-spectrum information and the pitch information required in the synthesizer. Measurements indicate that the speech quality is more natural than the conventional channel vocoder.

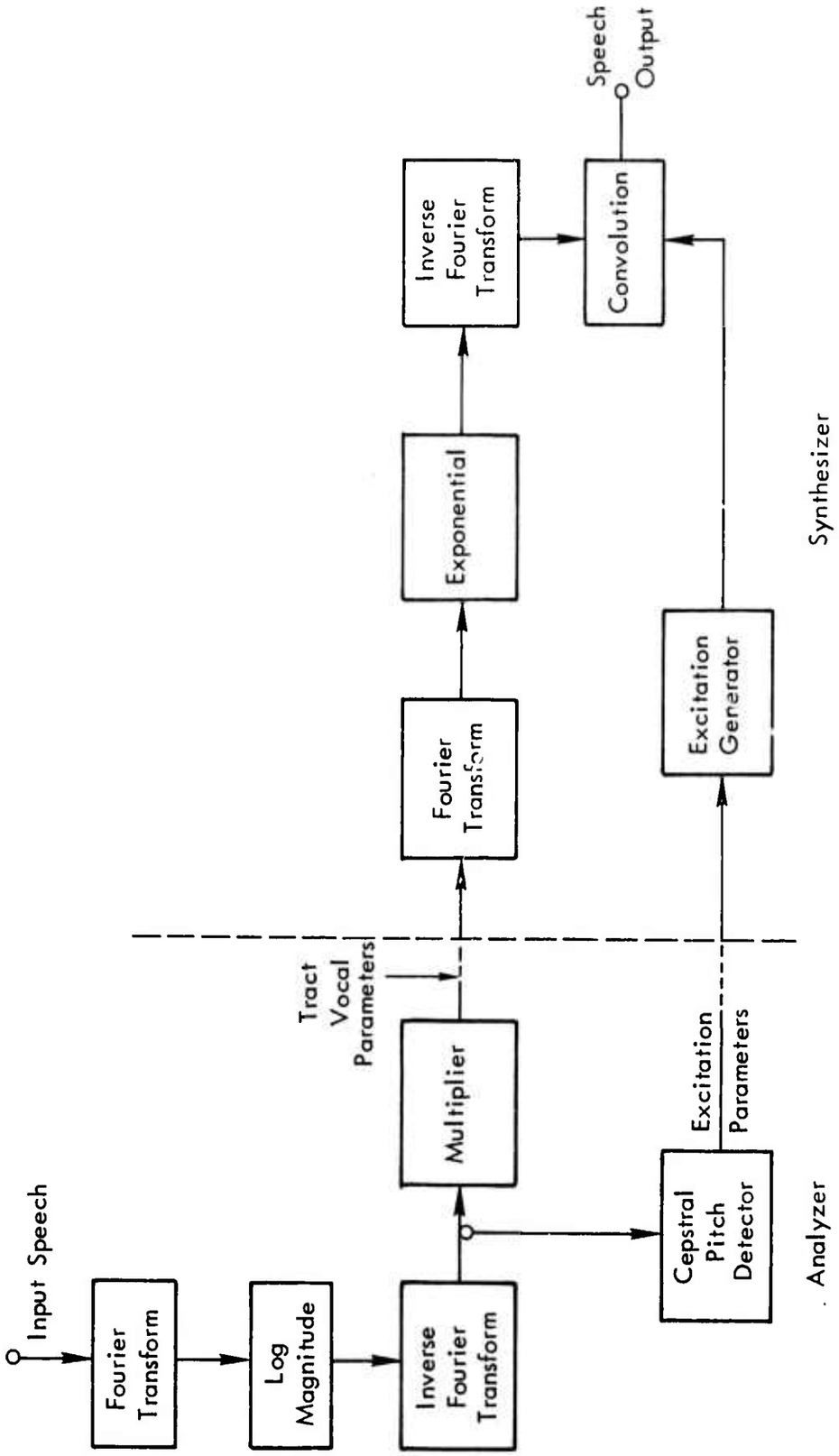

Fig. 4—Cepstrum vocoder general configuration

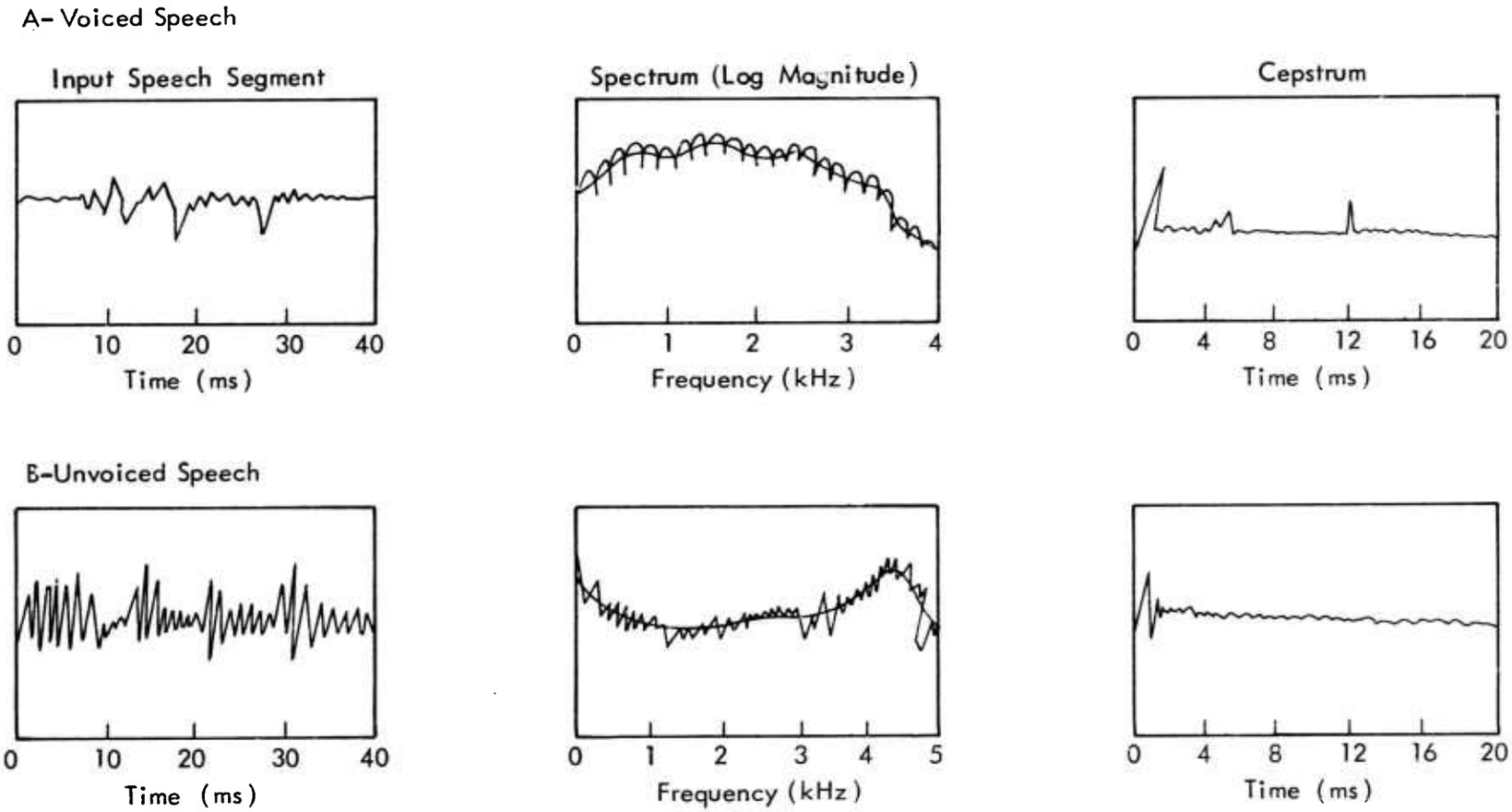

Fig. 5—Cepstrum vocoder speech waveforms

LINEAR PREDICTIVE CODING (LPC)

The LPC technique as applied to speech processing has received wide attention recently. This technique can reduce the redundancy in the speech temporal waveform by subtracting from the speech waveform the portion that can be predicted from its past. This technique differs from the vocoders and vocoder-like techniques discussed so far in that it does not rely on the rigid parameterization of the speech signal dictated by a physiologically based model such as that of Fig. 1.

Instead, it uses a functional model of the speech production mechanism based on the linear-prediction representation of the speech wave [16,17] as shown in Fig. 6. A single recursive (all-pole) filter represents the combined contributions of glottal flow, the time-varying vocal tract structure, and the oral radiation. The problem of separating the source function from the vocal tract function is entirely avoided. The vocal chord excitation for voiced sounds comes from an adjustable amplitude and period pulse generator, while the unvoiced excitation comes from a white noise source. The linear predictor, P, is a transversal filter with p delays of one sample interval each. It forms a weighted sum of the past p samples at the input to the predictor. The output of the linear filter at the n^{th} sampling instant is given [16] by

$$s_n = \sum_{k=1}^{p} a_k s_{n-k} + \delta_n \quad ,$$

where the a_k are the predictor coefficients accounting for the combined filtering action of the glottal flow, the vocal tract, and the radiation. The δ_n represents the n^{th} sample of the excitation. The number, p, of coefficients necessary to specify a given speech segment is determined by the glottal volume flow function, the resonances and anti-resonances of the vocal tract in the frequency range of interest, the radiation function, and the sampling frequency. For example, at a sampling frequency of 10 kHz, $p = 12$ seems to be adequate [16] in most cases. During the time that the vocal tract shape is constant, a complete representation of the speech waveform is provided by the predictor coefficients, a_k, the pitch period, the root-mean-square value of the speech samples, and a voiced-unvoiced indicator. Typically, these parameters must be readjusted every 5 to 10 ms. The synthesis procedure is just the inverse of the analysis, using a single recursive filter without requiring information about individual formants. Listener tests show little perceptible degradation [16,17] in the quality of the synthesized speech.

It is possible to reduce the data rate [16] to about 2400 bit/s without producing significant degradation in speech quality. Other speech characteristics and parameterizations are easily obtained from this encoding technique with little additional computation, e.g., formant frequencies and bandwidths, the spectral envelope, and the autocorrelation function.

There has been work in applying the concepts of LPC to more conventional encoding techniques [17]. For example, differential PCM and adaptive delta modulation have been extended to lower data rates (e.g., 9600 bit/s) by using linear predictive techniques. Predictive coding of the channel signals of a cepstral vocoder has produced bit-rate reductions from 7800 to 4000 bit/s with little speech degradation [18]. Formant tracking using an LPC algorithm has been accomplished by a number of researchers, particularly in the SUR community [19]. A brief description of how particular ARPA SUR contractors are using this technique in their work follows.

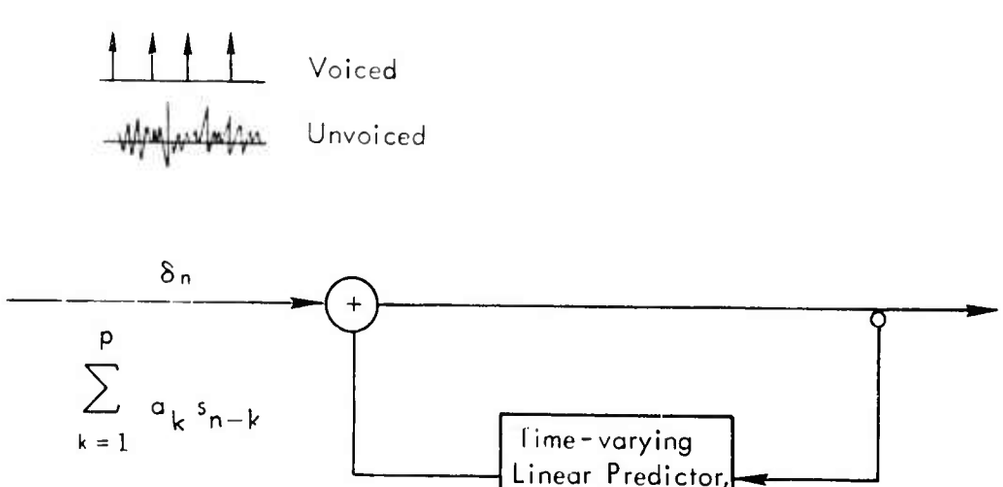

Fig. 6—Functional model of speech production based on the linear prediction representation (adapted from Ref. 17)

In summarizing some of the early LPC developments, a very early predictive coder (Bell Telephone Laboratories) rendered a voice signal at 9.6 kbit/s with the equivalent description of 6-bit logarithmic PCM. Its performance was extremely good providing highly intelligible speech. A 2.4-kbit/s predictive coder/decoder has been simulated on a computer with excellent results. A modification of the Bell predictive coding system has been breadboarded by ITT [6]. It had a 9.6-kbit/s data signal with performance equivalent to 4-bit PCM. The current state-of-the-art of LPC modems seems to be about 3600 bit/s.

SPEECH ANALYSIS/SYNTHESIS SUMMARY

The voice analysis/synthesis systems discussed above are low data-rate digital systems with a more sophisticated coding level. The choice of the appropriate voice coding technique for SUS or, for that matter, any application, obviously depends on many things. It depends on whether the system is to be all digital or whether a mixture of analog and digital information is to be transmitted and processed. It depends on whether the system is single channel or multi-channel, or whether the processing is parallel or serial, etc. Digital systems are convenient because they lend themselves to digital regeneration, wireline transmission, and forward error control techniques. On the other hand, modems for wireline and forward error control involve additional cost and analog transmission systems and are very economical and require less space, weight, and power. Furthermore, analog systems do not exhibit the thresholding effect peculiar to digital systems, and their reliability is inherently higher because of their basic simplicity of design. Limitations imposed

on analog systems include their incompatibility with the higher order processing of SUS and their inability to be regenerated.

In the SUS application, the question is not which voice encoding technique is most appropriate for the synthesis of high quality speech, but more a question of extracting meaningful parameters from the acoustic signal. It is well known that the concept of acoustically invariant parameters of speech is not an easy one with which to work, but some sort of parameterization must be implemented even though the higher level processing might make that specific choice a noncritical one. The acoustic analysis technique must also be compatible with a reasonable amount of digital processing. A wide dynamic range and a high sampling rate (with a corresponding minimized amount of coding) might make the determination of acoustic parameters easier, but probably at the cost of an intolerable amount of acoustic processing and, more importantly, with no improvement in the basic SUS task. At this point, it might be profitable to describe briefly the approaches to the acoustic processing that various researchers have followed. This summary is not meant to be inclusive or complete, but only to indicate current trends. Details of the individual approaches can be found in the appropriate references.

The Carnegie-Mellon University system's acoustic processing [20] takes the speech input through a 5 octave bandpass filter (spanning the range of 200 to 6400 Hz) and an unfiltered band. During each 10-ms interval, a measurement is made of the maximum intensity and the number of zero crossings within each band. The resulting 12 parameters every 10 ms are smoothed and log-transformed, and a subset of the parameters is used for further processing. Each 10-ms unit is classified by comparison with a standard set of parameter vectors. The standard set of vectors is obtained by selecting cluster centers from a training set of utterances containing various phonemes in a neutral context. The continuous parameter sequence is segmented into discrete phoneme-size chunks based on an acoustic similarity measure. The acoustic recognizer has available three sources for the generation and verification of hypotheses: (1) acoustic knowledge in the form of expected parameters for a phoneme in a neutral context, (2) a coarticulation model that modifies the expected features based on context, and (3) a vocabulary restriction in the form of a valid subset of words in the lexicon that contain a given sequence of features. The acoustic hypothesizer retrieves those words of the lexicon that are consistent with the gross features present. The acoustic verifier then determines whether a given hypothesis is consistent with the context presently available to it.

The Lincoln Laboratory approach [19] to acoustic analysis in their SUS uses as input an LPC spectogram, the LPC coefficients, and a segmentation indicator. A formant tracking algorithm yields the first three formant positions and amplitudes. The algorithm relies heavily on the output of previous processed frames to determine the output from the present frame so that the process is stared at a point in time where the formants are most likely to be correct. A linear predictive spectogram is computed at 5-ms intervals via an inverse filter method. An estimate of the fundamental frequency as well as a voiced-unvoiced decision is also computed at 5-ms intervals. The segmentation algorithm and determination of segment class relies on spectral measurements performed on the linear predictive spectra.

Stanford Research Institute's acoustic processing is based [21] on LPC for formant tracking. The LPC analysis of the raw time-series data provides spectral peaks that are stored as three formant frequencies and amplitudes. Each 10-ms time

segment is given a preliminary classification into one of six categories. Procedures for word verification relate the words predicted by syntactic and semantic processing to the acoustic data. Current efforts are aimed at increasing the sophistication in the acoustic processing by adding more subroutines for acoustic parameterization in order to refine the initial classification and to provide additional formant data.

In the Systems Development Corporation (SDC) approach [22], the digitized speech signal, after optional preemphasis, goes through three bandpass filters from which 6 parameters are extracted: amplitude and zero-crossing counts for each of the three filters. These six parameters, updated every 10 ms, form a unit. Linear predictive coding is used to obtain the first three formants. SDC has also used several FFT algorithms as well as an algorithm for cepstral pitch determination.

The group at Texas Instruments, which is not really a part of the ARPA SUR community, has relied heavily [23] on acoustic processing. Only short-term amplitude spectra of the speech signal are used. The preemphasized speech signal is resolved into 16 frequency bands between 300 and 3000 Hz. They have had good success both in terms of recognition and in being highly resistant to channel noise.

The Sperry Univac recognition approach [24] uses prosodically detected stress patterns and syntactic structure in aiding a partial distinctive-feature-estimation procedure. Smoothed frequency spectra needed for formant tracking are obtained from LPC using 14 predictor coefficients. Simple peak-picking is used for formant estimation. Energy and fundamental frequency time functions are computed for prosodic features analysis. Total energy is computed every 10 ms, as well as the fundamental frequency (by autocorrelating the center-clipped acoustic time waveform).

From this brief summary of the acoustic approach of various researchers, it is seen that the various SUS workers seem to be using LPC techniques and formant tracking at the SUS front end. Some of the speech workers outside this community seem to be following the trend (e.g., NSA, Bell Laboratories) while others seem to be taking other approaches (e.g., Texas Instruments, various Japanese workers).

Other factors of importance in the acoustic analysis include the segmentation questions as well as those having to do with the background noise, speaker characteristics, and prosodic features. These factors are discussed in Sec. V after a brief look at the importance of some speech synthesis work in Sec. IV.

IV. SPEECH SYNTHESIS AND PRODUCTION APPROACHES

So far, consideration has been limited to the analysis of speech input and, in some cases, its subsequent synthesis, or more precisely, its subsequent reconstitution. In this section, we consider a different, but not completely unrelated, process—using speech as a computer output medium. This process is usually called speech synthesis; but it should not be confused with the same term in the phrase "speech analysis/synthesis." The computer that produces speech output from internally stored data or text must also have stored information about linguistic rules as well as knowledge of the acoustic constraints. This is the same sort of information that an SUS must use in its processing except that the two problems are the inverse of one another. This section summarizes a few key developments in the area of speech synthesis that are of interest to the speech understanding problem. No attempt will be made to describe the various kinds of speaking machines that have been or are being built since this is certainly beyond the scope of this report.*

In synthesis applications requiring a large and flexible vocabulary, speech information must be stored in a form that makes it easy to produce a great variety of messages in a relatively efficient manner with an acceptable speech quality. Two methods, among others, seem to fulfill these requirements. The first method is based on formant synthesis in which word libraries are stored in terms of formant frequencies. An utterance is formed by concatenation of word-length formant data with word boundaries, pitch, and various prosodic features calculated according to linguistic rules. The second method is a synthesis from output text. An automatic syntax analysis of the text is made with sound pitch and duration computed from stored linguistic rules. Vocal tract shapes are then calculated for the utterances.

FORMANT SYNTHESIS

In the formant synthesis method, typical parameters might include the pitch period being specified to the nearest 0.1 ms, the gain to 1 part in 100, and the formant frequencies to the nearest 1 Hz with a parameter update 100 times per second. The formant synthesizer used in an experimental system is shown in Fig. 7. The synthe-

* Several excellent references are available on speaking machines, e.g., Refs. 25 and 26.

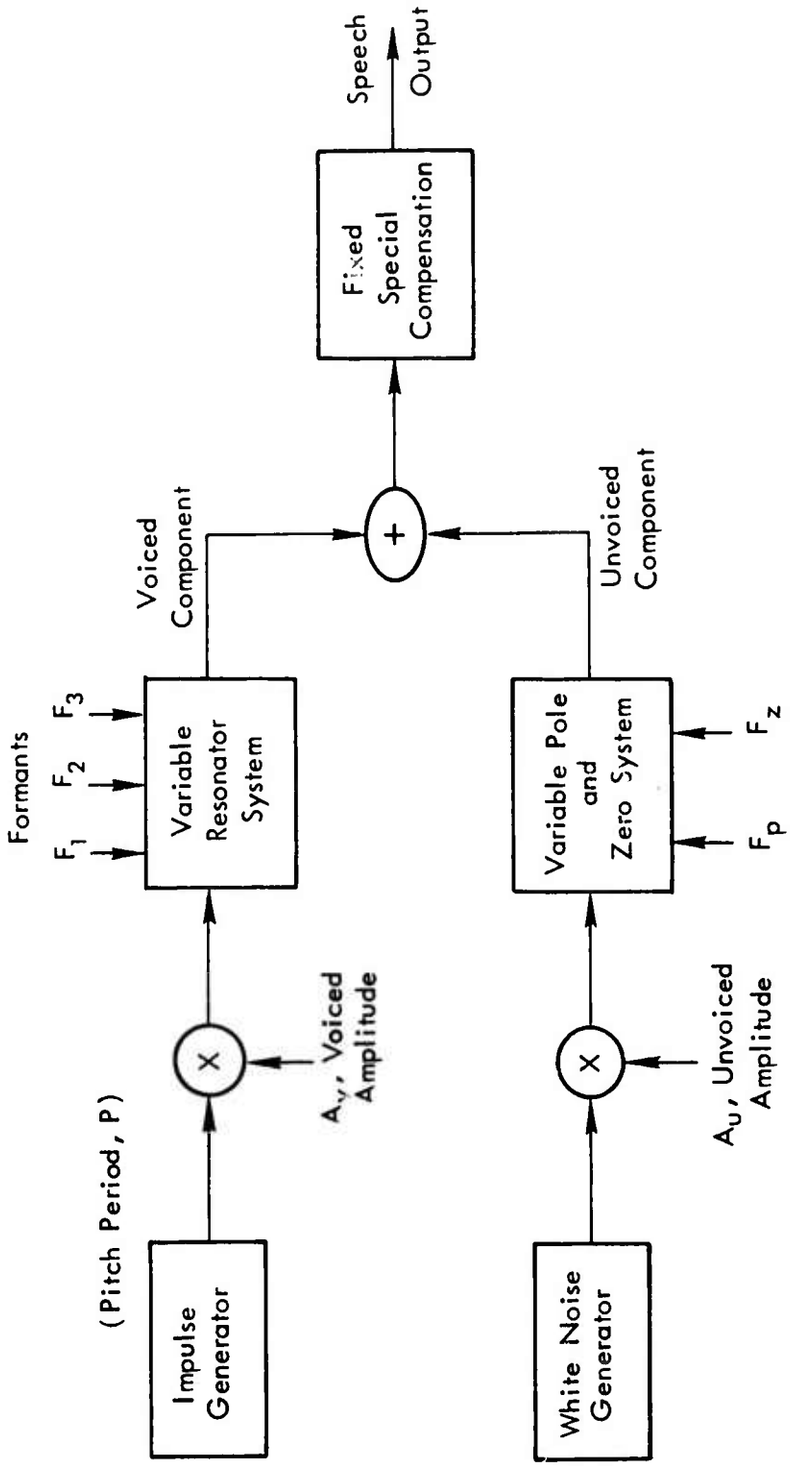

Fig. 7—Digital formant synthesizer (adapted from Ref. 22)

sized messages in this case are fabricated from pre-analyzed spoken isolated words. A message is synthesized by the concatenation program (illustrated in Fig. 8) that has the printed word string as input. The program uses separate strategies for prosodic features and segmental features. The final smoothly varying formant-synthesis parameters are then supplied to the digital synthesizer for actual synthesis. Timing information is derived from an external specification of the duration of each word that is dependent on context or is derived from calculations based on linguistic rules. The pitch contour is determined by using an archtypal contour or by rule. The formant contours are changed in duration when necessary and merged to form smooth continuous transitions. The speech is synthesized using the chosen prosodic and segmental features. The speech is generated in real-time at a 10-kHz sampling rate. It is thought [25] that simple rules for pitch and timing are sufficient for reasonable synthesis in certain limited-context applications.

SYNTHESIS FROM PRINTED TEXT

Synthesis from printed text [25] is another speech synthesis technique that is of interest. In this technique, there is a requirement for a large storage capacity. It uses a phoneme dictionary and contains automatic rules for conversion from written text to speech with reasonable timing and intonation. It synthesizes unrestricted speech from a dynamic characterization of the human articulatory system. The vocal tract model for phoneme synthesis is described by seven parameters that actually are cross sectional areas along the vocal tract. The articulatory model requires an input of discrete phonetic symbols as well as pitch and duration data as shown in Fig. 9. Printed text is converted into discrete phonetic symbols along with pause, stress, timing, and pitch assignments. These rules, by the way, are also applicable to the formant synthesis technique described above. The dictionary provides a phonemic transcription of each word with lexical stress marks, usage probabilities, and content-function distinctions. The syntax analyzer assigns pause probabilities, selects alternate word pronunciations according to usage, and alters stress. A decision on the kind of pitch contour and pause for each grammatical unit is made at the stress and pause assignment block. Pitch marks and timing control marks are assigned to each phoneme using a word boundary rule and a stress and termination rule. The text-conversion program is complete at this point and control goes to the articulatory model that completes the connected speech synthesis.

Both of these techniques for speech synthesis use linguistic as well as acoustic information in their operation. The acoustic processing techniques necessary to characterize the formant description in one case and the phonemic description in the other case are not much different than some of the analysis/synthesis techniques that were considered in Sec. III. The sequence of events and the governing rules are reversed in the synthesis procedure from what they are in the analysis procedure.

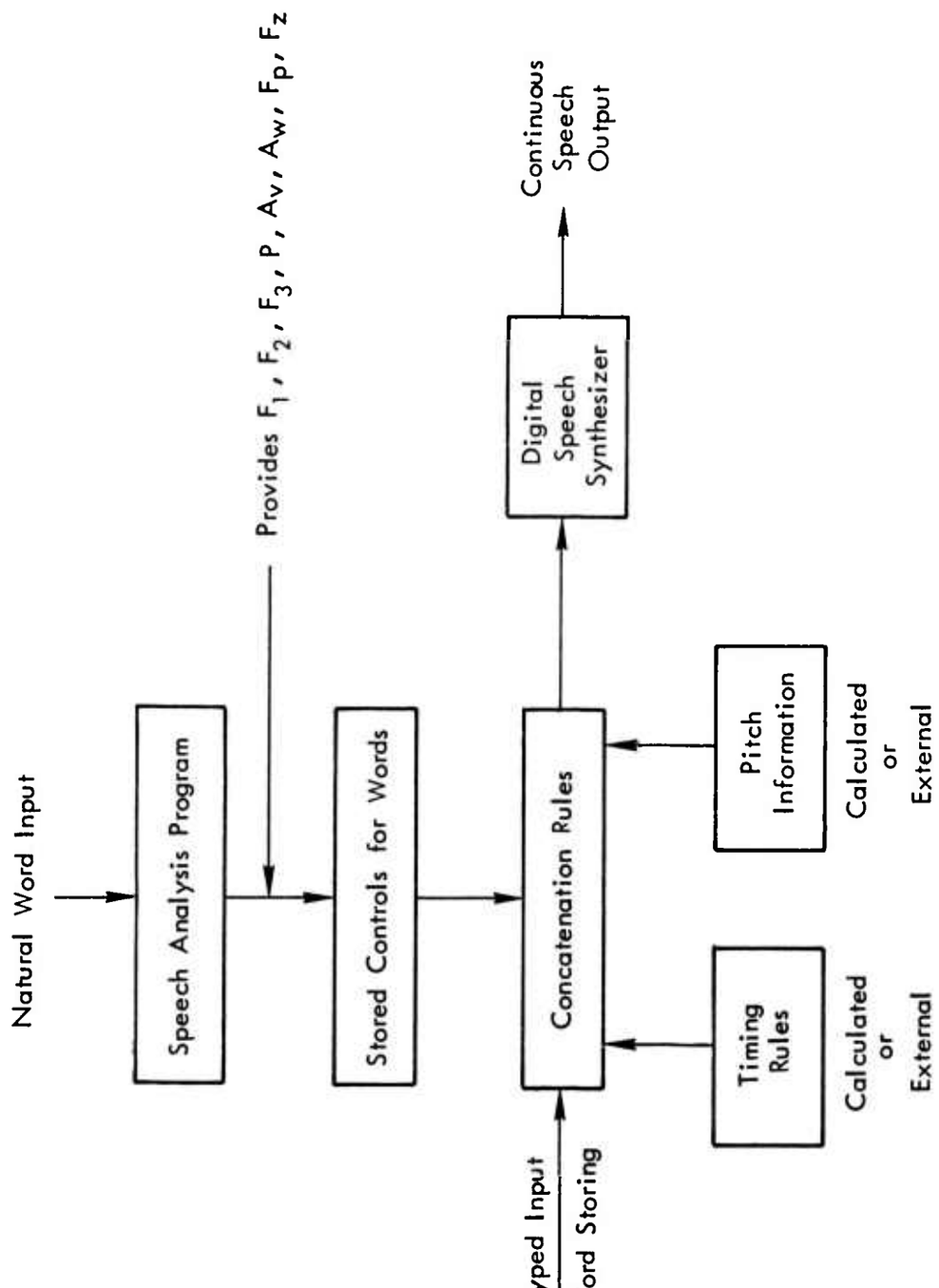

Fig. 8—Concatenation program (adapted from Ref. 22)

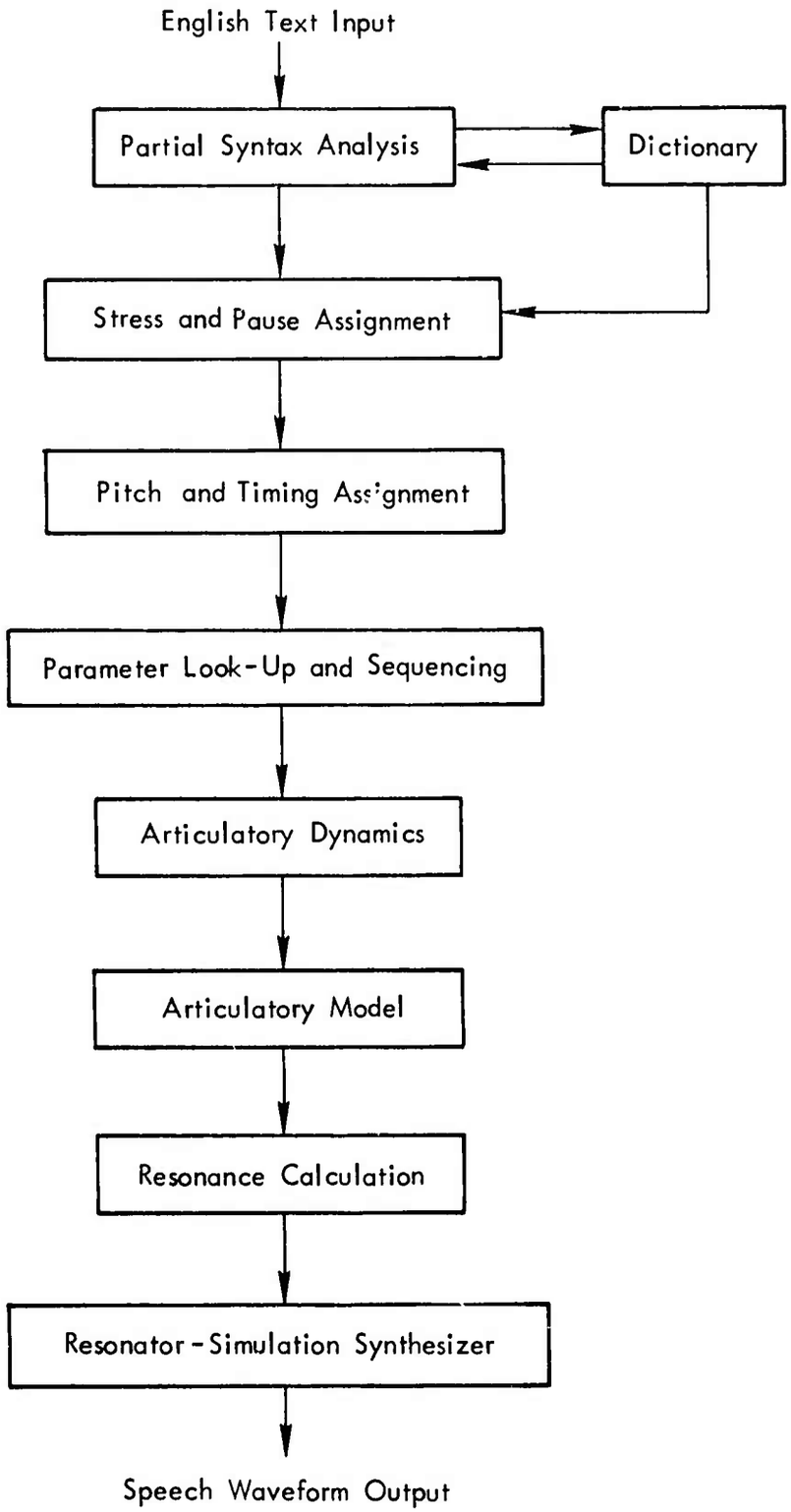

Fig. 9—Synthesis from text (adapted from Ref. 22)

V. COORDINATING THE ACOUSTIC LEVEL PROCESSING WITH THE HIGHER LEVELS

Now that some consideration has been given to the parameterization of the acoustic signal, it seems appropriate to consider how this can be tied more closely to the higher level processing in SUSs. The major questions here involve not only the choice of the acoustic processing technique and its implementation, which have already been discussed at length, but also the determination of the optimum mix between acoustic processing and higher level processing. The specific question areas of the acoustic processing are by now well known: parametric variability due to segmental, syntactic, and semantic context; multiple speakers; speaker variations; environmental noise; transducer characteristics; prosodic variations; etc. It would also be desirable to be able to put limits on the amount of processing, the amount of memory, and the cost of the acoustic processing in terms of these problems and question areas. Consequently, we will consider some of the system parameters involved.

TRANSDUCER

The specific transducer as well as the design philosophy of the acoustic processor are of prime importance. A telephone, because of its wide utility, is preferred over a high quality microphone for many applications. Unfortunately, the dynamic range and bandwidth limitations of a telephone, as well as other signal distortions, impose constraints on the acoustic processing. The limitations on bandwidth give rise to difficulties in detection of some fricatives, for example, since the primary cues are often beyond the upper cut-off frequency. In addition, there is attenuation distortion that is frequency dependent. Both the attenuation distortion and the bandwidth limitations are themselves equipment dependent. Further, noise considerations are important—random noise due to the digitalization process and discretization noise associated with the digital transmission over long-distance lines. There is also an envelope delay distortion that is a phase distortion in which frequencies at the low and high ends exhibit envelope delays relative to those at mid-band. Crosstalk is another problem, and a difficult one since it cannot be readily handled by appropriate preprocessing. The transducer problems, while annoying, are not beyond solu-

tion with current techniques at a cost that is felt to be modest compared to the higher level processing and memory costs.

COOPERATIVE/UNCOOPERATIVE SPEAKER

A cooperative speaker is certainly an imperative requirement for present SUSs to avoid the need for the wide dynamic range necessary for accomodating uncooperative speakers (i.e., a speaker can raise his level over about a 60 dB interval, from a whisper to shouting, if he so desires). Even at conversational levels, an individual may exhibit a 30 to 40 dB dynamic range. At conversational levels, the speaker-to-speaker variation (in terms of average sound levels) may be anywhere from 50 to 75 dB. Thus, there are conflicting requirements between a wide dynamic range, which would accommodate these varying kinds of input, on one hand, and a low-bit-rate processor on the other hand.

NOISE SOURCES

Another potential major problem is that of discrimination between the speech source and various noise sources. Noise, in this context, is anything that might interfere with the desired speaker's acoustic signal, including environmental noise (e.g., air conditioners, printer noise), system noise (e.g., quantization noise), and other speakers in the area, as well as extraneous sounds from the speaker himself (e.g., clearing the throat, "ahs," "hmmms"). Directional microphones can be of considerable help with some of these problems. A telephone as input transducer would make the problem more difficult in some respects. A noise-free acoustic chamber, while alleviating much of the difficulty, is unreasonable in an operational sense. In any case, the noise limitation does not appear to be a very serious one.

SEGMENTATION

Segmentation is another area of concern. It is a difficult problem, as attested to by the large number of approaches that have been proposed. Because of its potential for reducing the search space, segmentation remains a very important area for investigation.

The acoustic data contain very few cues about segmentation and word boundaries. However, there are some rules relating the parameters of the speech wave to the behavior of the articulatory system (acoustic-phonetic rules) and the rules that specify the phonetic segments from the lexical (phonemic) context (phonological rules). With proper normalization, it may be possible to make some of these rules speaker independent. One of the most important sources of error is the variability of segmental parameters of a given phoneme in different contexts. The acoustic-phonetic rules are, in part, attempts at predicting this coarticulation or overlapping of adjacent phonemes. These rules, however, are not yet systematically formulated

in a way that is appropriate for speech analysis. Significant parametric variations are also due to syntactic and semantic context, and much of this behavior is governed by phonological rules.

PROSODIC INFORMATION

Prosodic information can, in part, be determined from the raw acoustic data and, in part, from linguistic and context data. This is one area where feedback from higher levels of the recognition program could be used in meeting hypotheses on various entries. Prosodic features generally include timing, amplitude, and pitch data, as well as information on pauses. These features can add supra-segmental variability and may contain information that is critical for an actual understanding of the utterance.

Researchers have tried various approaches [27] to the prosodic feature question. One approach [28] has been the investigation of the interaction between segmental and supra-segmental features in the word spotting application—i.e., detecting a specified word in the context of continuous speech. It makes use of a primary recognition program at the acoustic level providing the data source for algorithms that recognize strings of segments or words. Segment boundaries are determined by a change of classification between two adjacent time samples where the classification is presently in terms of wide-band spectral characteristics (e.g., silence, vocal murmur, voiceless stop, vowel). There is work at Univac under ARPA sponsorship [24] using prosodically-detected stress patterns and syntactic structure in aiding a partial distinctive-feature-estimation procedure. There is also interest in using prosodically-detected syntactic structure to aid syntactic parsers and semantic processors.

IMPLEMENTATION

The processing, memory, and cost constraints have received very little attention, and there is good reason for this—not much is really known. The processing power is primarily devoted (1) to the parameter extraction process, and (2) to the searching and matching operations. In the first instance, specialized hardware is likely to be used, and this may not be too different than the special-purpose FFT units now proliferating. The question of memory structure is open depending on available technology (e.g., optical store, film store) and logical processing structure. Likewise, the overall cost question is open since it not only depends on the hardware costs of the speech understanding computers, but also on the various cost factors that go into the particular application. This will be more fully discussed in subsequent reports.

SUMMARY

It would seem that many of the problems at the acoustic front end of the SUS are under control. The transducer limitations can be handled by current techniques (e.g., Texas Instruments' use of the telephone in its recognition schemes). Noise will always be of some concern (e.g., crosstalk on a telephone line), but certainly not of overriding concern. Progress is being made in segmentation, but this remains a prime area of investigation. The use of prosodic features is of importance as witnessed by the advances in that area being made by some of the ARPA contractors. Speaker variability and speaker-to-speaker variations, while not discussed fully in this report, remain a problem. The processing, memory, and cost constraints of the acoustic processing and the SUS, itself, will be more fully addressed in subsequent reports.

REFERENCES

1. Rothauser, E., "Speech in Digital Communications Systems," *IEEE Trans. Audio & Electroacoustics,* Vol. AU-21, No. 1, February 1973, pp. 21-26.
2. Schaefer, R. W., "A Survey of Digital Speech Processing Techniques" *IEEE Trans. Audio & Electroacoustics,* Vol. AU-20, No. 1, March 1972, pp. 28-35.
3. Schroeder, M. R., "Vocoders: Analysis and Synthesis of Speech," *Proc. IEEE,* Vol. 54, May 1966, pp. 720-734.
4. Atal, B. S., and M. R. Schroeder, "Adaptive Predictive Coding of Speech Signals," *BSTJ,* October 1970, pp. 1973-1986.
5. Gold, B., "Techniques for Speech Bandwidth Compression," *JASA,* Vol. 38, No. 1, July 1965, pp. 2-10.
6. Birch, J. N., and N. R. Getzin, *Voice Coding and Intelligibility Testing for Satellite Based Air Traffic Control Systems,* NASA Contract NAS5-20168, Final Report, Magnavox Company, Silver Springs, Maryland, 1971.
7. Tierney, J., et al., "Channel Vocoder with Digital Pitch Extractor," *JASA,* Vol. 36, No. 10, October 1964, pp. 1901-1905.
8. Flanagan, J., *Speech Analysis, Synthesis, and Perception,* McGraw-Hill Book Company, Inc., New York, 1966.
9. Carlson, J., "Digitized Phase Vocoder," *1967 IEEE Conference on Speech Processing,* Boston, Massachusetts, November 1967, pp. 292-302.
10. Bell, C. B., et al., "Reduction of Speech Spectra by Analysis-by-Synthesis Techniques," *JASA,* Vol. 33, No. 12, December 1961, pp. 1725-1736.
11. Coker, C. H., "Computer Simulated Analyzer for a Formant Vocoder," *JASA,* Vol. 35(A), No. 1, October 1963, pp. 1911-1957.
12. Schaefer, R. W., and L. R. Rabiner, "System for Automatic Formant Analysis of Voiced Speech," *JASA,* Vol. 48, No. 2 (Part 2), 1970, pp. 634-648.
13. Bially, T., and W. M. Anderson, "A Digital Channel Vocoder," *IEEE Trans. Comm. Tech.,* Vol. COM-18, No. 4, August 1970, pp. 435-442.
14. Noll, A., "Short-Time Spectrum and Cepstrum Techniques for Vocal-Pitch Detection," *JASA,* Vol. 36, No. 2, February 1964, pp. 296-302.
15. Oppenheim, A., "Speech Analysis-Synthesis Based on Homomorphic Filtering," *JASA,* Vol. 45, No. 2, February 1969, pp. 458-465.
16. Atal, B. S., and S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," *JASA,* Vol. 50, No. 2, Part 2, 1971, pp. 637-655.

17. Dunn, J. G., "An Experimental 9600-bits/s Voice Digitizer Employing Adaptive Prediction," *IEEE Trans. Comm. Tech.*, Vol. COM-19, No. 16, December 1971, pp. 1021-1032.
18. Weinstein, C., and A. Oppenheim, "Predictive Coding in a Homomorphic Vocoder," *IEEE Trans. Audio & Electroacoustics*, Vol. AU-19, No. 3, September 1971, pp. 243-248.
19. McCandless, S., et al., SURNOTE 71, *Fundamental Frequency Formant, and Segmentation Data on Six Data Base Test Sequences*, Massachusetts Institute of Technology, Lincoln Laboratory, February 27, 1973.
20. Reddy, D. R., et al., SURNOTE 44, *A Model and a System for Machine Recognition of Speech*, Carnegie-Mellon University, Pittsburgh, Pennsylvania, September 1972.
21. Walker, D. E., *Speech Understanding Research*, Stanford University Annual Technical Report, ARPA Contract DAHC04-72-C-0009, February 1973.
22. Ritea, Barry, SURNOTE 7, *System Description and Data Representations for SUS*, Systems Development Corporation, December 9, 1971.
23. Doddington, G., *Final Demonstration, Speaker Verification Study*, Texas Instruments, RADC Contract F30602-72-C-0294, May 23, 1973.
24. Lea, W. A., et al., *Prosodic Aids to Speech Recognition, Final Report*, Univac Corporation, ARPA Contract DAHC15-72-C-0138, April 15, 1973.
25. Flanagan, J. L., et al., "Synthetic Voices for Computers," *IEEE Spectrum*, Vol. 7, No. 10, October 1970, pp. 22-43.
26. Flanagan, J. L., "Voices of Men and Machines," *JASA*, Vol. 51, No. 5, May 1972, pp. 1375-1387.
27. Lindgren, N., "Machine Recognition of Human Language: Part I," *Spectrum*, Vol. 2, No. 3, March 1965, pp. 115-136; "Machine Recognition of Human Language: Part II," *Spectrum*, Vol. 2, No. 4, April 1965, pp. 45-59.
28. Li, K., G. W. Hughes, and T. Snow, "Segment Classification in Continuous Speech," *IEEE Trans. Audio & Electroacoustics*, Vol. AU-21, No. 1, February 1973, pp. 50-57.