AUTOMATIC SPEECH RECOGNITION BASED ON A NEW SEGMENTATION PROCEDURE

Thesis for the Degree of Ph. D.

MICHIGAN STATE UNIVERSITY
EARL JOHN CRAIGHILL

1971

This is to certify that the

thesis entitled

Automatic Speech Recognition Based On

A New Segmentation Procedure

presented by

Earl J. Craighill

has been accepted towards fulfillment
of the requirements for

Ph. D. degree in EE

William Kilmer

Major professor

Date 4/13/71

O-7639

ABSTRACT

AUTOMATIC SPEECH RECOGNITION BASED ON
A NEW SEGMENTATION PROCEDURE

By

Earl J. Craighill

A procedure for segmentation of an acoustical speech signal is crucial to the design of any system for automatic speech recognition (ASR), yet no adequate scheme currently exists. This study proposes and investigates the implementation of a procedure for segmenting input in the form of connected speech from divers speakers using unlimited vocabularies.

A segmentation procedure which assigns linguistic elements, such as phonemes, to contiguous acoustical signal intervals would be hopelessly complex because of the many-to-many correspondence between currently used linguistic elements and portions of the acoustical signal. Instead, we propose a method for dividing the acoustical signal into analysis epochs with minimal linguistic specification so that they are independent of speaker and context.

Each epoch is defined by homogeneous signal characteristics; that is, a generation model is identified with associated parameters, and nonlinear time-varying differential equations are derived for these parameters. The equations are used to track the parameter values, and an epoch boundary is set at the point where they no longer predict (within a threshold) the characteristics of the observed speech signal. From the functional forms of the differential equations, we derive further processing algorithms

(analogous to data-dependent adaptive filters) for each epoch. Identification of the functional forms gives a gross linguistic classification which forms the basis for classification of the epoch.

The differential equations are characterized in terms of sliding moment averages of envelope and zero-crossing estimates on bandpass-filtered speech signals. This method of estimation is amenable to low-cost hardware implementation and requires few computations; thus, connected speech may be analyzed in real time without overloading a standard general-purpose computer. Asynchronous, real-time classification is achieved by decomposition of the decision algorithm by a process similar to that used in Kilmer's model of the reticular formation.

Overlapping bandpass filters are used to give an initial separation of acoustical features. Experimental evidence shows how this reduces the speaker dependence of further acoustical measurements. A decision logic structure is specified and discussed, showing that it is possible to select appropriate preprocessing procedures to focus attention on significant features of an acoustical signal epoch and to accentuate signal characteristics closely correlated with linguistic features. This preprocessing, when coupled with the syntactical structures developed from theoretical linguistics, is hopefully a first step in recognizing human connected speech from different speakers.

AUTOMATIC SPEECH RECOGNITION

BASED ON A NEW SEGMENTATION PROCEDURE

By

Earl John Craighill

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Electrical Engineering

1971

ACKNOWLEDGEMENTS

The Author wishes to thank Professor William Kilmer for his continuing support, guidance and patience during the preparation of this thesis. These qualities were shown both in and out of the classroom and are deeply appreciated.

For their constructive comments and evaluation of this thesis, the Author wishes to acknowledge the other members of his doctoral committee: R. C. Dubes, T. Guinn, C. L. Park, R. F. Reid, and H. Salehi and also colleagues at Stanford Research Institute: W. F. Foy and W. P. Rupert.

The research was supported at Michigan State University under Air Force Contracts No. Af-AFOSR-1023-66,67,68 and at Stanford Research Institute under IR&D Project No. 656531-329.

Without the support and encouragement of many people, this thesis would not have been finished. A few of these people are: my supervisor, D. F. Babcock; secretaries, A. Guinn, S. Peterson and K. Spence; my mother and father and my wife Karilyn.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

I-A Overview

A procedure for segmentation of an acoustical signal is crucial to the design of automatic speech recognition (ASR) systems. As yet, however, no adequate procedure exists for real-time automatic recognition of connected human speech from several speakers. Principles from communication theory and linguistic theory must be incorporated in order to derive an efficient segmentation procedure. The language of modern communication theory, familiar to the electrical engineer, most appropriately describes the input with which we are concerned. For this study, we limit the input to connected phrases of naturally spoken human language that have been transduced into time-varying analog voltages. The output of an ASR system, usually in the form of a sequence of linguistic elements*, is generally described in the framework of linguistic theory, primarily phonology.

At first glance, the goals of communication theory and phonology (namely, an accurate description of the current state of the process, acoustical signal, or sequential linguistic elements) seem to be compatible. However, when one considers the large number of variations in the acoustical signal possible for any given linguistic element, the situation becomes hopelessly complex. Many attempts have been made to eliminate this variation and thereby preserve only the meaningful

*These linguistic elements may be phonemes, distinctive features or words. We are specifically thinking of only one level of classification rather than a composite process such as identification of phonemes and then morphemes. Our recommendation for a first element is smaller than the usual phoneme or distinctive feature.

relationships of the linguistic elements. Successful decoding of this complex acoustical signal by human listeners involves at least the application of knowledge acquired from previous experiences of hearing and speaking natural language and the listener's expectation of what will be said. Thus, at this level, the basic assumptions of engineering communication theory are no longer valid, and there is no applicable strong property of ergodicity.

The purpose of this thesis is to describe a segmentation procedure which not only specifies basic units for recognition but also gives an adequate description of the complicated speech acoustical signal. This description is prescribed by the requirements of further linguistic decoding (words, phrases, ...). Further, the segmentation procedure that identifies lower units will direct the higher levels of decoding so that the search space is kept within practical bounds. The segmentation procedure requires three subsystems based on a parametric generation model of the acoustical signal:

(1) Initial estimation of parameters.

(2) A classification based on parameter estimates for signal types.

(3) Selection of appropriate time-varying filters operating on the input to give refined parameter measurements.

The requirements of these diverse topics are discussed in terms of a representation of the acoustical signal which is developed from the viewpoint of time-varying differential operators. Its use in deriving estimators and detecting initial changes in these estimators is verified experimentally.

In the remaining sections of this chapter, currently used segmentation procedures are discussed in light of the complex nature of the information-bearing features present in the human speech acoustical signal. A parallel interrelated feature structure is described that is capable of recognizing a shift of the pertinent information from one feature to another. Linguistic information is conveyed with respect to two levels (the vowels of an utterance form a primary, and the consonants are incorporated by perturbations of this primary substrate). In order to unravel this complicated structure, broad classes of speech sounds that represent different types of signal characteristics must be defined; this classification can then be used to direct further analysis for recognition. By this method, formant* theory is related to higher levels of linguistic decoding. Various preprocessing schemes are considered which are commonly applied to ASR systems for the purpose of isolating individual formants. To satisfy the requirement for real-time operation, a preprocessing scheme is chosen which uses a bank of overlapping wideband filters (with sufficient bandwidth to avoid distortion) to remove noise and to provide a compact representation of the salient features required for the recognition task. Real-time operation requires decomposition of the decision process resulting in fewer computations and a recognition structure tailored to the complicated overlapping nature of the speech signal.

---

*By a formant, we mean the resulting time waveform for one cavity of the vocal tract excited by glottal pulses, or frication noise.

In Chapter Two, the acoustical properties of the speech signal are modeled as a composite nonstationary stochastic process and the mathematics of communication theory are used formally to describe the process's complicated nature. One isolated formant is modeled by a time-varying differential operator involving envelope, frequency, and bandwidth parameters. The inadequacies of fixed-frequency types of analysis (such as sliding Fourier transforms) are discussed, and requirements for low-distortion filtering are derived. Then the transient response of linear filters to envelope and frequency changes found in typical acoustical signals is derived in a way that offers new insight into the behavior of analysis procedures and defines requirements for the preprocessing wideband filters. Formulas for real-time pointwise estimators of the significant parameters are derived, and a predictive differential equation segmentation procedure is specified which will specify epochs in the acoustical signal having homogeneous signal characteristics.

In Chapter Three, this segmentation procedure is discussed within the framework of traditional linguistic theories. The complicated structure of human communications requires additional mechanisms

(1) To determine the linguistically significant changes in signal parameters, and

(2) To incorporate contextual information into the decision process (which, in turn, resolves ambiguities and directs further classification).

Structural theories are modified to include recognition and to show the effects of linguistic rules on lower elements (effects of stress

on vowels, etc.). The use of the segmentation recognition procedure proposed here is basic to a feed forward system, thus eliminating complicated feedback analysis-by-synthesis techniques.

In Chapter Four, the formant representation and segmentation results allow application of state-of-the-art detection/recognition techniques* to a restricted speech signal (without the complex interrelationships between features). Study of the Bayes minimum risk solution reveals that the primary concept is a probability mixture formula for the outputs of nonlinear estimation filters, each tailored to a possible generating model for the input signal and (correlated) noise. Several difficulties are noted for implementation of this optimal solution: realization of the nonlinear filters, correlation between different (suboptimum) filter outputs, and conflict between classifications on different filter outputs.

It is concluded that a heuristic recognition scheme tailored more to the filter bank used in this study would be a better choice. Techniques are developed to reduce the dependence among the probabilities computed on the different filter outputs.

A first-level recognition system which can operate asynchronously in real time is described. A nonlinear iterative structure determines which filters have pertinent formant information. Specialized algorithms derived from linguistic rules are then applied to these filter outputs to determine the needed information for classification of this particular

---

*Section I-E contains a discussion of terminology that is used in this study for the pattern recognition discussions.

signal epoch. The output is a classification which is compatible with higher levels of linguistic analysis. A second stage with formant tracking filters guided by the initial classification gives the ability to focus attention on only the desired acoustical features. Thus, the complex acoustical signal can be segmented in time into homogeneous epochs and also concurrent features of varying frequency with well-defined mathematical models and time-varying parameters.

A total system design incorporating this segmentation procedure as a first step will facilitate the use of human speech as input to machines for robot control, text manipulation, command and control of space vehicles, and many other man/machine tasks.

I-B THE STRUCTURE AND INTERRELATIONS OF ACOUSTICAL FEATURES IN HUMAN SPEECH SIGNALS

The object of an ASR system is to determine recurrent elements from measurements made on acoustical speech signals. Figure 1 shows a composite of several approaches to Automatic Speech Recognition based on the theoretical encoding of speech shown in the upper block. This theoretical encoding is motivated by Hockett's[1] discussion of a GHQ (grammatical headquarters) emitting a discrete flow of morphemes which are encoded into a discrete flow of phonemes. Then, a speech transmitter converts the discrete flow of phonemes into a continuous speech signal.

The determination of parameter values for each idealized element is motivated by the following studies. Peterson and Barney[2] measured first and second formant frequencies of nine English vowels in a fixed consonantal context (the word h____d). Gerstman[3] re-worked their data, normalizing for each speaker, showing a sufficient amount of separability of the measurements for vowel classification (in a fixed context for isolated words). The correspondence between a fixed frequency or hub of origin and consonants was first proposed by Potter, Kopp, and Green.[4] Classification of stop consonants by association with a frequency value was modified by Cooper et al[5] and Yilmaz.[6] They proposed consistent measurements for stop-consonant classification could be made relative to the following vowel formant frequencies. The slurring box accounts for perturbations (hopefully slight) of these parameter values caused by environment and speaker variations.

The first step in recognition is a division of the acoustical speech into time epochs. The segments studied may be separated by epochs (portions

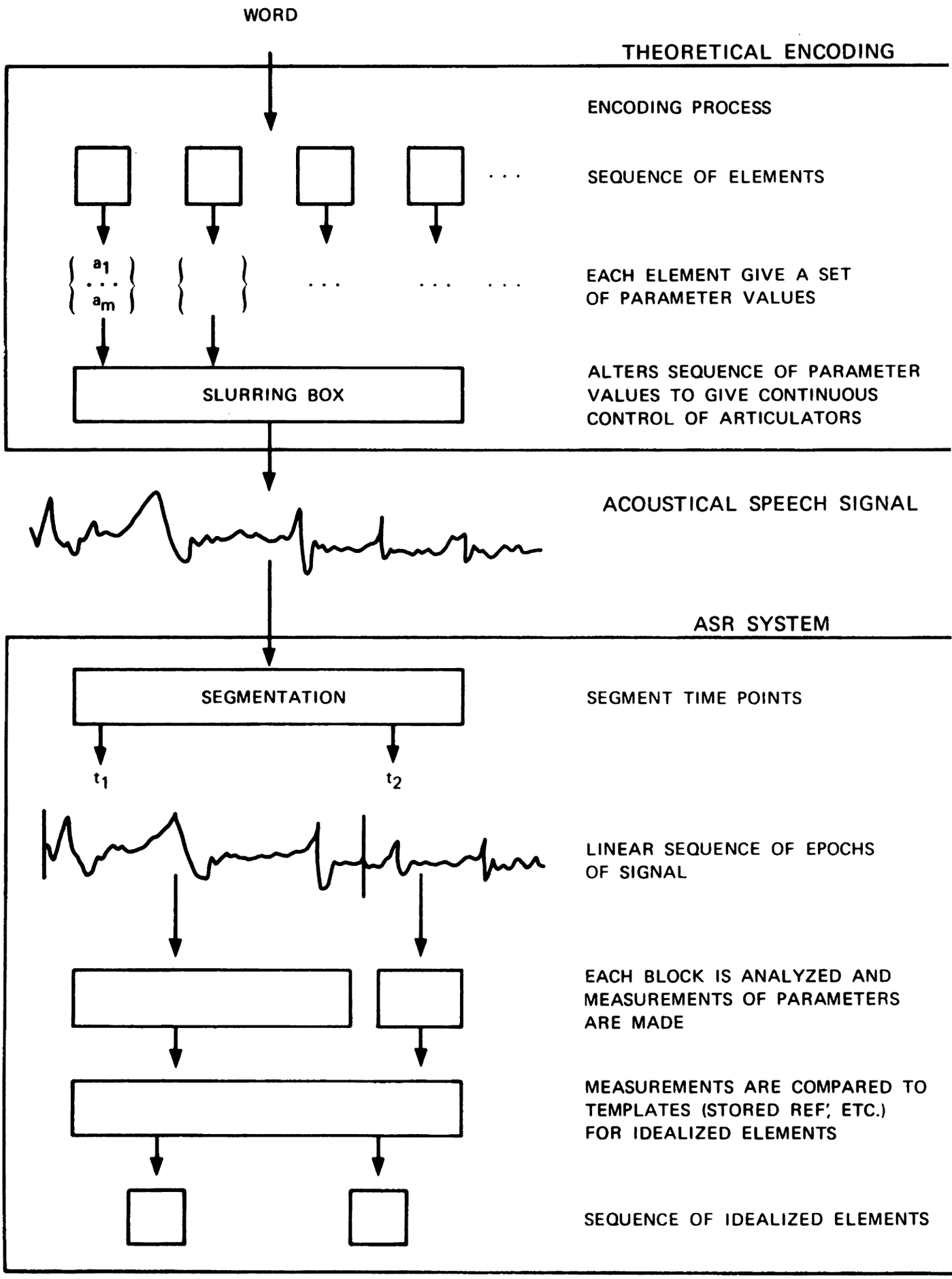


FIGURE 1 TYPICAL ASR SYSTEM BASED ON DISCRETE ENCODING MODEL

of signal) rather than points, as in the case of Reddy[7] who analyzes only steady-state portions (i.e., portions with constant values of envelope and frequency) and ignores the transition portions between them. The opposite approach is taken by Dixon et al.[8] in their analysis and segmentation procedure. They define a new element called the transeme, which is a "dynamic segment describable on a production basis as the transition from one relatively steady-state articulatory configuration to another."

The criterion for segmentation and further analysis may not be related to linguistic elements at all, as in the case of Gazdag.[9] His segmentation points are determined completely in terms of the measurement procedure that he uses to analyze the speech waveform; hence they are independent of any exterior linguistic criterion. ASR systems developed along these lines have no ability to ignore speaker and environment variations or free phonetic variation; i.e., in midwest English, prevoicing before [b] or [d] is optional. Usually a separate "case" (pattern class) is set up for each; hence the success that these various ASR systems have in isolated sound situations or in one-person conversational speech cannot easily be extended to connected conversational speech for many speakers.

Harris[10] has discussed the extremely difficult problem of trying to define linguistic elements as direct descriptions of portions of the flow of speech. He finds it convenient in his analysis to define certain elements which extend over quite long periods and others which extend over short periods. "In the course of reducing our elements to simpler combinations of more fundamental elements, we set up entities

such as junctures and long components which can only with difficulty be considered as variables directly representing any member of a class of portions of the flow of speech." (p. 18) A similar formalization in the early work of Fant, Jakobson et al.[11] describes distinctive features that are parallel rather than serial descriptions of the acoustical waveform. Extensions of this approach by Chomsky and Halle[12] are discussed at the end of this section. Bobrow, Klatt, and Hartley[13] have proposed an ASR system based on this idea and derived independent parallel features from the acoustical signal and performed classification on those features. Other ASR systems using independent features have been proposed by Hill[14] and Focht.[15] Bobrow et al. discuss the difficulties of recognizing conversational speech for divers speakers in terms of:

(1) Consistency of each speaker in repeating words for training (giving rise to phonetic variation)

(2) Speaker-dependent variation in their measurements (shifts in formant frequency location)

(3) Segmentation of longer utterances.

These difficulties are caused in part by the extremely complex nature of parallel features and the interrelations between them. Öhman[16] has studied various vowel/consonant/vowel (VCV) combinations and has stated that it is impossible to treat even these short utterances as three successive gestures. It is possible to analyze them only by considering the stop-consonantal gesture as superimposed on the substratum determined by the two vowels and the transition between them. Houde[17]

has investigated this further by means of X-ray movies of the configuration of the tongue during articulation. The dynamic trajectories of points on the tongue during articulation of VcVcV nonsense words can be decomposed into target-directed (targets are long duration steady-state vowel positions) and deviation (90^{o} to target-direction) components.* Five facts are clear:

(1) The deviation component is characteristic of the consonant ([b] and [g] were used).

(2) The characteristic deviation for [b] and [g] was not toward a target or hub but rather a consistent deformation of articulator (primarily tongue) configuration.

(3) Targets of preceding vowels are changed by the consonant (i.e., I in [Ige] has a different steady state position than I in [I b e]).

(4) Stress placement affects vowel target positions.

(5) Timing of target-directed component was dependent only on distance between target positions and not on speed of articulation, speaker or consonantal environment for the limited data investigated.

We can discuss these results in a way more compatible with linguistic theory by use of Lamb's[18] concept of a medium as a most unrestricted (or most predictable) form and then describe the pertinent features which convey information as perturbations of that medium. He defines a phonetic

*This decomposition is slightly different from Houde's, in order to demonstrate the concept of overlapping features.

feature as distinctive if its presence is not determined by its environment. This idea may be extended to explain the Ohman and Houde data by stating that the vowel-to-vowel transition is actually the medium for the consonantal distinctions.

We should define acoustical features more generally than just those defining linguistic events.* These acoustical features may be classified as:

(1) Linguistic

(2) Speaker signature

(3) Speaker emotional state.

The interrelationship of all these features that are present simultaneously, preceding or following in time, may be correlated with the dominant (distinctive) feature, but this correlation is usually situation (speaker, context) dependent and thus can introduce much variation in determining recurrent elements. It has been pointed out by Harris that time of start and stop of different acoustical features may not be coincident in time. Thomas[19] suggests that a speaker is able to adjust only one formant frequency; other frequencies are allowed to fall where they may. He states further that this formant is always the second, but the data presented by Ohman does not support this. Rupert[20] has studied isolated words spoken by three males and two females; he suggested that:

*By "linguistic" we mean the specific content of the speech waveform that is being used to communicate a discourse or text. For the purposes of man/machine communication, this definition will be sufficient. We do not wish to get into a discussion of various gestures, intonations, etc., which can also convey information.

(1) Each speaker does consistently control at least one acoustical (linguistic) feature which is usually less than the entire acoustical signal (i.e., one or two formants).

(2) Although the controlled feature(s) (say, second formant) may not be the same in absolute value for all speakers, the time patterns are similar and can be identified by their recurrent nature.

(3) There is a high degree of recurrence across speakers of these controlled features.

(4) Other acoustical features (may be correlated with linguistic) that occur vary considerably according to speaker, phonetic environment, etc.

Öhman has proposed a motor-control model to partially explain his data as saying that for a VcV sound there are independent signals (or parameters in our theoretical model) for the first vowel, the consonant and the second vowel. The various muscles work in a coordinated fashion to produce continuous changes in articulatory configuration. This approach has actually been used to some extent in the work of Reddy. He first classifies his segments into phoneme classes (vowel, fricative, stop, nasal, liquid) and then performs a specialized analysis on each segment which is directed by the phoneme class label.

Based on this discussion we formulate the following premises about a feature description of the speech signal:

(1) Only a subset of the acoustical features present in a time epoch of speech are linguistically significant;

this subset can be recognized by the precise repeatable nature of its members. We do not mean precise values (formant frequencies = 500 hz, 1500 hz, and 2400 hz) but rather, precise time behavior within physical (motor control) and linguistic* constraints.

(2) Epochs of the acoustical signal can be equivalenced to classes determined by a subset of linguistic acoustical features. These classes can be defined (by the choice of the subset of features) in such a way that they are situation (context, speaker) independent. Roughly, the class labels are a generalization of the consonant, vowel labels used by linguists and also a refinement of Reddy's phoneme classes and Rupert's production modes (PM's).

(3) Further feature analysis is simplified considerably, and a more precise syllable (canonical form) analysis can be performed by a directed-search technique based on the above classification. This removes the inherent circularity in many classification schemes involving normalization (analogous to the visual recognition problem of finding an object of interest to focus on while it is out of focus).

(4) Once vowel (peak of syllable -- Hockett) classes are specified, they set up a primary formant transition structure.

---

*As noted by Öhman, consonantal variations of formant transitions are different for Russian speakers than for English speakers.

(5) Consonantal modifications are with respect to the primary formant structure and hence will be termed secondary.

(6) There is interaction between primary and secondary acoustical features, but the class labels can be assigned independent of this interaction.

The concept of precisely controlled features determined by phonetic environment at first appears similar to the distinctive features matrices proposed by Chomsky and Halle as the final linguistic idealized description of the speech waveform. However, there are two crucial distinctions:

(1) Significant features are chosen, and other (redundant) features are eliminated based on the simplicity of description and reduction of logical complexity in the encoding process. In speech recognition, the human is generally unaware of mathematical formulations when he is learning to speak; hence, the features he selects to emphasize and control precisely are chosen for communication with another human being and immunity to noise for that communication. Hence, an ASR system must determine the precisely controlled features that are present rather than formulate hypotheses about which ones would be easiest to analyze if they were present.

(2) Their concept of opposition is with respect to elements that can occupy the same time epoch (minimal pair). This involves a comparison of definite (albeit

situation-dependent) measurements of the present input with some representative set of measurements for the opposing element. Many investigators have noted the difficulty in this approach (Hemdal and Hughes[21]). The relative opposition concept of Rupert and Yilmaz does not have this difficulty, because a time epoch is compared to the preceding and successive epochs for its relevant opposition measurements. Hence, normalization becomes less of a problem.

In the following sections, we will expand these premises and show experimental evidence indicating a different description of the acoustical speech signal is necessary for an ASR system which more accurately measures timing and frequency characteristics.

I-C SEGMENTATION OF THE ACOUSTIC SPEECH SIGNAL INTO ANALYSIS EPOCHS

The optimistic goal of some segmentation procedures is to define time points and the acoustical signal such that the resulting sequence of signal epochs will correspond to a sequence of idealized linguistic elements. One then simply decides which linguistic elements each epoch is most like. In the previous section we discussed this approach and the resulting difficulties, especially in conversational speech involving long phrases. Bobrow et al. state that the purpose of segmentation should be a selection of appropriate measurements to be made, dependent on the phonetic context. Reddy's phoneme classes are directive in the sense that they select appropriate decision procedures to be used in analyzing each of his segments. We are thus led to a procedure that will define time boundaries and also prescribe a particular type of analysis to be performed between these time boundaries. The resulting epochs may not necessarily correspond one-to-one to the final sequence of linguistic elements. As an example, we might consider a word such as "back" spelled phonetically [b a k] that has been modified by tape cutting at the beginning and the end to remove all noise bursts related to the consonants. The resulting acoustical signal would contain only a vowel-like portion, and only two time boundaries would occur at the beginning and end of this epoch. However, if the tape cutting has not been too severe, a person would still perceive the entire word; hence, further analysis should determine from the transitions that the generating sequence of linguistic elements is more like three: consonant/vowel/consonant, rather than one vowel.

A segmentation procedure should also identify the significant controlled acoustical feature within the time boundaries. Rupert discusses how this reduces the variability induced by situation-dependent acoustical features. This would amount to attention focusing that includes as a special case formant tracking. By ignoring all but the distinctive controlled features, a large amount of noise rejection can be accomplished. Further segmentation need not be impaired by this attention focusing, because, as proposed by Rupert, it should be the precisely controlled features that govern the segmentation. However, the beginning of new features outside the area of attention must be able to "capture" the recognition choice so that a feature does not dominate long after it has ceased being significant.

The object of our segmentation procedure, to act as a direction for analysis, must then be able to isolate homogeneous epochs of signal, since in order to make reliable measurements we must have a tailored measurement algorithm (i.e., it is extremely difficult to track a formant during a fricative or noise-like portion of the acoustical signal, Thomas[19]). This suggests a representation of the acoustical waveform that shows isolated acoustical features and gives an adequate description of the signal properties so that segmentation and class identification can be performed.

The concept of homogeneous segments must be augmented somewhat because of the special nature of speech signals. In order to analyze a generalized acoustical signal generated by a complex scheme, as in human speech, one could use standard communication theory techniques of identifying a state model for each epoch (i.e., a set of differential

equations, n-degree polynomial fit, etc.) and then say the epoch has physical homogeneity as long as the model is valid. Then the switching times or segmentation points will correspond to changes in models. We must also consider linguistic homogeneity as discussed previously; there are several portions (acoustical features) of the total speech signal which are not linguistically significant. Therefore, the homogeneous property is with respect to both the physical measurements of the signal and the linguistic significance of those measurements.

I-D PREPROCESSING OF THE ACOUSTICAL SPEECH SIGNAL

Preprocessing of acoustical speech signals, when inspired by modern communication theory techniques, has been dictated more by what is available rather than by what is appropriate. Researchers have attempted to justify application of existing techniques by analogy with color (light frequency) perception (Yilmaz) or human perceptual experiments. The former approach can be though of as looking at the world through rose-colored (harmonic) glasses. The latter technique must be used with caution, since the capabilities of the human brain are not available in an ASR system.

The complicated nature of speech signals involves a predominant pitch frequency, which does not contain linguistic information (at a lower unit level), plus several components with time-varying frequencies. An acceptable analysis is possible but requires much computation (Schafer and Rabiner[22]). A real-time ASR system intended to make efficient use of a machine cannot afford this luxury. The problem involves more than waiting for a faster computer or a trickier algorithm when one wants to recognize connected speech from several speakers. In this section we will discuss the complicated nature of human speech signals and form a basis for specification of a preprocessing scheme tailored to the nature of ASR requirements.

The primary goal of preprocessing is to specify a transformation (filtering) which will: (1) remove noise (including other, confounding features of speech as discussed in the previous section); and (2) provide a compact representation of the salient features required for the recognition task. We cannot expect a straightforward application of standard

techniques based on homogeneous models* to achieve these goals. The generation of the acoustical speech signal is best modeled as a composite stochastic process (that is, a heterogeneous mixture of several interdependent time-varying systems). In addition, experiments measuring human perception of acoustical events indicate that man's ability to discriminate frequency is more acute than his perception of differences in intensity (Flanagan)[24]. We will show that the commonly used filtering techniques have poor frequency resolution, which adversely affects ASR system performance in natural human conversation.

If we assume that the signal is generated by a homogeneous process, the most efficient transformation would match this generation process, as attempted by Weiner-Hopf or Karhunen-Loeve[25] filtering. The difficulty (and success) in using these methods depends on the initial selection of the representation criterion and representation constraints.

The formation of the input signal minimizes, according to the chosen criterion, the differences between the output and an idealized signal. The criterion chosen has a considerable effect on the final form of the filter. There are many problems in which the mean squared error formulation is required in order to obtain any useful mathematical results. However, another criterion may be better suited to a particular estimation problem. For example, a filter designed for minimum mean squared error would be used successfully in the case of a stochastic signal (fricative), where the mean value and bandwidth of the frequency energy distribution are sufficient statistics. On the other hand, in the case of a vowel formant

---

* One characterization of a homogeneous process is a set of differential equations of a prescribed form with (time-varying) parameters and a fixed forcing function.

the peak of the frequency energy distribution is much more important than the mean value, necessitating a maximal likelihood criterion. Thus, even assuming that we can apply the more sophisticated techniques of communication theory to the speech preprocessing problem, we will generally need more than one "optimum" filter for a speech signal because of the changing nature of the speech acoustical signal.

The set of all possible inputs must be limited (by the filtering operation) in order to achieve rejection of noise and unwanted signals. This "allowable" subset is usually defined by a set of constraints (differential equation in the Kalman[26] formulation). Along with providing rejection capabilities, this would make the recognition problem easier by limiting the search space. However, the set of constraint equations, in order to be useful, must be a very accurate description of the instantaneous (rather than some average) "state" of the speech signal, implying that the classification must be known in advance in order to perform the preprocessing transformation. Halle[27] has proposed a feedback type ASR system (analysis by synthesis) to perform this circular classification. However, in view of the large number of computations implied by such a procedure and the previous discussion of the nature of the speech signal, we would propose the following: At the marking of a change in the speech signal decide which of several classes the new epoch belongs to and which "portion" of the total signal energy contains the significant information. Then, tailor a "filter" to this portion and perform the required transformation for as long as the desired features remain in the signal (determined by observing the results of the transformation).

We have already discussed how different criteria lead to several filters or transformations. Also, the parallel nature of the acoustic feature in a speech acoustical signal indicates multifiltering as a first step. We can summarize some of the requirements of a multifiltering pre-processing to remove noise and unwanted signals.

(1) Simile - Preservation of the necessary characteristics of a selected portion of the total acoustical signal. The subspace resulting from the filter transformation should, at this stage, preserve the input's characteristics (for instance, if the filter were a bandpass, time-invariant filter, this criterion would require preservation of the amplitude and phase relationships of the input within the 3 dB bandwidth of the filter).

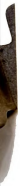
(2) Rejection - Removal of extraneous acoustical characteristics, including background noise and other speech features, such as other formants or the pitch component (for bandpass filters, this would require extremely good attentuation outside the 3 dB bandwidth).

(3) Continuity - At least one of the filters should contain a feature throughout its duration (for bandpass filters with a vowel glide of the second formant in the input signal that extends from 1400 Hz to 2800 Hz, at least one of the bandpass filters should have 3 dB bandwidth to encompass this range). This is desirable because we do not want artifact boundaries particular to a specific set of filters introduced when a feature transverses filter

boundaries. If this condition is not satisfied, a much more complicated decision network must be used to eliminate these artifact boundaries.

Further complications arise because of the wide frequency range, extending over many octaves, and the extreme variations in amplitude. Five contiguous 1/1 octave filters are required to cover the intelligible range of speech (one more if high-quality speech transmission is required), and the amplitude ranges over 120 dB with short-term variations on the order of 20-30 dB. One of the most popular instruments for displaying and representing speech signals is the sonagram, a 2-dimensional graphical display of frequency versus time, with intensity indicated by shading on the display. It has been shown that the sonagram is a physical approximation of the generalized sliding Fourier series (Lerner[28]), that is, a Fourier series computed over a time interval that is stepped along the acoustical signal. The difficulties in analyzing speech can be discussed in terms of the sliding Fourier series and the parameters involved. First, the length of the interval over which the series coefficients are computed must be greater than the period of the lowest frequency component of interest. Measurement of formant frequencies is further complicated during vowel-like portions by the pitch frequency (proportional to the repetition rate of the glottal pulses). The range of these pitch frequencies is from 80 to 400 Hz. The time period over which the Fourier series coefficients are computed must be greater than the pitch period (say two or three times the largest, $\sim$ 25-30 ms), or a great deal of variation will occur depending on the phase of the pitch frequency*. Thus, there is a lower bound on

* The ideal situation would be to synchronize the Fourier series computation period with the pitch periods. This requires a pitch detector and a device to decide on presence of pitch periods. The resulting frequency resolution is still on the order of the pitch frequency.



frequency resolution on the order of the pitch frequency. Sliding Fourier power spectra for both wideband (65-6500 Hz) and bandpass filtered vowel glides are shown in Appendix D. The irregular form of the spectra is due to the pitch component. Also, the high power of this component relative to higher frequency components (which carry the linguistic information) requires a significant dynamic range (50 dB is shown in Fig. D-1); even then, formant frequencies are difficult to identify. It would be expected that bandpass filtering should isolate these peaks, as is seen in Figure D-2. However, we should note that there are several problems that still are not solved:

(1) When two energy peaks are in the same filter, a decision must be made as to which peak corresponds to a formant and whether the other peak is simply a harmonic of the pitch frequency or a second formant. Ideally, it would be nice to treat one formant in every filter; however, this is overly optimistic.

(2) Measurement Resolution - This is possibly a special case of (1) in that the measurement scheme (sliding Fourier series, for instance) has a certain resolution; i.e., a certain minimum distance must be present between two peaks for them to be recognized as two separate peaks. The problem that can occur here is that different speakers may have different spacing, so that for one a two-formant "sound" may appear as a broad single peak while for the other the same "sound" will appear as two close narrow peaks.

(3) Frequency Glides (large values of derivatives of frequency) that move in and out of filters and across filter boundaries.

The ideal approach, of course, is to treat a feature as a continuous event, independent of the filter bandwidths, so that artifacts would not be introduced.

(4) Correlation of Formants in Adjoining Filters - Since the filters are overlapping, the formant could be present in two filters; important types of information may be found by comparing adjoining filters (Hanne[23]).

(5) Requirements (2) and (3) above are actually contradictory and cause, in the case of bandpass filters, a situation where in order to contain a formant glide within one filter, the bandwidth would be entirely too wide for adequate rejection and for emphasis of the many types of speech features encountered.

(6) The effects of the pitch component are not completely removed by 25-30 ms computations, time window tapering, or bandpass filtering as has been suggested by researchers.

These problems for bandpass filters, or, as has been shown by Schafer and Rabiner[22], for even more sophisticated types of frequency analysis, are caused by the inappropriate nature of any fixed-frequency type of analysis for speech processing. The criteria for using such analysis on (1) steady-state phenomena, such as constant vowels or nasals, (2) vowel glides (great changes in frequency of formants) and (3) noise-like signals, very quick , random transient-type phenomena, are in general quite incompatible. Further, it has been shown by Hanne that for several measurement schemes, the estimation of formant frequencies (natural modes) of the acoustical signal approaches a harmonic of the pitch frequency rather than the true value.

A recent article by Lecours and Sparkes[30] has indicated that narrow-band filters enhance the frequency pattern of vowels, whereas wideband filters more accurately show the transient time behavior of stop consonants (rapid envelope onset -- a fact well known to users of sonagraphs). Hanne has pursued this prefiltering idea further with a more sophisticated system of overlapping filters to estimate first formant frequencies within 3 percent. Flanagan's[23,24] study indicates that this approach is closer to the frequency estimation error in human recognition. Thomas[19] has also used wideband filters to emphasize frequency regions to show second-formant variations more clearly. Both Hanne and Thomas have argued that the effect of filtering speech signals can be predicted or inferred from usual steady-state filter analysis. However, Fig. 2 shows a sonagram of a common English word, indicating a frequency derivative on the order of 10,000 Hz per second. This high value of frequency derivative is known to give quite unpredictable and unexpected outputs from time-invariant linear filters (Baghdady[31], Wiener and Leone[32], Cannon and Duncan[33]). One should reexamine the criterion for filter bandwidth in terms of the time-varying properties that can occur in speech signals. The inverse relationship between rise time and bandwidth indicates that a fixed bandwidth bank of filters must be a compromise at best. The effect of an analysis period on the order of 25-30 ms is to average or smear quick transient phenomena. Discussion of recognition errors in various systems using this type of techniques (Reddy[7]) indicates that many consonants, especially stop consonants, are missed due to this smearing or averaging. The usual reason given for the recognition errors is the low energy and short duration of these speech sounds. One possible solution would be to vary the computing period inversely with frequency

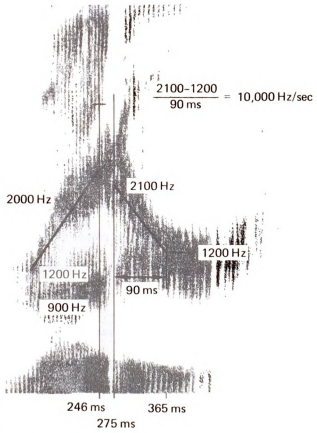


FIGURE 2 SONOGRAM OF ENGLISH WORD, RUDDER, SHOWING HIGH VALUE OF FREQUENCY DERIVATIVE

(long periods for low frequency and short periods for high). The resulting coefficients would not be for an orthogonal expansion, and also vastly differing waveforms can occur in the same frequency region.

Figure 3 shows both wideband (65 to 6500 Hz) and bandpassed time acoustical signals from recordings of four speakers saying medial [b] from [umbif] (see Appendix A for a description of the experimental pseudo-language used).

Bandpass filters can emphasize characteristics in the real-time waveform of extremely short transient-type bursts (release of the stop consonant [b] for different speakers, both male and female). Although the wideband waveforms (Figure 3, first page) show very little similarity, it is possible by bandpass filtering to find similar waveforms for the different speakers (Figure 4). The rejection of other features in the acoustical signal, as well as noise, by the filtering has made this possible. It will be noted that the most consistent and similar waveforms across speakers need not, and often do not, occur in the same frequency range (filter).

It has been argued by researchers that other acoustical clues for the perception of a stop consonant exist, namely the transition into the following vowel. Cooper et al.[5] investigated perception of synthetic initial vowels with frequency glide onsets. Öhman[16] has shown with sonagrams of actual speech that these results may not apply to connected human speech sounds. His data showed that, for medial stop consonants, the common notion of a formant hub does not hold; that is, there is no consistent point of origin for a given consonant, say [b], to which and from which vowel formants tend.

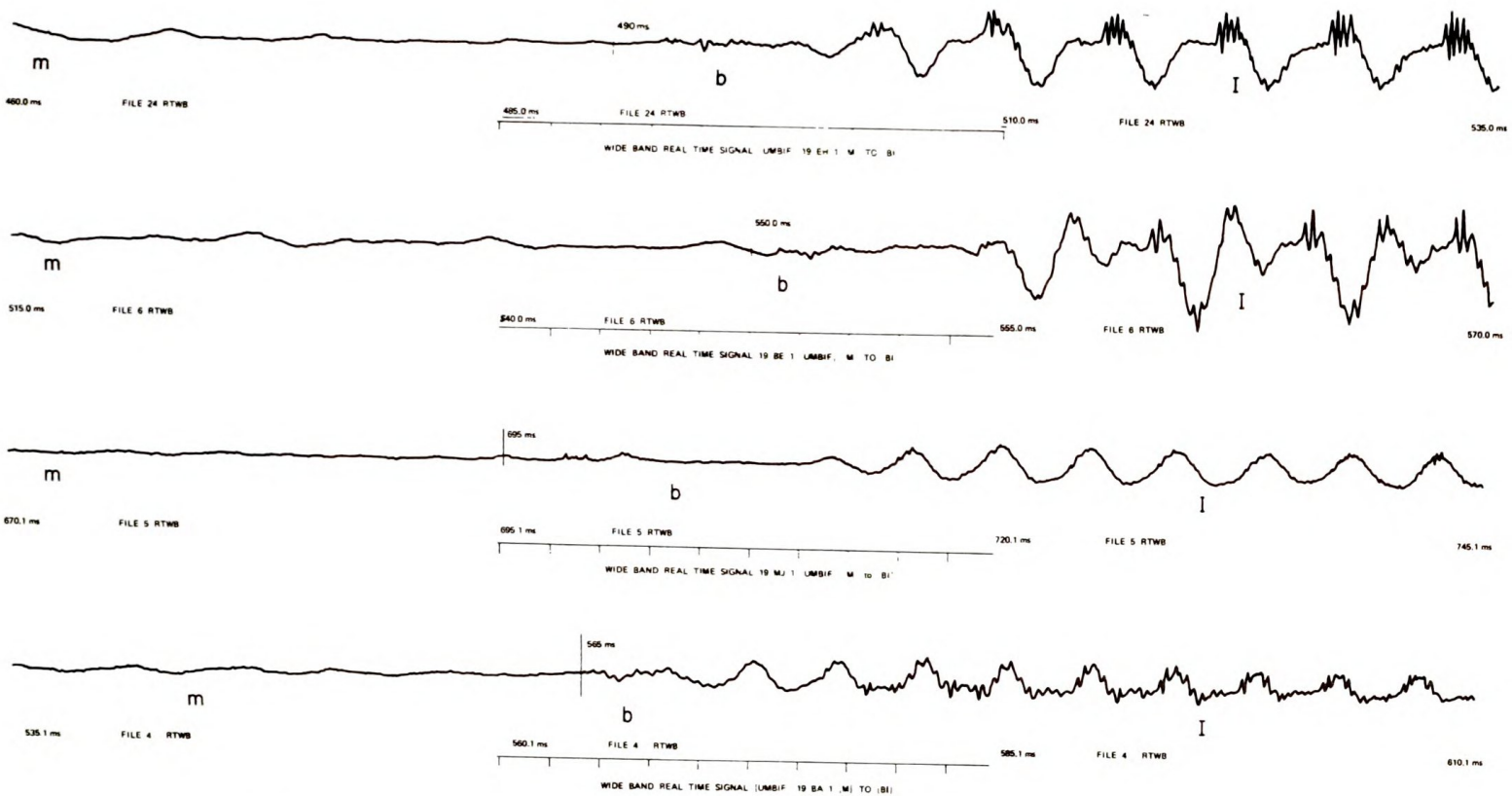
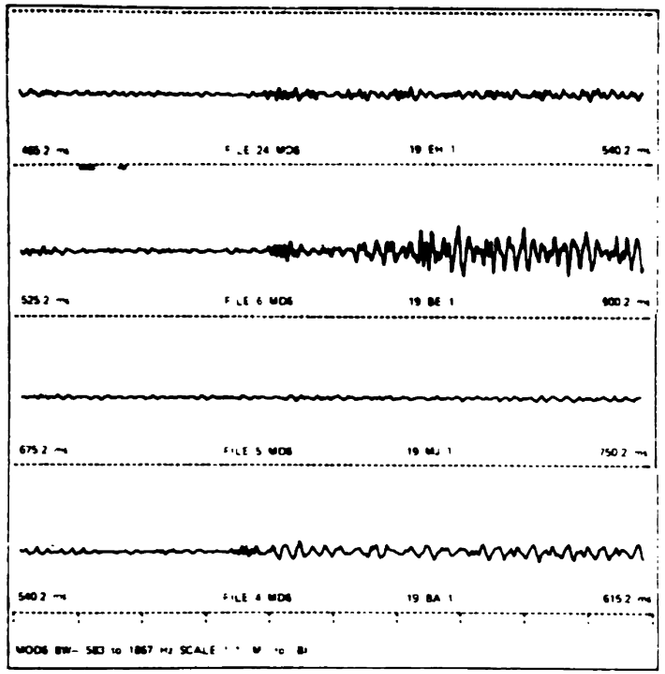
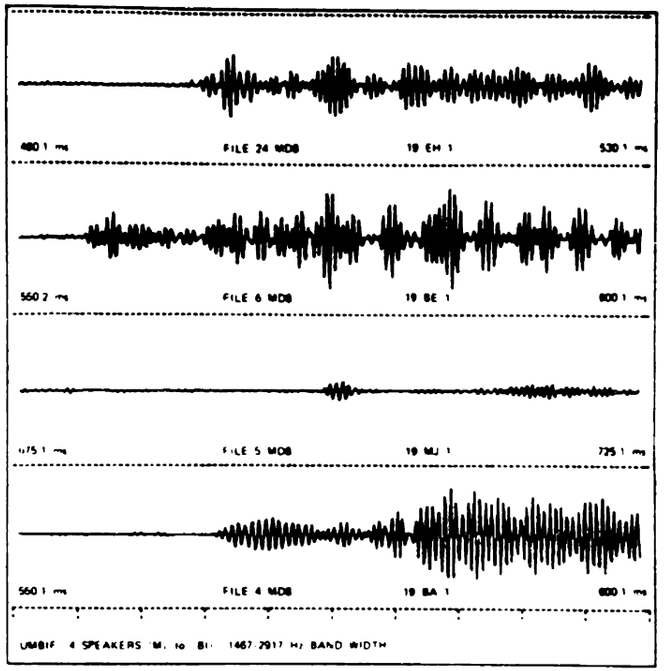
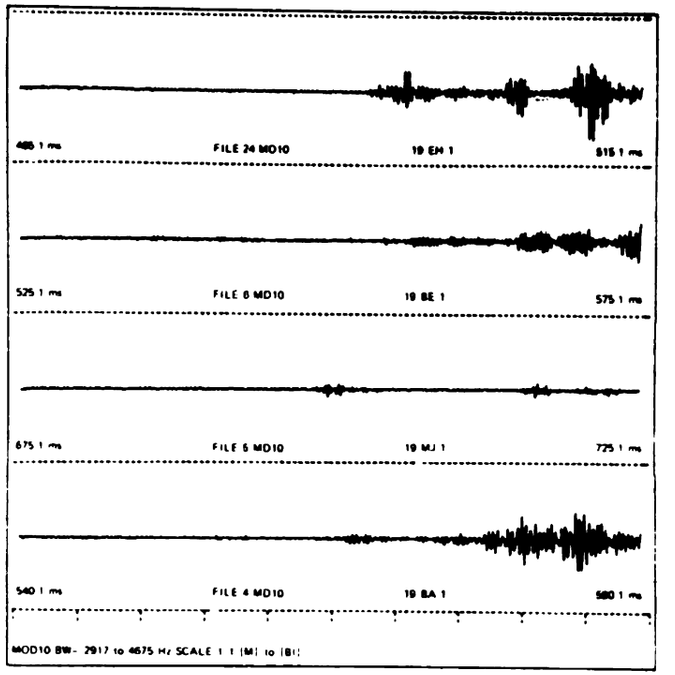
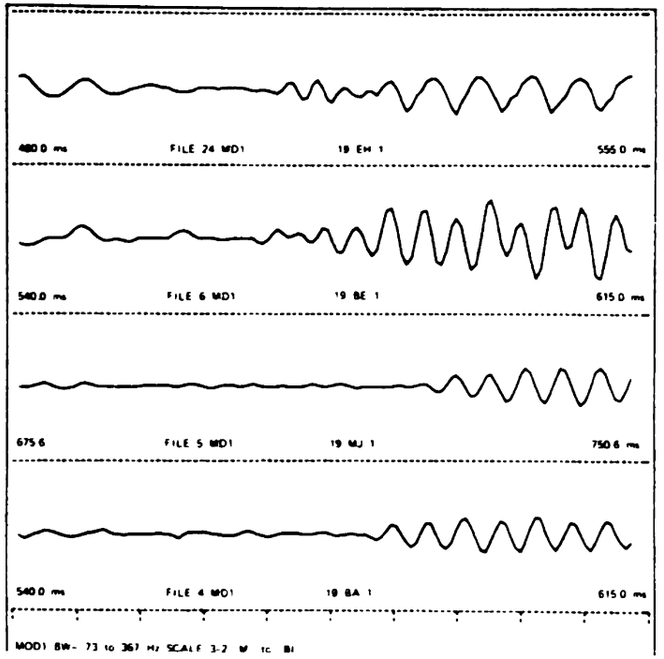


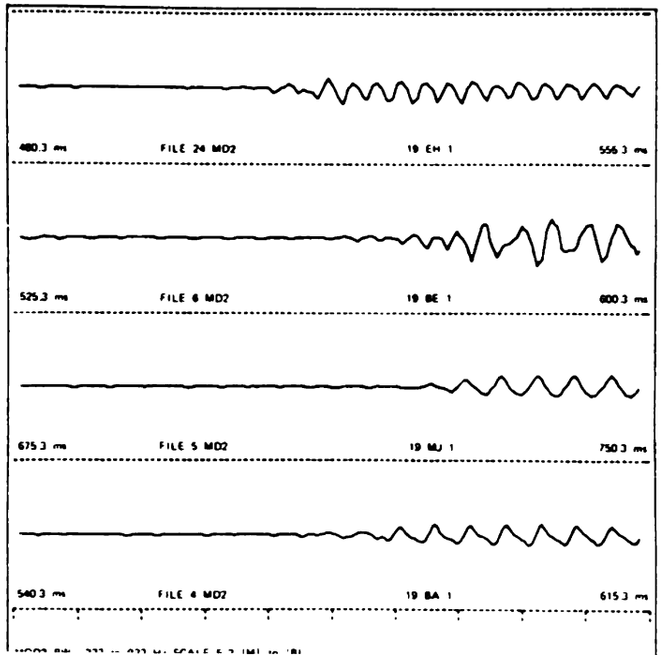
FIGURE 3 SPEECH ACOUSTICAL SIGNAL SHOWING SHORT TRANSIENT PHENOMENA

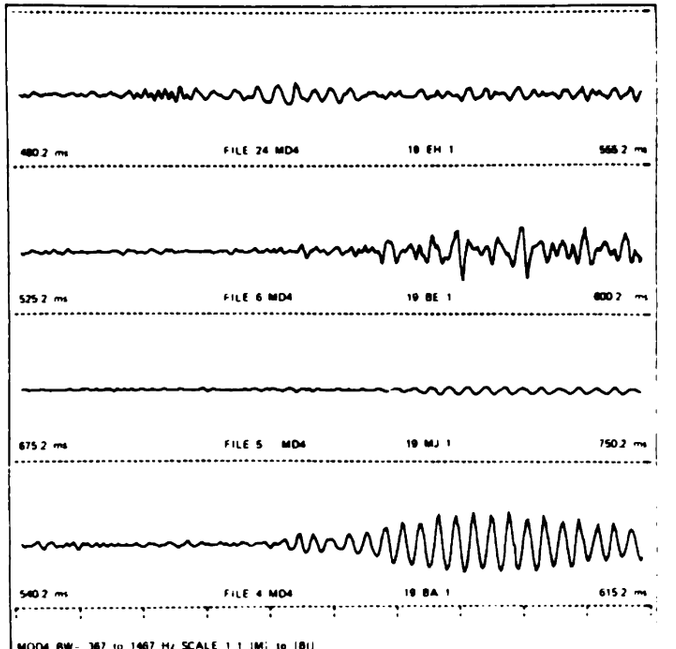

FILE 24 MD6
19 EH 1
FILE 6 MD6
19 BE 1
FILE 5 MD6
19 MJ 1
FILE 4 MD6
19 BA 1

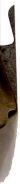

FILE 24 MD8
19 EH 1
FILE 6 MD8
19 BE 1
FILE 5 MD8
19 MJ 1
FILE 4 MD8
19 BA 1


FILE 24 MD10
19 EH 1
FILE 6 MD10
19 BE 1
FILE 5 MD10
19 MJ 1
FILE 4 MD10
19 BA 1


FILE 24 MD1
19 EH 1
FILE 6 MD1
19 BE 1
FILE 5 MD1
19 MJ 1
FILE 4 MD1
19 BA 1


FILE 24 MD2
19 EH 1
FILE 6 MD2
19 BE 1
FILE 5 MD2
19 MJ 1
FILE 4 MD2
19 BA 1


FILE 24 MD4
19 EH 1
FILE 6 MD4
19 BE 1
FILE 5 MD4
19 MJ 1
FILE 4 MD4
19 BA 1



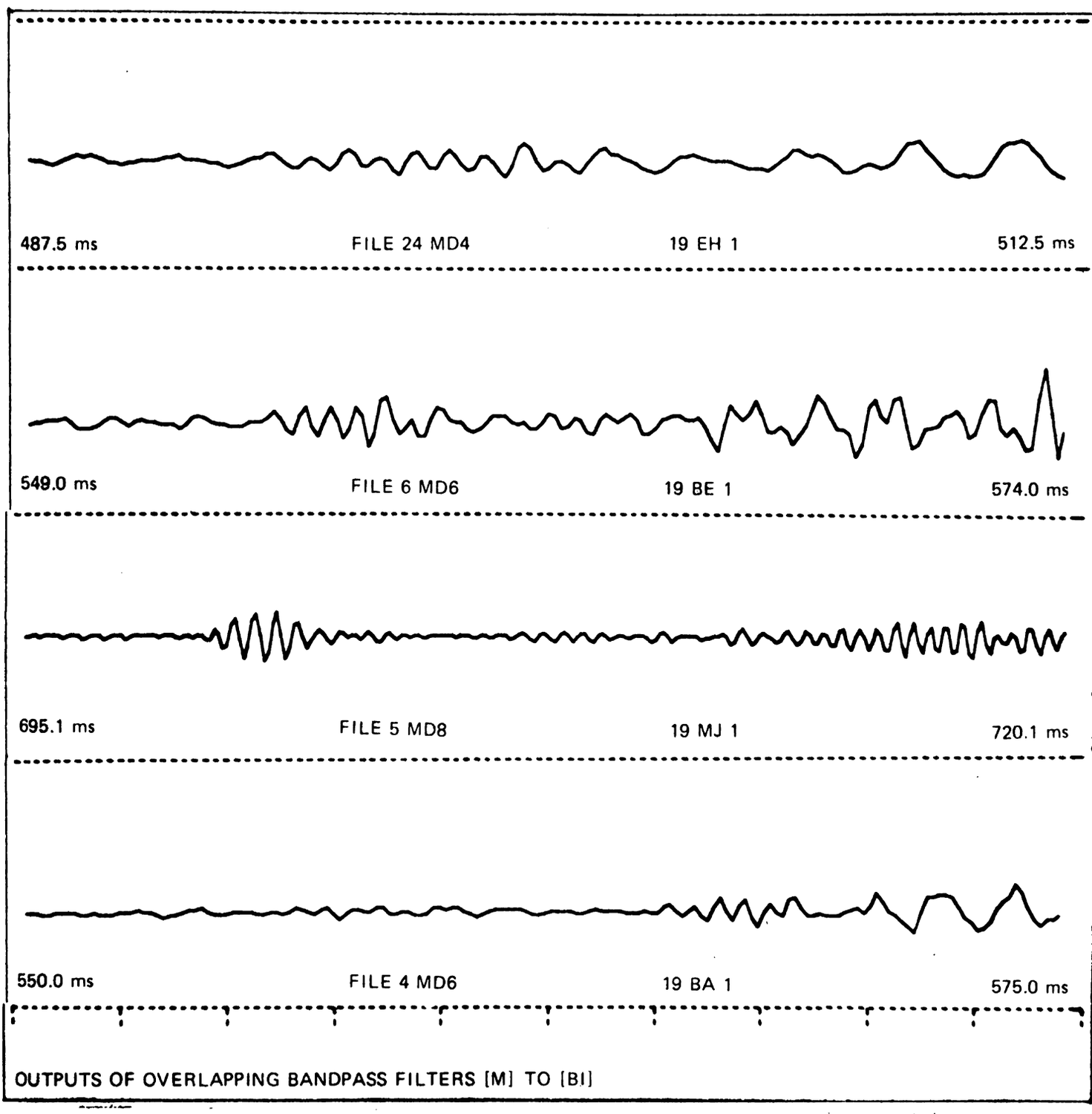


FIGURE 4 CONSISTENT TIME WAVEFORMS FOR SEVERAL SPEAKERS FROM DIFFERENT BANDPASS FILTERS

The choice of a set of filters for preprocessing the acoustical signal ranges from a set tailored to several classes of acoustical signals, possibly along with different representation criteria, to a set of contiguous narrow bandpass filters. The second approach has been extremely popular, especially for speech synthesis using vocoders. A familiar characteristic of narrowband filters, i.e., ringing, when excited with a sharp increase or decrease in amplitude or frequency is not consistent with the requirement of simile. For a period of time after a sudden change in amplitude or frequency, the output of the filter is not representative of the input. This problem will be discussed in the next chapter.

To avoid these difficulties, we have chosen a wideband (half-power bandwidth greater than one third the center frequency) overlapping filter bank (See Appendix B). The particular choice of the number of filters and the bandwidth of each filter was made in order to satisfy the three stated requirements. We shall see that the type of multiband filtering used here fulfills these requirements to a certain degree but has several limitations which must be corrected in the decision algorithm that follows the multiband filtering. The reason for these limitations is obvious. A time-invariant filter based on steady-state sinusoidal considerations obviously is not representative of the speech acoustical signal. However, there are several reasons for this choice over the admittedly better set of tailored filters. These reasons include:

(1) The hardware is readily accessible

(2) A large number of investigators have used wideband filtering in proposing and implementing preprocessing schemes, including Hanne[20], Reddy[7], Thomas[19], Gazdag[9], Shafer et al[22] and Yilmaz[6].

(3) Adequate representations have not been tailored to the time-varying acoustical signal.

(4) Few decision structures have been studied which are tailored to this type of multiband filtering preprocessing.

Another popular related analysis tool is the Fourier transform, especially since the introduction of the "Fast Fourier Transform" algorithm by Cooley and Tukey.[34] The Fourier transform equations can be modified so that each coefficient computation may be thought of as a (digital) filter operation. Hence, the complete transform computation may be considered a multi-bandpass filter processing.*

Much can be learned by considering a multi-bandpass filtering scheme with the intention of using it only as a first step and deriving from it further requirements for a tailored multi-filtering scheme.

A popular approach for parameterization of the filter outputs is to compute coefficients for an orthogonal series representation. However, the criterion commonly used for these computations is complete representation of the entire signal and minimization of the error between the orthogonal series and the original signal. This is not what is needed for an input to an ASR system. We would rather like to see only those parameters necessary for recognition. Flanagan[24] has modelled the speech-generation process as either a two-pole linear filter excited by the glottal pulses (for vowels and oral continuants) or a filter excited by white noise with variable bandwidth in the center frequency (for fricatives, stop consonants). The various parameters of input, envelope and

---

* In the next chapter we will see that the Fourier coefficient computation acts like a narrowband digital filter and hence is subject to "ringing."

filter bandwidth and center frequencies are considered to be time-varying. Thus, he would propose two parameters for each of our acoustical features, related to center frequency and bandwidth. Rupert has suggested that there are also consistent spectral shapes to the acoustical features which have been only slightly considered by previous investigators. These shape functions appear to be easily described by at most four parameters; say, the first four moments of the spectral density. They were first derived from sonagrams, but inspection of machine-calculated power spectra (Appendix D) shows that they may be more artifacts of the hardware than consistent features of speech acoustical signals. However, Sitton[45] has studied the first four moments of reciprocal zero-crossing distributions and found more consistent results.

Thus, one is led to different estimates of center frequency (and higher moments) for a narrowband (unimodal) spectral density. Zero-crossing counts immediately come to mind. There are many schemes and investigations of zero-crossings for analysis of speech signals (Cherry and Philips[36]). However, these measures were usually made on the total signal and, as can be seen by considering the sum of two sinusoids with variable amplitudes, the resulting output can be very difficult to interpret unless the signal has its spectral energy concentrated in a narrow frequency band. Thomas has used zero-crossing analysis on the output of his bandpass filter to estimate second formants; he finds an extremely good representation for vowels and indicates trouble only for very low power portions of the acoustical signal (fricatives, stop consonants). The use of bandpass filters followed by zero-crossing counts to estimate the frequency structure of formants has been demonstrated (Peterson[37], Hanne[29]).

Recently, Scarr[34] has discussed the fine structure of zero-crossings for speech-like signals having formants and pitch frequency components. He uses wide (1 octave) filters to isolate formants and shows the effect of pitch periods on formant frequency estimation. The errors involved in zero-crossing analysis are on the order of 1/number of zero crossings and therefore proportional to frequency. The case with Fourier series analysis is different, in that the frequency-location error is fixed at $\pm\frac{1}{2}$ the lowest frequency component (in this case, the pitch frequency).

Zero-crossing counts can be related to instantaneous frequencies (Baghdady[31], Lerner[28]) and thus incorporated into a discussion of quasi-stationary response of linear filters. However, few investigators have pursued this approach in the case of speech signals. Reddy[7] uses zero-crossing measures as an estimation of steady-state frequencies and also some envelope measurements (primarily relative envelope changes). We will discuss on a slightly more theoretical basis the relationships between zero-crossing measures and instantaneous envelope measurements in the next chapter. There are obvious benefits to be derived from the use of both derived time series in that the interpretation of zero-crossing counts is greatly enhanced by specification of the nature of the speech signal (i.e., if it is a vowel portion or a fricative portion, etc.), which can be determined by investigation of the envelope time series.

The subject we will investigate in the following chapters involves prefiltering by a bank of overlapping bandpass filters with the criterion that significant acoustical features appear in at least one of the filters over their duration. This presents a new type of recognition problem, involving the logic to decide which filter has the significant output and to perform a preliminary classification as discussed previously. This is the topic of the next section.

I-E DECOMPOSITION OF PATTERN RECOGNITION ALGORITHMS

The use of multiband overlapping filters to preprocess speech signals presents a specialized type of pattern-recognition processor. For the sake of clarity, we will adopt the widely used mathematical formulation in our discussion of this problem: The inputs to pattern-recognition devices are parameters, distinguishing characteristics of a physical event. A measurement is the numerical value of a parameter. A pattern vector, then, is an ordered set of measurements of a physical event; each measurement can be thought of as a component. The distance in pattern vector space between two vectors is a geometric measure of their closeness. A typical, but not always appropriate, distance is the standard Euclidean sum of squared differences of each component. A pattern-recognition algorithm is an assignment of class labels to the pattern vectors. In a typical pattern-recognition algorithm, each input pattern vector to be classified is compared with a number of reference vectors by a distance measure. The input vector is then assigned the label of that reference vector for which the distance is minimized. An ideal pattern-recognition algorithm would result in a dichotomization of the pattern vector space with unique class labels for each disjoint region. In the cases where this is not possible, the output of our pattern-recognition (PR) algorithm can be a degree of presence (DOP) vector, which has one component for each class label. The DOP vectors indicate the relative assignment for each class (say, normalized distances) and hence are a generalization of the single class label output.

A directed search is a special type of pattern-recognition algorithm that trades sequential operations for multidimensional single operations;

i.e., in the reference vector comparison case, a subset of reference vectors is selected by first examining few components and eliminating large portions of the pattern vector space from further search. Plasticity is a description of a particular type of pattern-recognition algorithm that allows changes in the pattern vector to DOP vector mapping, depending on a subset or all of the pattern vectors (the terms "learning" and "adapting" have been used for this process). A deterministic pattern-recognition algorithm is one which has no plasticity; that is, an a priori fixed mapping of vectors into classes, possibly by setting thresholds on measurements. Normalization is a process which we will distinguish from the pattern-recognition algorithm as being more concerned with the derivation of the parameter measurements. Although there are analogous types of standardization processes that do occur in pattern-recognition algorithms, it will facilitate the discussion to make this distinction.

We can now consider a schematic of the logic required for a pattern-recognition algorithm for our multi-bandpass filters and its operation.

In Figure 5, the output of each bandpass filter goes into a measurement device, producing an n-dimensional pattern vector for a time epoch (physical event) of the acoustical signal. These may be coefficients of an orthogonal expansion over a certain time interval, coefficients of a differential equation or another set of appropriate measurements (mean values, maximum value derivatives, maximum value standard deviations, etc.). For a continuous output of the bandpass filter, these types of measurement require time interval marks, which we will assume for now are generated elsewhere or are a part of the measurement scheme. The output DOP vector is of dimension r, the number of speech sound classes, discussed in

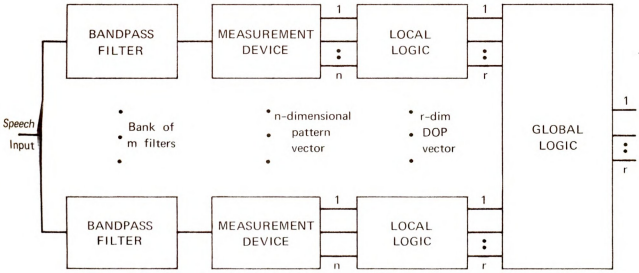


FIGURE 5 SCHEMATIC OF MULTIFILTER RECOGNITION LOGIC

Section D (on the order of 4-6). When referring to operations or properties of individual filter outputs, we will denote these as local, and when talking about properties of the entire bank of filters, we will denote these as global. By the particular choice of our filters, we see that a local property is one that is restricted to a certain frequency range. We will talk formally about "closeness" of pattern vectors in terms of clusters in the sense of Ball and Hall.[39] That is, we will say a set of pattern vectors is clustered if the intra-cluster distances are small (relative to a threshold, or to inter-cluster distances). The homogeneous property, which we introduced in our definition of acoustical segments, is with respect to both the physical measurements of the signal and the linguistic significance of these measurements. We might reformulate that property in terms of our definitions; physical measurements have some significance and consistency if they form a cluster (denoted as a physical cluster) in the pattern vector space. It may not always be the case that these physical clusters have linguistic significance. For example, a frequency measurement on a low-order filter primarily exhibits the pitch frequency. In this case, the physical clusters would correspond to different pitch frequencies and not to different linguistic events. At the opposite extreme, a physical cluster might be related to two distinct linguistic events, such as a medial [b], which has a very small amount of silence before the burst release, or a great amount of background noise such that it is difficult to distinguish from a fricative such as [f]. The resulting measurements for both the [b] and the [f] would tend to lie "close" to each other and, hence, lie in one physical cluster. Thus, the linguistic clusters would correspond to one physical cluster. At first, it appears that appropriate class labelling of the physical clusters

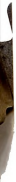
would define the linguistic significance; however, as indicated previously, the difficult task of assigning an exterior linguistic criterion to physical measurements subject to speaker, environment, and free phonetic variations will require a more sophisticated, plastic type of correspondence. The intention of keeping the actual decision algorithm simple so that it may be implemented in real time (with a minimal amount of computation) requires a better solution to the problem than simply keeping track of all the physical clusters and then making a correspondence to a set of linguistic labels. This type of approach requires, for example, storage of a large number of reference vectors (say, one for each physical cluster), comparison to these at each step of the decision algorithm, and a continual updating of these reference vectors due to slowly drifting measurements. In our problem, this approach is not feasible because of the variations due to different speakers. Bobrow and Klatt [13] have shown a decision algorithm (applied to the speech recognition problem) which is a directed search using decision-tree type logic that reduces the computational limitations (amount of storage, number of comparison speed of classification) of the usual multidimensional pattern recognition algorithm. Their procedure, applied to a speech measurement situation in which the variations discussed above are removed, would result in an effective ASR algorithm. Their technique, of course, will fail in the situation where a large number of reference vectors are saved for comparison.

The concept of precisely controlled features can be related here, also, to physical clusters, in that if other perturbing influences are removed, these precisely controlled features should result in "tight" physical clusters. This approach in itself should reduce considerably the amount of variation and hence the number of physical clusters needed

for description. This is then what we mean by attention focusing; i.e., the selection of a portion of the speech signal with precisely controlled features and tight physical clusters for further processing.

The complexity of the decision logic in Fig. 5 for an ASR system is dependent upon whether a decision for assigning a class label can be dichotomized into a number of local decisions followed by a global decision (analogous to the Zeiger[40] decomposition of automata), i.e., Is the dimensionality of the pattern vectors on the order of mxn or n (where m is the number of modules and n is the number of measurements in the input pattern vector for each module)? In the situation where two estimation criteria are appropriate (not necessarily simultaneously) for an n parameter problem, hence leading to two "filters", (as discussed in Section D), we would say the dimensionality is n rather than 2n, but "shifts" according to the input. The local decision would be based on "best" estimate according to the local criterion and the global decision would then be the choice of which estimator was most appropriate by examining the variance of the parameter estimator, for instance.

This variance measure of the estimation process can be generalized to handle the many more difficult and varied situations in ASR systems. We can also measure the quality of the DOP vector, e.g., the peakedness measure introduced by Kilmer et al[41]. A quality measure of the specific classification of an input pattern vector indicates the significance of the estimation of the measurements and consistency of the pattern vector w.r.t previous classifications.



Knowing that the complexity of PR algorithms goes up exponentially with the number of dimensions, a decomposition can result in real-time computations. The discussion of the previous sections indicates that this is the case for speech, in that the entire wideband acoustical signal is not precisely controlled and does not contribute in its entirety to the linguistic information. The choice of a logic structure, then, depends on this decomposition. We propose to show in Chapter IV that this is valid and indeed enhances the physical measurements in such a way as to reduce variations and improve the probability of success of classification.

Kilmer et al. have studied parallel recognition structures of the type shown in Fig. 6 and have demonstrated that an iterative nonlinear shakedown net (called S-RETIC)* is capable of arriving at a consensus of opinion among the local pattern-recognition elements (denoted modules), solving conflicts that may arise and selectively tuning to particular modules that have made a high-quality decision. We feel that this type of logic-structure is ideally adapted to the requirements of an ASR system. In particular, the bandpass overlapping filters have a mixture of correlation with neighboring filters and a high degree of local specificity because of the precisely-controlled features in speech signals (corresponding to the local redundancy of potential command concept of the S-RETIC). The parallel computations involving low-dimensionality (on the order of the dimensionality of each module) allow a minimal amount of computation.

*By S-RETIC, we mean the algorithm that performs the iterative nonlinear shakedown as described in Kilmer et al. (1967) and not the complete simulation study. Effectively, we denote S-RETIC for the computer program which corresponds to the B parts of the modules with their interconnections.

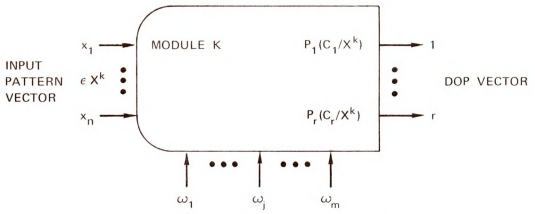


FIGURE 6 QUASI-STATISTICAL FORMULATION OF PR ALGORITHM

In order to get some feeling as to how S-RETIC arrives at its decision and also to consider an alternative procedure for using a number of pattern-recognition elements in unision, we can consider the probability distribution approximation techniques first discussed by Lewis[42] and Brown[43]. In order to apply their techniques to PR problems, we will consider each component of the DOP vector as being a conditional probability distribution $P_k(C_\ell/X_k)$, $\ell = 1, \ldots, r$, defined over the (module) input pattern vector space, X^k ($x^k \epsilon X^k$) for each class, C_ℓ (see Fig. 6). The DOP vector is computed from stored conditional distributions $P_k(X^k/C_\ell)$ for an input x^k by Bayes formula (assume $P(C_\ell) = 1/r$).

$$P_k(C_\ell/x^k) = P_k(x^k/C_\ell) / \sum_{\ell=1}^{r} P_k(x^k/C_\ell) \qquad \text{(I-E-1)}$$

The only requirements on the stored distributions is that they be non-negative for all C_ℓ, x^k and normalized such that

$$\sum_{X^k} P_k(x^k/C_\ell) = 1 \qquad \ell = 1, \ldots, r \qquad \text{(I-E-2)}$$

We can apply Lewis and Brown's techniques to $P_k(X^k/C_\ell)$, k=1, . . . m. for one class by considering each pattern-recognition module as computing a low order approximation to the true distribution. Chow[45] defines the structure of a pattern recognition algorithm as the function form of the probability distributions, particularly the condition dependencies among the components of the pattern vectors. He describes the Lewis-Brown approximation as structure adaptation. Hence, a parallel net of modules with lateral communication between local PR computations allows at least m different structures for each class. S-RETIC then selects the appropriate structure.

So far in our discussion, we have been considering decision structures that, except for the possibility of operating with minimal computations and less complexity, appear similar to those termed template-matching in Section B. This static type of pattern classification has little hope of working with connected conversational speech. The structure we are proposing has more flexibility built into it and operates like the PR algorithm we have described for isolated sounds where timing marks are well defined. The philosophy behind the design of the STL-RETIC program was to operate in an asynchronous manner, rolling over from one decision to another based on input changes. This structure is exactly the type that is needed for dynamic speech recognition; when one classification is chosen, such as silence preceding a word, and a new feature begins. It has been demonstrated by Kilmer that the change in the input (as reflected by the change in the local DOP vectors) is sufficient to cause a change in the overall global DOP vector. It will possibly be necessary to also determine changes in the input measurements. We propose to do this by detecting inherent changes in the physical characteristics of the signal and then deciding if these changes are significant enough to cause a recomputation of the global decision.

We will return to these questions in Chapter IV. First, however, we consider in Chapter II the nature of the acoustical waveform and discuss a procedure for detecting inherent changes in that waveform. In order to specify a training procedure for a plastic PR algorithm, an external classification criterion is needed. The lack of a one-to-one correspondence between acoustical and linguistic events rules out completely unsupervised learning. In Chapter III, structural linguistics is discussed in order to provide this criterion.

II REPRESENTATION OF TIME-VARYING SIGNALS

Representation of signals that result from transformations by a time-varying differential operator of standard signals present many difficulties, particularly to engineers with backgrounds in linear time-invariant differential operator analysis. Two representations are commonly used, the analytic signal[46] and the sliding Fourier transform methods.[22]

II-A Analytic Signals

The analytic signal representation is an attempt to define precisely the empirical notions of envelope and frequency. The primary advantage of this representation is that it separates the envelope and phase portions of the signal; in addition, the resulting spectrum is one sided (i.e., there is no mirror negative frequency portion). This corresponds to most spectra "pictures" and makes various moment calculations practical.

The spectrum of a real signal u(t) for t $\epsilon(-\infty,\infty)$ is the Fourier transform

$$U(j\omega) = \int_{-\infty}^{\infty} u(t)\, e^{j\omega t}\, dt \quad .^{*}$$

The Hilbert transform of the real signal x(t) defined on the interval $-\infty < t < \infty$ as the Cauchy principal value of the integral

$$x^{h}(t) \stackrel{d}{=} \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{x(\sigma)}{t-\sigma}\, d\sigma \quad , \qquad -\infty < t < \infty \qquad \text{(II-A-1)}$$

---

*We will adopt the convention of denoting the spectrum of a real function of time by capital letters.

is another useful transform. The new real signal $x^h(t)$ has the following properties (Titchmarsh[47]):

(1) $x(t) = \cos(\omega t + \varphi)$ $\qquad x^h(t) = \sin(\omega t + \varphi)$

(2) Under rather general conditions if $\mathbf{x} = y^h$; $x^h = -y$

(3) $X^h(f) = -jX(f); \quad f > 0$

$= \quad 0 \; ; \quad f = 0$

$= \; jX(f); \quad f < 0$

We can now define the analytic signal corresponding to x(t)

$$\hat{\mathbf{x}}(t) \stackrel{d}{=} \mathbf{x}(t) + jx^h(t) \qquad \text{(II-A-2a)}$$

$$= a(t)\, e^{j\alpha(t)} \qquad \text{(II-A-2b)}$$

where

$$a(t) = \sqrt{x^2(t) + x^{h^2}(t)} \qquad \text{(II-A-2c)}$$

$$\alpha(t) = \arctan\left\{x^h(t)/x(t)\right\} \qquad \text{(II-A-2d)}$$

The analytic signal $\hat{x}(t)$ has the one-sided spectra mentioned before, because of Property (3) and the definition. This signal is complex (the real portion is the original signal). Since the process of taking the real part of a complex function is a linear operation, it commutes with other linear operations such as convolution, differentiation, and integration.

Equation II-A-2b gives us an interpretation of the analytic signal representation as a phasor in the complex plane with time-varying magnitude and angle (with respect to the real axis). We may denote these quantities as the envelope and phase functions, terms motivated by the use of the analytic signal in various modulation studies (Baghdady[31], Weiner and Leon[32]). The instantaneous frequency is defined as the time derivative of the phase function

$$\omega_i(t) \triangleq d\alpha(t)/dt \qquad \text{(II-A-3)}$$

The analytic signal, although giving an instantaneous time description, can be used effectively for only a limited set of signals, namely those with slowly varying envelope and frequency functions. In order to enlarge this set of signals, we will introduce another definition which will be useful in discussing second-order time-varying differential operators. The derivative of an analytic signal may be written as a product of the analytic signal and a new signal, $b_x(t)$, which we will denote the prebandwidth signal.

$$\frac{d\hat{x}(t)}{dt} = \frac{d}{dt}\left\{a(t)e^{j\alpha(t)}\right\} = \left\{\frac{1}{a(t)}\frac{da(t)}{dt} + j\frac{d\alpha(t)}{dt}\right\}\hat{x}(t) \qquad \text{(II-A-4)}$$

where

$$b_x(t) \triangleq \frac{1}{a(t)}\frac{da(t)}{dt} + j\frac{d\alpha(t)}{dt}$$

The name of this function follows the convention of Deutsch* and Gabor's[46] definition of effective bandwidth. First shift the spectrum

---

*Deutsch[48] denotes $\hat{x}(t)$ as the pre-envelope signal because its magnitude is the envelope.

of $\hat{x}(t)$ to its center frequency. This frequency shift can be included in $b_x(t)$ by a property of Fourier transforms

$$X^s(j\omega) \triangleq X\{j(\omega + \omega_o)\} \leftrightarrow e^{-j\omega_o t}\,\hat{x}(t) = \hat{x}^s(t) \qquad \text{(II-A-5a)}$$

$$b_{x^s}(t) = \dot{a}(t)/a(t) + j(\dot{\alpha}(t) - \omega_o) \qquad \text{(II-A-5b)}$$

When ω_o is the center frequency of $X(j\omega)$, the complex portion of b_{x^s} reflects the time variations of instantaneous frequency about the mean. The effective bandwidth, BW, is the second moment of the spectrum about the mean.

$$BW^2 \triangleq \frac{\int_{-\infty}^{\infty} \omega^2 \left|X^s(\omega)\right|^2 d\omega}{\int_{-\infty}^{\infty} \left|X^s(\omega)\right|^2 d\omega} = \frac{\int_{-\infty}^{\infty} \left|\frac{d}{dt}\hat{x}^s(t)\right|^2 dt}{\int_{-\infty}^{\infty} \left|\hat{x}^s(t)\right|^2 dt}$$

$$= \frac{\int_{-\infty}^{\infty} \left|b_{x^s}(t)\,\hat{x}^s(t)\right|^2 dt}{\int_{-\infty}^{\infty} a^2(t)\,dt} \qquad \text{(II-A-6)}$$

The magnitude of b_{x^s} is an upper bound for the effective bandwidth by the Schwarz inequality and thus is a measure of an <u>instantaneous bandwidth</u>

$$BW^2 = \int_{-\infty}^{\infty} \left|b_{x^s}(t)\right|^2 a^2(t)\,dt \Big/ \int_{-\infty}^{\infty} a^2(t)\,dt$$

$$\leq \int_{-\infty}^{\infty} \left|b_{x^s}(t)\right|^2 dt \qquad \text{(II-A-7)}$$

Another interesting relationship between $b_x(t)$ and $\hat{x}(t)$ is (for $\dot{\hat{x}}(t) \neq 0$):

$$b_x(t) = \frac{d}{dt}\hat{x}(t) \Big/ \dot{\hat{x}}(t) = \frac{d}{dt}\{\log \hat{x}(t)\} \qquad \text{(II-A-8)}$$

In speech analysis, a logarithmic scale for amplitude (loudness) has often been used. By taking a derivative (with appropriate definitions for the complex logarithm) we can replace the transcendental function with a function more easily computed on a digital machine.

Now, consider a second-order time-varying linear differential equation (DE).

$$\ddot{\hat{x}} + a_1(t)\dot{\hat{x}} + a_2(t)\hat{x} = \hat{u}(t) \qquad \text{(II-A-9)}$$

where $a_1(t)$ and $a_2(t)$ are real functions denoting the time-varying parameters (for example, of a formant-producing cavity in speech generation). $\hat{u}(t)$ is an excitation function which may be stochastic (fricatives) or deterministic (glottal pulses). Introducing the prebandwidth function,

$$\left[\dot{b}_x(t) + b_x^2(t) + a_1(t)\, b_x(t) + a_2(t)\right]\hat{x}(t) = \hat{u}(t) \qquad \text{(II-A-10)}$$

The homogeneous solution of the reduced DE ($\hat{u}(t) = 0$) involves solution of a Ricatti equation for b_x, which can be solved if a_1 and a_2 are constant.

$$\dot{b}_x + b_x^2 + a_1 b_x + a_2 = \dot{b}_x + (b_x + (b_x + c_1)(b_x + c_1^*)) = 0$$

$$b_x = -c_1;\ -c_1^*$$

where

$$c_1 = \tfrac{1}{2}a_1 + \sqrt{a_1^2/4 - a_2} \qquad \text{(II-A-11a)}$$

$$c_1^* = \tfrac{1}{2}a_1 + \sqrt{a_1^2/4 - a_2} \qquad \text{(II-A-11b)}$$

c_1 and c_1^* are the pole locations for the time-invariant system given by Eqn. (II-A-10). When the constants c_1 and c_1^* are complex, the magnitude of b_{x_s}, the shifted prebandwidth function, has the damping factor $a_1/2$, which is an accepted "bandwidth" for this system. Thus, our definition is useful in relating bandwidth to a system that may have an infinite value for BW (this happens for certain values of a_1 and a_2).

When $a_1(t)$ and $a_2(t)$ vary slowly with time, so that $b_x \approx 0$, we can still define $c_1(t)$ and $c_1^*(t)$ by Eqn. (II-A-11) and we can define time-varying poles without Fourier transforms. In general, Eqn. (II-A-10) must be solved by numerical integration, but the function b_x is related to the crucial parameters of a system described by Eqn. (II-A-9) and can provide insight into the system's behavior. Analysis of higher-order time-varying systems by this approach is not as easy as the analysis of time-invariant systems, where reduction to second-order systems is achieved by partial fraction expansions. The lack of a superposition principle, plus the computational difficulty with sums of analytic functions, further complicates the generalization.

The analysis of the dynamic characteristics of one isolated formant is possible (and more tractable) with the introduction of the prebandwidth function. Real differential equations (DE) for the envelope and frequency functions can be derived by substituting the definition of b_x from Eqn. (II-A-4) into Eqn. (II-A-10) and separating the result into real and imaginary parts, giving

$$\left[\ddot{a}(t) + a_1(t)\dot{a}(t) + \left\{a_2(t) - \omega^2(t)\right\} a(t)\right]\left[\cos\left\{\alpha(t)\right\} - \sin\left\{\alpha(t)\right\}\right]$$

$$= g(t)\cos\{\gamma(t)\} \qquad \text{(II-A-12a)}$$

$$\left[\dot{\omega}(t) + 2\omega(t)\dot{a}(t)/a(t) + a_1(t)\omega(t)\right]a(t)\left[\cos\{\alpha(t)\} + \sin\{\alpha(t)\}\right]$$

$$= g(t)\sin\{\gamma(t)\} \qquad \text{(II-A-12b)}$$

where

$$\hat{x}(t) = a(t)e^{j\alpha(t)}$$

$$\hat{u}(t) = g(t)e^{j\gamma(t)}$$

$$\omega(t) = \dot{\alpha}(t)$$

The equation for the envelope (II-A-12a) is of the same form as the total signal DE with a "natural frequency" reduced by $\omega^2(t)$. The DE for the frequency is nonlinear in ω and a and shows the effect of damping on the natural frequency.

We can change (II-A-12a) by substituting for the second derivative of the envelope

$$\ddot{a}(t)/a(t) = \frac{d}{dt}\left\{\dot{a}(t)/a(t)\right\} + \left\{\dot{a}(t)/a(t)\right\}^2 .$$

Then we can rewrite (II-A-12) as

$$\frac{d}{dt}\left\{\dot{a}/a\right\} = g/a\left[\cos\gamma/\{\cos\alpha - \sin\alpha\}\right] + \omega^2 - a_2 - a_1\dot{a}/a - \left\{\dot{a}/a\right\}^2 \qquad \text{(II-A-13a)}$$

$$\frac{d}{dt}\{\omega\} = g/a\left[\sin\gamma/\{\cos\alpha + \sin\alpha\}\right] - 2\omega\dot{a}/a + a_1\omega \ . \qquad \text{(II-A-13b)}$$

If we identify ω and $\dot{a}/a$ as state variables, then Eqn. (II-A-13) is in the form of a nonlinear vector differential equation. For speech acoustical signal representation, these state variables are invariant to amplitude scale changes as seen from their differential equations; further, they form the real and imaginary part of the prebandwidth function.

As noted in Chapter 1, speech acoustical signals fall into a number of classes, depending on the values of the four signal parameters $a_1(t)$, $a_2(t)$, $g(t)$, and $\gamma(t)$ in our single formant model. Inspection of Eqn. (II-A-13) indicates that the derivatives of the two state variables depend only on the state variables and these four time-varying parameters. Thus, if we were to specify the two state variables and their derivatives as functions of time, we could perform the speech signal classification. This procedure does not require us to solve the complex nonlinear differential equations or to perform any type of matrix inversion that would be necessary to identify the time-varying model parameters.

When $\hat{u}(t)$ is a train of unipolar glottal pulses (each being 2 to 12 ms in duration), $\hat{u}(t)$ can be represented by the excitation envelope $g(t)$. For this situation, the sinusoidal oscillation terms can be removed from Eqn. (II-A-12). This is achieved by the physical process of envelope detection and lowpass filtering. In Section D, this filtering operation is investigated and a criterion for selecting the cutoff frequency is given to minimize distortion of the solution of the differential equation and maximize the smoothing of the oscillation terms.

When the excitation signal is stochastic, we cannot obviously reduce the complexity of the differential equation (i.e., $g(t)$ may

not adequately represent the total characteristics and $\gamma(t)$ may also be required to adequately describe the random fluctuations). Under certain conditions, it is possible to assume that the excitation function $\hat{u}(t)$ is a Gaussian random process with expected value of $0(E(\hat{u}) = 0)$ and has independent increments with a uniform energy versus frequency distribution (white noise). The differential operator described by Eqn. (II-A-10) will then specify an autocorrelation function for $\hat{x}(t)$. Kelly and Reed[49] show that the envelope and phase functions for $\hat{x}(t)$ and their derivatives have the following probability densities for each fixed t when $\hat{x}(t)$ is a stationary process.

$$p(a, \alpha, \dot{a}, \omega) = p(a)p(\alpha)p(\dot{a})p(\omega/a) \qquad \text{(II-A-14)}$$

where

$p(a) \rightarrow R(\sigma)$ Rayleigh with mean σ, $E\{x^2\} = \sigma^2$

$p(a) \rightarrow N(0, B_x^2)$ Normal with mean 0 and variance B_x^2.

$p(\alpha) \rightarrow U(0, 2\pi)$ Uniform between 0 and 2π

$p(\omega/a) \rightarrow N(\bar{\omega}, B_x^2/a^2)$

$\bar{\omega} \overset{d}{=} E\{|\omega|\}$

$B_x^2 \overset{d}{=} E\{x^2\}/[E\{\dot{x}^2\}] - \bar{\omega}^2$.

This indicates that the angle, envelope, and envelope derivative are statistically independent for each t (independent random variable). Thus, no information is lost by removing the oscillatory terms in Eqns. (II-A-12) and (II-A-13). For bandpass spectral densities (like those we are considering), where the energy is concentrated in a range $\Delta\omega$ about $\bar{\omega}$, the envelope and phase function energy distributions are concentrated in a similar range about $\omega = 0$ (Davenport and Root[25]). Also, the uniform distribution of the phase contains no parameters of the generating equations.

Abramson[50] has defined B_x^2 for stochastic processes as the mean square bandwdith. For ergodic stationary processes it is equal to the effective bandwidth, BW^2, given by Eqn. (II-A-6), which is applicable to deterministic processes. Thus the instantaneous bandwidth function, $b_x(t)$ is related to bandwidth measurements for deterministic and stochastic (stationary) processes. Further, for second-order differential operators, Eqn. (II-A-10), all the parameters of the process can be determined from first-order probability distributions (cf. Eqn. II-A-14). It is not necessary to estimate autocorrelation functions or spectral relationships between $b_x(t)$ and the parameters of the differential operator (Eqn. II-A-10). Since this operator determines the autocorrelation function, these remarks apply to nonstationary processes also.

Many speech sounds can be modeled by stochastic processes with stationary autocorrelation functions (giving time-invariant spectral densities). However, the short duration and low relative energy of these sounds does not allow a "steady-state" spectral density approach. Thus we must consider transient responses. In the next section we will discuss the problems of using spectral estimation techniques and the transient response of linear systems to envelope and frequency changes.

II-B. Sliding Fourier Series

The recent development of the Cooley-Tukey[34] algorithm for fast digital computation of Fourier series coefficients has caused much interest in Fourier frequency analysis. Modern communication literature uses "Fourier analysis" to refer to a particular use of any set of orthogonal functions to approximate a given signal by the following form:

$$f(t) \sim \sum_{k\geq 0} a_k \varphi_k(t) \qquad \text{(II-B-1)}$$

where the set of functions $\{\varphi_k(t)\}_{k\geq 0}$ is such that for some interval of time $[a,b]$ and some weight functions h(t) (definition of orthogonal functions)

$$\int_a^b h(t)\, \varphi_n(t)\, \varphi_m(t)\, dt = c_n^2 \delta_{nm} \qquad \big(h(t) > 0\big) \qquad \text{(II-B-2)}$$

where $\delta_{nm} = 1 \quad n=m$

$= 0 \quad$ otherwise;

the a_k's are constants.

For any N, and for any given finite energy function f(t), the integral weighted squared error defined by

$$\int_a^b h(t) \left| f(t) - \sum^{N} a_k \varphi_k(t) \right|^2 dt$$

is minimized by the constants

$$a_k = \int_a^b h(t) f(t) \varphi_k(t) dt \, . \qquad \text{(II-B-3)}$$

The most popular orthogonal set is the set of trigonometric functions, with $h(t) = 1$ over $[a,b]$. However, the trigonometric functions have finite energy only over finite intervals. Therefore, the class of functions we can represent by Eqn (II-B-1) with trigonometric functions must be non-zero only on a finite interval.

A finite energy representation over an infinite interval is achieved by defining the truncated time function

$$f_T(t) \stackrel{d}{=} f(t) \qquad -T/2 \le t \le T/2 \qquad \text{(II-B-4)}$$
$$\stackrel{d}{=} 0 \qquad \text{otherwise}$$

and then repeating $f_T(t)$ every T seconds.* A Fourier series of the form of Eqn. II-B-1 can be used, with

$$\varphi_{2k}(t) \stackrel{d}{=} \cos k\omega_o t$$
$$\varphi_{2k-1}(t) \stackrel{d}{=} \sin k\omega_o t \quad ; \quad \omega_o \stackrel{d}{=} 2\pi/T .$$

Some of the properties of the finite Fourier series are:

$$(1) \quad a_{2k} = \mathrm{Re}\int_{-T/2}^{T/2} f(t)\, e^{jk\omega_o t}\, dt$$
$$a_{2k-1} = -\mathrm{Im}\int_{-T/2}^{T/2} f(t)\, e^{jk\omega_o t}\, dt$$

$$(2) \quad f(t) \sim \sum_{k=0} \left\{a_{2k}\cos(k\omega_o t) + a_{2k-1}\sin k\omega_o t\right\} \quad ; \quad a_{-1} \stackrel{d}{=} 0$$
$$\sim \mathrm{Re}\sum c_k e^{j(k\omega_o t + \varphi_k)} = \sum c_k \cos\left\{k\omega_o t + \varphi_k\right\}$$

$$c_k \stackrel{d}{=} \sqrt{a_{2k}^2 + a_{2k-1}^2}$$

---

*This representation is a good approximation only over the interval $[-T/2, T/2]$.

$$\varphi_k \overset{d}{=} \arctan\left\{a_{2k-1}/a_{2k}\right\}$$

Notice in property (2) the resemblance to the form for analytic signals. The analytic signal corresponding to this series is*

$$f(t) \sim \sum_{k\geq 0} c_k e^{j\{k\omega_o t + \varphi_k\}} \qquad \text{(II-B-5)}$$

Now, consider some implications of these properties for time-varying signals, especially signals with varying frequency. Looking at Property (2) again, the series is a sum of cosine functions with constant amplitude and constant phase. (Guillemin[5.1]) states that the approximation of arbitrary functions by this type of series is due to constructive (and destructive) interference between sinusoidal functions of different frequencies. The natural association of the Fourier coefficients with a frequency distribution (analogous to Laplace and Fourier transform theory) causes some problems due to the interference phenomena. Figure 7 shows a particular waveform defined over a finite period $\left[T_a, T_b\right]$. The transform of y(t) (assume $T_a = \frac{1}{2}T_b$) is:

$$S_y = T_b e^{j f T_b/4} \frac{\sin\left\{2\pi(f-f_\ell)T_b/4\right\}}{2\pi(f-f_\ell)\,T_b/4} \qquad f_\ell = 1/T_\ell$$

---

*To put the series in true analytic form, Baghdady considers each term as a phasor and defines the amplitude and phase function for the resulting phasor sum, a construction that may have some intuitive appeal but is no help at all computationally.

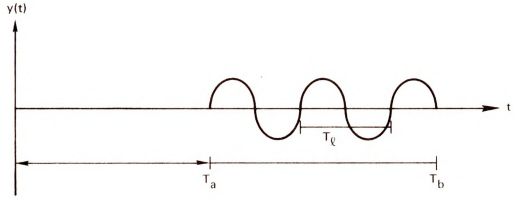


FIGURE 7 SHORT TRANSIENT PHENOMENON WHICH IS DIFFICULT TO ANALYZE WITH FOURIER SERIES

and indicates that Fourier coefficients computed over $[0,T]$, for $T=T_b$, would be significantly nonzero for several values of k other than $k_o = T/T_\ell$. The nonzero coefficients are necessary to cancel out $c_{k_o} \cos\{2\pi k_o t + \varphi_{k_o}\}$ over $0 \le t \le T/2$. The distribution of energy among the c_k's is misleading to an intuitive concept of frequency associated with y(t).

A remedy that has been suggested for these problems is to make T smaller (less than $T_b/2$) and compute a sliding Fourier series (i.e., starting the computation at increasing times). The resulting computations can be interpreted as "time-varying" c_k's and φ_k's. (However, this approach adversely affects the computational savings of the Cooley-Tukey method.) We may then ask if a representation of the form

$$\hat{x}(t) \sim \sum_{k\ge 0} c_k(t)\, e^{j\varphi_k(t)} \qquad \text{(II-B-6)}$$

would combine the properties of the analytic function and Fourier series. We can get some insight into the behavior of this series in the case when $\varphi_k(t) = \omega_k t + \theta_k$. The Fourier transform of $\hat{x}(t)$ in that case is

$$X(j\omega) \sim 2\pi \sum_{k\ge 0} e^{j\theta_k} C_k(\omega-\omega_k) \quad ; \qquad \omega > 0 \qquad \text{(II-B-7)}$$

$$\sim 0 \qquad\qquad \omega \le 0$$

where $C_k(\omega)$ is the Fourier transform of $c_k(t)$. Thus, the convolution sum in Eqn. (II-B-7) has smeared all the $c_k(t)$ functions together.

An example of a set of $c_k(t)$'s results from the "sliding" definition of Fourier coefficients,

$$c_k(t) = \int_{-\infty}^{t} \hat{x}(\sigma)\psi_k(t-\sigma)\,d\sigma \qquad \text{(II-B-8a)}$$

where $\psi_k(\sigma) = \varphi_k(\sigma)$, one of a set of orthogonal functions and the "duration" (non-zero time interval or effective time width) of ψ is much less than that of $\hat{x}$. In particular

$$c_k(t) = \frac{1}{T}\int_{-\infty}^{t} \hat{x}(\sigma)e^{j\omega_k(t-\sigma)}\,d\sigma \qquad \text{(II-B-8b)}$$

$$= e^{j\omega_k t}\int_{t-T}^{t} \hat{x}(\sigma)e^{-j\omega_k\sigma}\,d\sigma$$

We see that the calculation of sliding Fourier trigonometric coefficients can be interpreted as the output of the linear filter with input $\hat{x}(t)$ and impulse response

$$h(t) = \frac{1}{T}e^{j\omega_k t} \qquad 0\le t\le T \qquad \text{(II-B-9)}$$

$$= 0 \qquad \text{otherwise}$$

We might ask how $c_k(t)$ would look for various situations, especially for time-varying frequencies (as in speech formants, FM modulation systems, etc.). To answer that question precisely, we must develop some methods of looking at the response of linear filters to a general class of inputs. Before developing such a method, we might suggest what the $c_k(t)$'s should display.

Suppose the input $\hat{x}(t)$ is a constant amplitude sine function with a linearly varying frequency,* $\omega_i(t)$ ($\omega_i(t_k) = \omega_k$, k=1,2,3,4).

---

*We denote an instantaneous frequency function by $\omega_i(t)$ when it may be confused with values of frequency.

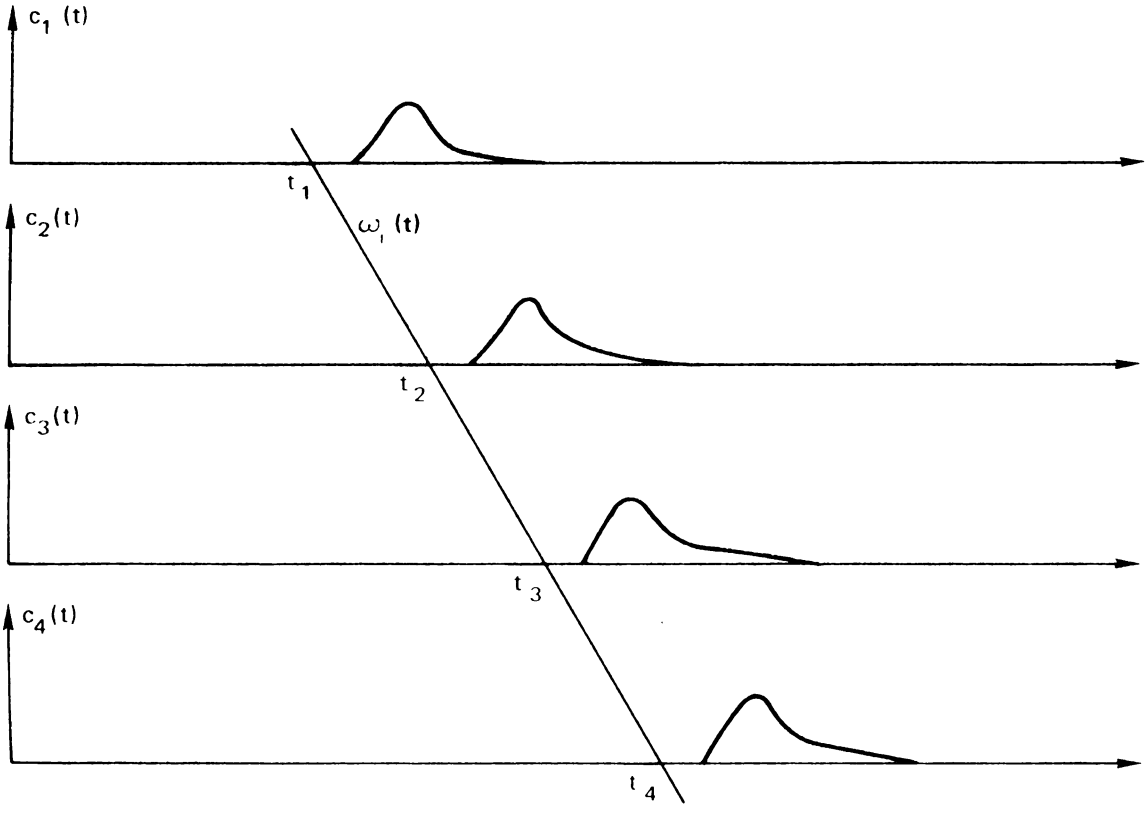


FIGURE 8 IDEALIZED FOURIER COEFFICIENT RESPONSE TO VARYING FREQUENCY INPUT

Then, each $c_k(t)$ corresponds to a frequency ω_k, k=1...4, which should ideally look like Figure 8. In the next section we show that this is possible only with restrictions which are too severe for the class of speech acoustical signals.

II-C Response of Linear Filters to Analytic Signals

When inputs to a linear filter (used to separate different formants in speech signals, say) contain amplitude and frequency derivatives of significant magnitude, the usual transform-superposition method of analysis becomes unwieldy, especially in determining transient response. Baghdady,[31] Leon and Weiner,[32] and Cannon[33] have suggested a different approach to this problem; they use the analytic signal and convolution integral to show the nature of the output of a linear filter in a more enlightening manner. Their approach is a generalization of standard sinusoidal analysis using Fourier series. If the input to a filter is a sinusoid that starts at t = 0

$$x(t) = ae^{[j\omega_o t]} \qquad t \geq 0 \qquad \text{(II-C-1a)}$$

$$= 0 \qquad t < 0$$

and the filter has Fourier transform $H(j\omega)$, which is rational, with simple poles at the point $s = s_1, s_2, \ldots, s_n$, then the output of the filter is

$$o(t) = aH(j\omega_o)e^{j\omega_o t} + a\sum_{k=1}^{n} A_k e^{s_k t} \qquad \text{(II-C-1b)}$$

with

$$A_k = \left[\frac{(s-s_k)\,H(s)}{s-j\omega_o}\right]_{s=s_k}$$

The first term in Eqn. (II-C-1b) is the steady-state or stationary solution, and the second is the transient term. The stationary solution is simply the input multiplied by the Fourier transform of the filter evaluated at the input frequency. When the input has a time-varying

amplitude and/or frequency, the form of Eqn. (II-C-1b) is duplicated by

$$o(t) = a(t)e^{j\alpha(t)} H(j\omega(t)) + \epsilon \qquad \text{(II-C-1c)}$$

where

$o(t)$	is the output of the filter
$a(t)e^{j\alpha(t)}$	is the analytic signal form of the input
$H(.)$	is the complex Fourier transform of the filter impulse response
$\omega(t)$	is the instantaneous frequency of the input
ϵ	is the transient or distortion term.

The first term, called the quasi-stationary term, is merely a complex number times the input, giving an amplitude and phase change. Thus, the idea of "frequency selection" by filtering has a definite meaning when ϵ is small compared to the quasi-stationary term. The transient or distortion term results from the filter's attempt to "follow" the changing input. Baghdady (and others) have bounded the distortion term and restricted the set of inputs to satisfy the bound in order to use the quasi-stationary term as an approximation to the output of the filter. The class of linear filters was limited in these studies to those described by rational functions of the frequency variable.

For the representation problems we are considering, this class of filters is not general enough (a "Fourier coefficient" filter is not of that type), nor do we have control over the class of inputs in the same manner. We will find the following definitions notationally (and, possibly, intuitively) convenient.

The Fourier transform pair for a real function h(t) is

$$H(j\gamma) \stackrel{d}{=} \int_{-\infty}^{\infty} h(t)\, e^{-j\omega t}\, dt \qquad \text{(II-C-2a)}$$

$$h(t) \stackrel{d}{=} \int_{-\infty}^{\infty} H(j\gamma)\, e^{-j\omega t}\, d\gamma \qquad \omega = 2\pi\gamma \qquad \text{(II-C-2b)}$$

Baghdady, Leon, and Cannon now define the quasi-stationary response of the filter as (for input instantaneous frequency, $\omega_i(t)$)

$$H(j\omega_i(t)) \stackrel{d}{=} \int_{-\infty}^{\infty} h(\sigma) e^{-j\omega_i(t)\sigma}\, d\sigma \qquad \text{(II-C-3)}$$

However, this is not a precise definition of a filter response to the instantaneous frequency unless the frequency changes slowly. Assume that h(t) is nonzero only over a finite interval $[0,T_h]$. Then, $\omega_i(t+\sigma)$ for $0 \le \sigma \le T_h$ is given (for ω_i analytic in $[0,T_h]$) by*

$$\omega_i(t+\sigma) = \omega_i(t) + \dot{\omega}_i(t)\sigma + \sum_{k\ge 2} \frac{\sigma^k}{k!}\, \omega_i^{(k)}(t)$$

and so a more exact definition results by using $\omega_i(t+\sigma)$.

$$H(j\omega_i(t)) \stackrel{d}{=} \int_0^{T_h} h(\sigma) e^{j\omega_i(t)\sigma} \left[\prod_{k\ge 1} e^{-\left(\frac{\sigma^{k+1}}{k!}\, \omega_i^{(k)}(t)\right)} \right] d\sigma \qquad \text{(II-C-3')}$$

This definition is unwieldy for situations with significant frequency derivatives, although it is more accurate than Eqn. (II-C-3). Of course, the two definitions are compatible if $T_h \dot{\omega}_i(t) \ll \omega_i(t)$.

---

*We use the notation $\dot{\omega}$ for the first derivative of ω with respect to its dependent variable and $\omega^{(k)}$ for higher derivatives.

Our approach will be to use Eqn. (II-C-3) as a definition, but with a generalized frequency term, i.e.

$$H(j\psi(t,t_o)) \stackrel{d}{=} \int_0^{T_h} h(\sigma) e^{j\psi(t,t_o)\sigma} \, d\sigma \qquad \text{(II-C-4)}$$

where

$$\psi(t,t_o) \stackrel{d}{=} f(\omega_i(t+t_o)) \qquad 0 \le t_o \le T_h \qquad t_o \text{ fixed}$$

We can illustrate by an example. The Fourier transformation of Eqn. (II-B-9) is:

$$H_k(j\omega) = \frac{1}{T} \int_0^T e^{j\omega_k(\sigma)} e^{-j\omega\sigma} \, d\sigma$$

$$= e^{-j(\omega-\omega_k)T/2} \left\{ \frac{\sin(\omega-\omega_k)T/2}{(\omega-\omega_k)T/2} \right\} \qquad \text{(II-C-5a)}$$

and for the frequency function $\psi(t,t_o)$

$$H_k(j\psi(t)) = e^{-j[\psi(t,t_o)-\omega_k]T/2} \left\{ \frac{\sin(\psi(t,t_o)-\omega_k)T/2}{(\psi(t,t_o)-\omega_k)T/2} \right\} \qquad \text{(II-C-5b)}$$

Thus, the coefficient c_k has a maximum value whenever $\psi(t,t_o) = \omega_k$ as we had shown in Figure 8.

The calculation of the distortion term will be facilitated by the definition

$$H(j\omega,t) \stackrel{d}{=} \int_{-\infty}^{t} h(\sigma) \, e^{-j\omega\sigma} \, d\sigma \qquad \text{(II-C-6a)}$$

$$= e^{-j\omega t} \int_{-\infty}^{t} h(\sigma) e^{j\omega(t-\sigma)} \, d\sigma \qquad \text{(II-C-6b)}$$

$$= \left\{ h(t) * e^{j\omega t} \right\} / e^{j\omega t} \qquad \text{(II-C-6c)}$$

Kharkevich[53] calls this function the "running spectrum." It can be shown that this definition introduces artifacts into the spectrum, although it does have the limiting property

$$(1) \qquad H(j\omega,\infty) = H(j\omega) = \int_{-\infty}^{\infty} e^{-j\omega\sigma} h(\sigma)d\sigma \quad .$$

Another property is

$$(2) \qquad \frac{d}{dt} H(j\omega,t) = h(t)\ e^{-j\omega t} \quad .$$

From Eqn. (II-C-6c) we can see that the running spectrum is a normalized transient response of a filter with impulse response h(t) to an analytic sinusoid signal.

The output of a filter with impulse response $h(\sigma)$ and input $\hat{x}(t) = a(t)e^{j\varphi(t)}$ is

$$o(t) = \int_{-\infty}^{t} a(t-\sigma)e^{j\alpha(t-\sigma)} h(\sigma)d\sigma \qquad \text{(II-C-7)}$$

We make the following assumptions

$$h(\sigma) = o; \qquad \sigma < o; \quad \sigma > T_h \qquad \text{(II-C-8a)}$$

$$\lim_{\epsilon \to 0} \int_{0}^{\epsilon} h(\sigma)e^{-j\omega\sigma} d\sigma = 0 \qquad \text{(II-C-8b)}$$

Equation (II-C-8a) is realistic, since most digital computer applications require this truncation. Equation (II-C-8b) simplifies the exposition by not allowing terms of the form $K_\infty\delta(t)$ in h(t). Using integration by parts, the output becomes

$$o(t) = \int_0^{T_h} a(t-\sigma)\, e^{j\alpha(t-\sigma)}\, h(\sigma)\, d\sigma$$

$$= \int_0^{T_h} a(t-\sigma)\, e^{j\alpha(t-\sigma)} \left\{ e^{j\psi\sigma}\, \frac{d}{d\sigma} H(j\psi,\sigma) \right\} d\sigma$$

$$= \left[a(t-\sigma)\, e^{j[\alpha(t-\sigma)+\psi\sigma]}\, H(j\psi,\sigma) \right]_0^{T_h}$$

$$- \int_0^{T_h} \left\{ \frac{\frac{d}{d\sigma} a(t-\sigma)}{a(t-\sigma)} + j\left[\frac{d}{d\sigma} \alpha(t-\sigma) - \psi \right] \right\} \hat{x}(t-\sigma)\, e^{j\psi\sigma}\, H(j\psi,\sigma)\, d\sigma$$

By the assumption in Eqn. (II-C-8b) the first term evaluated at zero is zero.

$$o(t) = a(t-T_h)\, e^{j[\alpha(t-T_h) + \psi T_h]}\, H(j\psi)$$

$$+ \int_0^{T_h} \{ b_x(t-\sigma) - j\psi \hat{x}(t-\sigma) \} \left[h(\sigma) * e^{j\psi\sigma} \right] d\sigma$$

$$= o_q(t) + o_d(t) \qquad \text{(II-C-9)}$$

We denote by $o_q(t)$ the quasi-stationary portion of the output transient response and by $o_d(t)$ the distortion term. The quasi-stationary term shows, explicitly, that the output is delayed from the input by an amount on the order of the interval over which $h(t) \neq 0$. A reference different than the one commonly used minimizes phase distortions occurring in $o_d(t)$ compared to use of the usual reference, t. The distortion term integrand is the prebandwidth function for the input times $\left[h(\sigma) * e^{j\psi\sigma} \right]$, a transient response term for the filter. For exponential filter functions (resulting from rational transfer functions), this term is

$$h(t) = e^{s_1 t}$$

$$h(t) * e^{j\psi t} = \frac{e^{j\psi t} - e^{s_1 t}}{(j\psi - s_1)} \qquad \text{(II-C-10)}$$

which corresponds to one factor in the distortion term in Cannon and Duncan's result when ψ is the instantaneous frequency.

The interpretation of $H(j\omega,\sigma)$ as a transient response (Eqn. II-C-6) shows us that the distortion term is a weighted average of the filter's ability to track frequency and amplitude changes. The term $\psi(t,t_o)$ is indicative, also, of the precautions necessary in interpreting the response. That is, for $\psi(t,t_o) = \omega_i(t) + t_o\dot{\omega}_i(t)$, we have a "pseudo-frequency," $t_o\dot{\omega}_i(t)$, biasing the instantaneous frequency $\omega_i(t)$. An attempt to include this bias in the distortion terms complicates the result tremendously. H(.) evaluated at the biased instantaneous frequency term is actually the predominant output when $t_o\dot{\omega}_i(t)$ is signficant. (See the following example.)

We could ask whether $t_o\dot{\omega}_i(t)$ is ever significant in the class of signals we wish to represent. Figure 2 (in Sec. I-D) shows a typical formant frequency transition from samples of the spoken word "rudder." This frequency transition has been inferred from a sonograph display. The range (over several speakers) of the frequency derivative $\omega_i(t)$ is from 5000 to 15000 Hz/sec or 5 to 15 Hz/msec. So, computation times on the order of 20 to 30 ms can have biases of ± 100 to 450 Hz. If we take an idealized "formant transition" of the form:

$$\hat{x}(t) = e^{j\varphi(t)} \qquad \text{(II-C-11)}$$

where $\frac{\dot{\varphi}(t)}{2\pi}$ = 2000 Hz $\qquad 0 \leq t \leq .020$

$= 2000 - 300[3(t/.030)^2 - 2(t/.030)^3] \qquad .020 \leq t \leq .050$

$= 1700$ Hz. $\qquad .050 \leq t \leq .070$

$\dot{\varphi}(t)$ gives a cubic transition from 2000 Hz to 1700 Hz with a maximum second derivative of 10,000 Hz/sec. (see Figure 9a).

Figure 9 compares the magnitude of the actual output, o(t), with the magnitude of the quasi-stationary term for five Fourier coefficient filters with 20 ms computing period. Also shown is a curve of the envelope maxima across the five filters. Figure 9a shows the quasi-stationary term evaluated at the input instantaneous frequency, $\psi = \varphi'(t)$. Figure 9b shows the quasi-stationary term evaluated at a biased instantaneous frequency.

$$\psi = \omega_i(t) + T_h/2\dot{\omega}_i(\beta) \quad , \qquad T_h = 20 \text{ ms} \qquad \text{(II-C-12)}$$
$$t - T_h \leq \beta \leq t$$

As is seen, this biased term gives a good correspondence between the quasi-stationary envelope maxima and the actual output envelope maxima. (Note that this delay distortion is <u>not</u> due to nonlinear delay versus frequency characteristics).

The implications of this analysis for the signals we are considering are obvious. Sliding Fourier spectra with computation periods on the order of 20 ms cannot adequately show frequency changes in the input without bias.

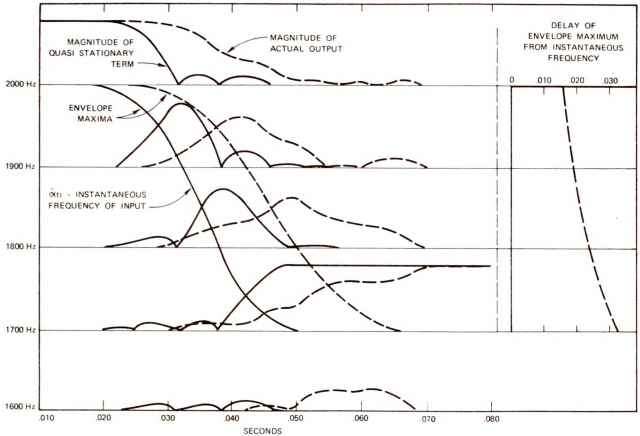


FIGURE 9 MAGNITUDE OF FOURIER COEFFICIENT OUTPUTS FOR TIME-VARYING FREQUENCY INPUT—ACTUAL AND QUASI-STATIONARY TERMS USING INSTANTANEOUS FREQUENCY OF INPUT

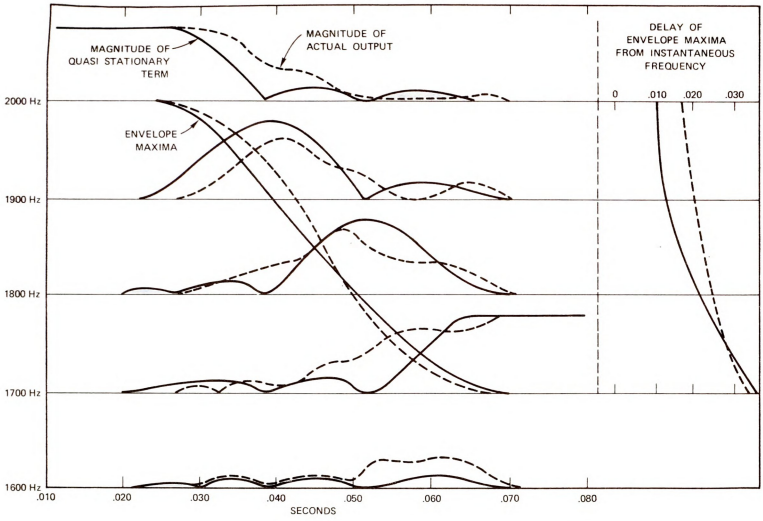
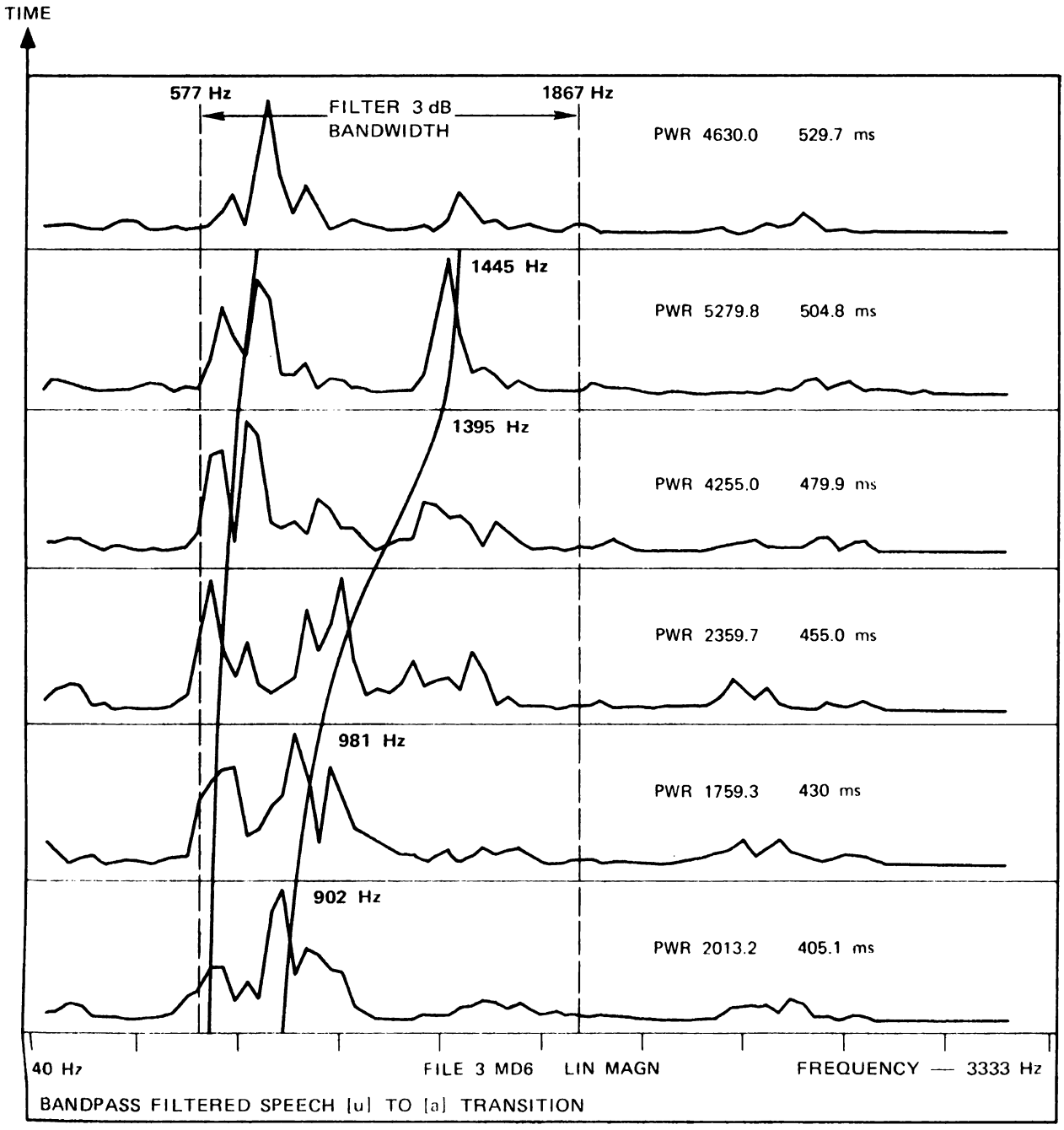


FIGURE 9 MAGNITUDE OF FOURIER COEFFICIENT OUTPUTS FOR TIME-VARYING FREQUENCY INPUT—ACTUAL AND QUASI-STATIONARY TERMS USING INSTANTANEOUS FREQUENCY OF INPUT (Concluded)

Another example of envelope delay changes due to changing frequency can be seen in Figure 10. Power spectra for a vowel transition of band-limited speech (module 6; bandpass, 577-1867 Hz) are shown. The vowel transition for a male speaker, from unstressed /u/ to stressed /a/, is also shown in Appendix D. The sliding spectra are computed every 15 ms over a 25 ms interval. Two vowel peaks are present in this filter, with one peak changing in center frequency from 902 Hz at 405.1 ms to 1445 Hz at 530 ms. The estimated frequency derivative at 480 ms is 9200 Hz/s. The absence of a significant second peak at 480 ms (relative to the lower peak) can be explained by envelope delay, which is caused by the bias frequency due to the great change in both frequency and envelope of the corresponding formant. These "holes" occur frequently in spectra of speech signals, as is noted by Schafer and Rabiner[22] and require complicated logic to avoid errors in formant peak tracking systems. The technique used by Schafer and Rabiner gives better frequency resolution at the cost of numerous computations (4 minutes on a GE-635 computer to compute two formants and pitch period for two seconds of speech). They first compute a cepstrum* to reduce the influence of the pitch frequency and then display the magnitude of the cepstrum transform along a spiral arc (see Figure 11) rather than along the unit circle. The spiral arc corresponds to a straight line in the s-plane. (Schafer and Rabiner call this procedure the chirp z-transform.) Improved frequency resolution results from passing close

---

*A cepstrum is computed taking the log of the Fourier transform. Two convolved time waveforms can then be separated if their frequency distributions are approximately disjoint.



(See Appendix D for description of labeling.)

FIGURE 10 FORMAT ENVELOPE AND FREQUENCY TRANSITION CAUSING DELAY DISTORTION

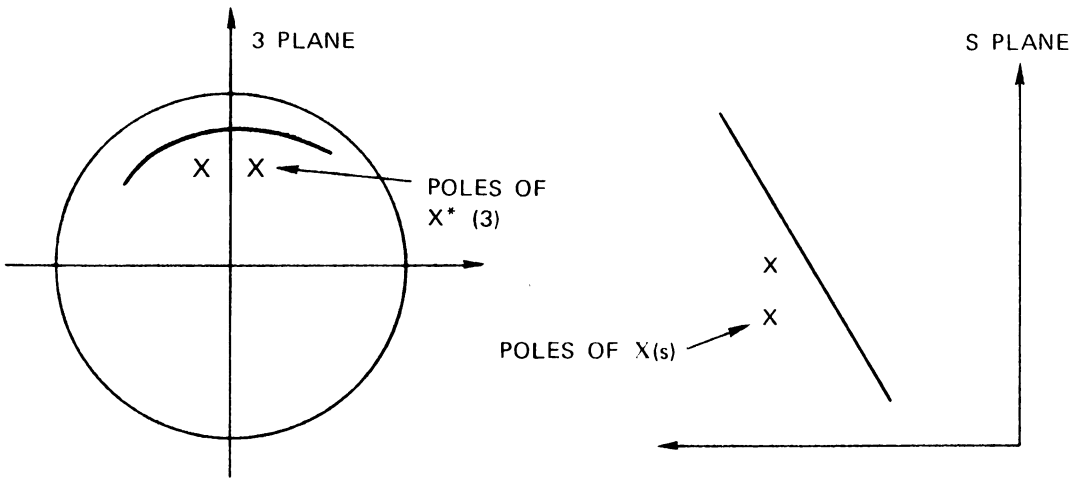


FIGURE 11 CORRESPONDENCE OF Z-PLANE SPIRALS AND S-PLANE LINES FOR THE CHIRP Z-TRANSFORM (From Rabiner, Schafer and Radar)

to the poles of $X(j\omega)$. We can express this concept in our formulation by letting ψ be a complex variable rather than purely imaginary. The choice of ψ to minimize $o_d(t)$ would be an approximation $b_x(t)$ throughout $[t-T_h, t]$. When $\hat{x}(t)$ is generated by a second-order linear time-invariant operator, the real part of b_x is the real part of the complex pole in $\hat{X}(j\omega)$.

We can see, then, how frequency resolution is enhanced, although envelope delay is still present since no provision was made for frequency derivative compensation. Schafer and Rabiner's techniques require a classification process to limit the input signals to male, nonnasal vowels and an iterative process to find the best spiral arc to achieve good discrimination. If the characteristics of the input were known, we could reduce the number of computations by appropriate choice of filter transfer function. This is the technique used in modern scan-frequency analysers where the phase of the IF filter transfer function is matched to the frequency derivative (scan rate) of the input. (Kincholoe[54]). If we knew $b_x(t)$ and the signal class, we could parallel Kincholoe's techniques by adjusting a time-varying filter to select a formant by center frequency tracking, minimize delay distortion by adjustment of the filter phase function to match the frequency derivative, and improve frequency discrimination (rejection of other formants) by bandwidth matching.* Such a scheme is shown in Figure 12.

The estimation of $b_x(t)$ requires classification of the input signal (as we discussed in Chapter I) and results in a "rough" initial estimate $\tilde{b}_x(t)$ which is used to generate a mixing signal $\exp\{-\tilde{b}_x(t)\}$.

---

*As noted by Kincholoe, the matched phase function would attenuate adjacent formants whose frequency derivatives are not matched in the same manner shown in Figure 9.

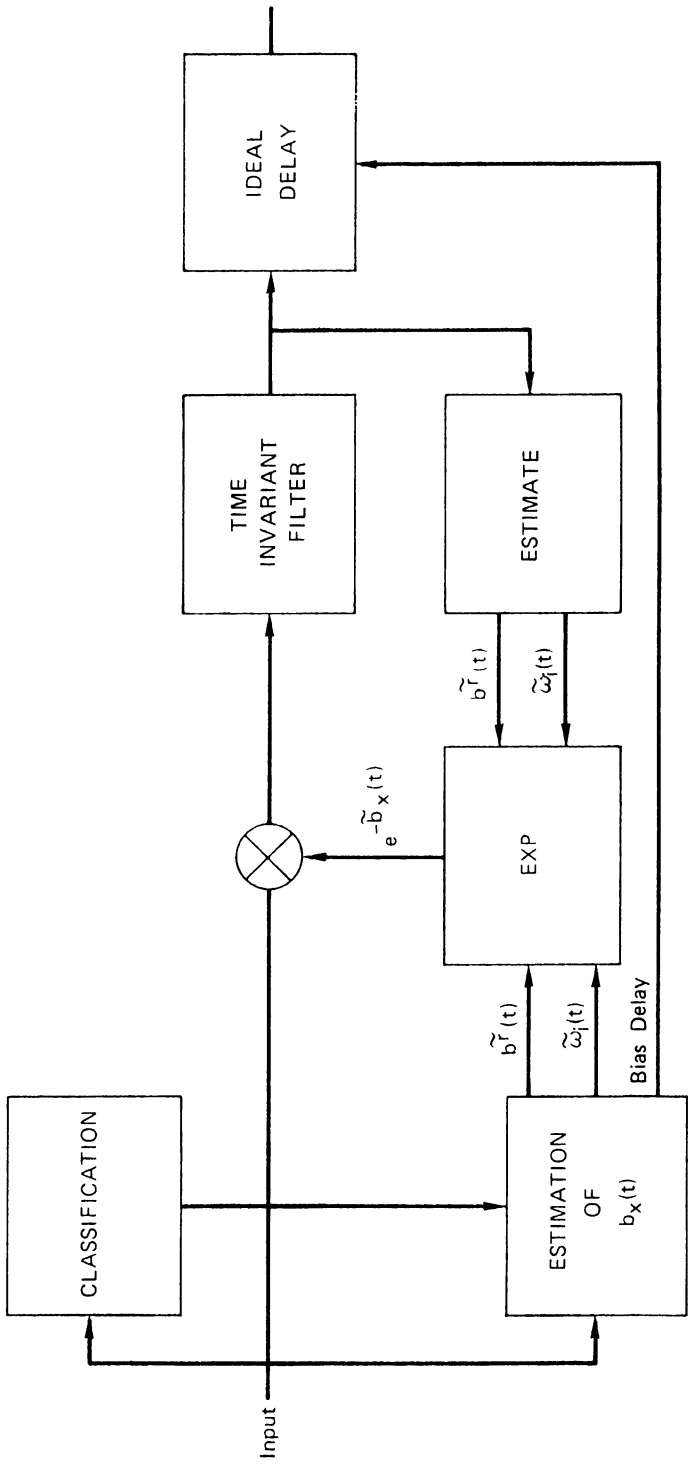


FIGURE 12 TIME-VARYING FILTER FOR FORMAT PARAMETER ESTIMATION

The filter can then be specified using standard Laplace transform tec niques where the dependent complex variable of the transfer function the difference between the mixing signal's complex "frequency" $\{\tilde{b}_x(t)$ and that of the input. The estimate is improved by a feedback loop. The delay distortion caused by frequency and amplitude changes is est mated and then corrected by a variable ideal delay. Equation (II-C-9 can be used to analyze the feedback loop, but it can also provide a synthesis procedure for a digital algorithm which significantly reduc the computations necessary to implement the scheme shown in Figure 12 Assuming that $\tilde{b}_x(t)$ is given, the majority of the computations are re to implement the filtering (mixing and delay require one operation, e per point of time).

There are two types of digital filter algorithms, transversa recursive. Transversal filters compute an output value from delayed values and are basically discrete convolutions (or correlations) of t form:

$$o_k = \sum_{j=b}^{N-1} a_j i_{k-j} \qquad k = 1, 2, \ldots \qquad \text{(II-}$$

The number of operations (one addition and one multiplication) per po of time is N.* Recursive filters compute an output value from delayed and output values. The algorithm is derived from the z transform of

*The Cooley-Tukey algorithm for computing Fourier coefficients is of form and for this special case requires only $\sim \log_r N$, where r is the greatest divisor of N.

filter time function.

$$\frac{O(z^{-1})}{I(z^{-1})} = \frac{P(z^{-1})}{Q(z^{-1})} = \frac{a_o + a_1 z^{-1} + \ldots + a_m z^{-m}}{1 + b_1 z^{-1} + \ldots + b_n z^{-n}}$$

$$O_k = \sum_{j=0}^{m} a_j i_{k-j} + \sum_{j=0}^{n} b_j O_{k-j} \qquad \text{(II-C-14)}$$

where $z^{-1} = e^{-j\Delta s}$ is an ideal delay of time Δ

m is the number of zeroes

n is the number of poles.

The number of operations per time point is m+n.

We can use the quasistationary term from Eqn. II-C-9 to approximate the filter operation in one operation per time point. The prefiltering classification and estimation of $b_x(t)$, along with feedback correction, allow this approximation to yield precise frequency tracking (the amplitude distortion is not relevant). The appropriate (narrowband) filter characteristics are stored by means of the complex transfer function H(·). The value of the input at each time instant is multiplied by the value of this function at the estimated bias frequency. This method combines the relatively low number of operations of the recursive filter with a desirable feature of the transversal filter. This feature is its ability to change the filter coefficients. If this is done with a recursive filter, an additional transient distortion is introduced. Thus we can achieve an approximate time-varying digital transfer function with a low number of operations, given an estimate of $b_x(t)$ and a classification of the input.

The classification system must be able to determine rough, but unbiased estimates of parameters of the incoming signal. The overlapping filter bank discussed in Chapter I can provide a basis for the estimation with some restrictions. In order to maintain simile* between the outputs and input of a filter (within the effective bandwidth), $T_h \dot{\omega}_i(t)$ must be less than the acceptable frequency resolution error. Thus a "worst case" bandwidth requirement can be derived which would introduce negligible frequency bias for all speech signals (although the bandwidth would be excessive for some).

Inspection of sonograms of English words spoken by several speakers indicates that the maximum value of $\dot{\omega}_i(t)$, 15,000 Hz/second, occurs frequently from 800 Hz up to 3,000 Hz. (Above this frequency it is hard to make reliable inferences.) Figure 13 shows bandwidth requirements for several percentage resolution errors. The bandwidth is determined from T_h (approx. rise-time) for linear-phase filters by the relation

$$B\tau \approx 1 \qquad \text{(II-C-13)}$$

where

$$B = \int_0^\infty A(\omega)\, d\omega / A(o)$$

$$\tau = \int_0^\infty h(s)\, ds / h(o) \approx .8T_h$$

$$A(\omega) = \left| k \int_{-\infty}^{\infty} h(t) e^{-j\omega t}\, dt \right|$$

B gives a measure of bandwidth that is approximately equal to the half power and effective bandwidths (for filters with very sharp rolloff like those we are using, this approximation is better). τ is a measure of

*Defined in Section I-D, p. 23.

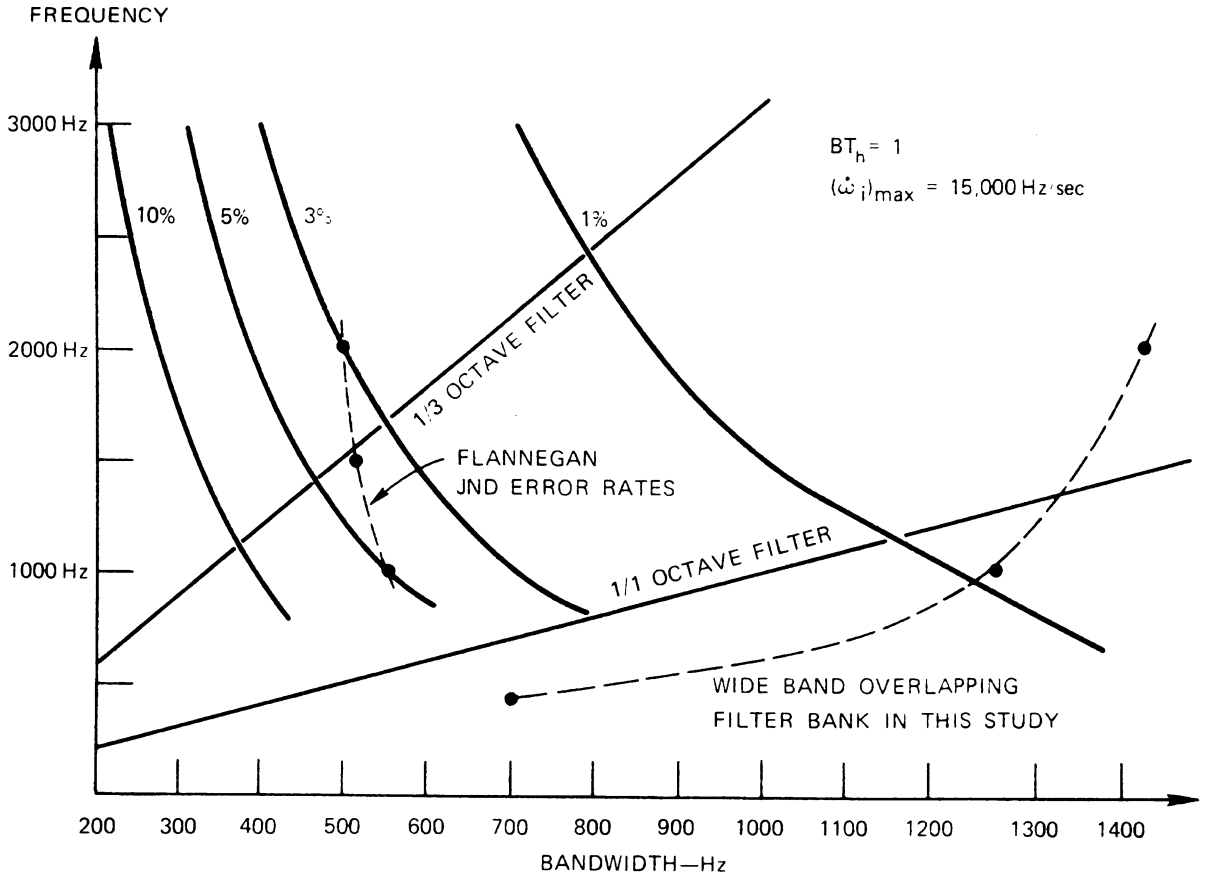


FIGURE 13 BANDWIDTH REQUIREMENTS FOR LARGE FREQUENCY REQUIREMENTS

rise time, usually between the 10 percent and 90 percent points on a step-response envelope curve. Figure 13 shows that our choice of bandwidths (Sec. I-D) is adequate in view of the inference from Flannagan's [23] data that a just noticeable difference in frequency for human experiments ranges from approximately 5 percent at 1000 Hz to approximately 3 percent at 2000 Hz. The data for this experiment results from individual variation of the first and second formant frequencies in a four-formant synthesized vowel.

In the next section we look at the outputs of such a filter bank and attempt to segment the speech signal into homogeneous epochs with center frequency and bandwidth as parameters.

II-D ESTIMATION AND SEGMENTATION OF INSTANTANEOUS SIGNAL PARAMETERS

The preceding section demonstrates how complex acoustical signals, such as those encountered in speech analysis, are represented most appropriately by instantaneous time functions related to the envelope, instantaneous frequency, and pre-bandwidth function. Differential equations for these functions have been derived for a single isolated formant. The bandpass pre-filtering that we have specified in Appendix B attempts to isolate formants. However, the inadequacies of fixed-frequency bandpass filters and the presence of inherent background noise in any realistic environment indicate that these differential equations will not be an exact representation. Therefore, a general form for these differential equations that can be expected to describe the signal parameters as seen on the outputs of our bandpass filters is more appropriate. In the following, we will denote the ratio $\dot{a}/a$ as b^r (the real part of b_x).

$$\frac{d}{dt} b^r = f_1(b^r, \omega, g/a, \gamma, \eta_1, t) \qquad \text{(II-D-1a)}$$

$$\frac{d\omega}{dt} = f_2(b^r, \omega, g/a, \gamma, \eta_2, t) \qquad \text{(II-D-1b)}$$

where f_1 and f_2 are nonlinear time-varying functions for the derivatives of the state variables. η_1 and η_2 are stochastic processes which represent the unwanted signals and other noise.

The classical theorems on "best" estimators deal with asymptotic properties as the number of samples becomes large. These results are of

II-D ESTIMATION AND SEGMENTATION OF INSTANTANEOUS SIGNAL PARAMETERS

The preceding section demonstrates how complex acoustical signals, such as those encountered in speech analysis, are represented most appropriately by instantaneous time functions related to the envelope, instantaneous frequency, and pre-bandwidth function. Differential equations for these functions have been derived for a single isolated formant. The bandpass pre-filtering that we have specified in Appendix B attempts to isolate formants. However, the inadequacies of fixed-frequency bandpass filters and the presence of inherent background noise in any realistic environment indicate that these differential equations will not be an exact representation. Therefore, a general form for these differential equations that can be expected to describe the signal parameters as seen on the outputs of our bandpass filters is more appropriate. In the following, we will denote the ratio $\dot{a}/a$ as b^r (the real part of b_x).

$$\frac{d}{dt} b^r = f_1(b^r, \omega, g/a, \gamma, \eta_1, t) \qquad \text{(II-D-1a)}$$

$$\frac{d\omega}{dt} = f_2(b^r, \omega, g/a, \gamma, \eta_2, t) \qquad \text{(II-D 1b)}$$

where f_1 and f_2 are nonlinear time-varying functions for the derivatives of the state variables. η_1 and η_2 are stochastic processes which represent the unwanted signals and other noise.

The classical theorems on "best" estimators deal with asymptotic properties as the number of samples becomes large. These results are of

little help in estimating instantaneous values. A multiple regression analysis would fit a polynomial of specified degree to the observations over a fixed interval. However, this method requires a priori knowledge that is not available (maximum degree of the polynomial and a fixed interval for fit) and much computation (usually a matrix inversion (Donahue[71]). Thus, pointwise estimates are required.

For time-invariant differential operators with either stochastic* or deterministic excitation, the two common parameters are mean frequency ($\bar{\omega}$) and bandwidth (BW^2). The mean frequency for analytic signals is well defined in terms of the spectrum $\hat{X}(\omega)$. We can derive a formula in terms of the time functions $a(t)$, $\dot{a}(t)$ and $\omega(t)$.

$$\bar{\omega} = \frac{\int_{-\infty}^{\infty} \omega X(\omega)\, X^*(\omega)\, d\omega}{\int_{-\infty}^{\infty} a^2(t)dt} = \frac{1}{j}\int_0^{\infty} \hat{x}^* \frac{d}{dt}\hat{x}(t)\, dt \Big/ \int_{-\infty}^{\infty} a^2(t)dt$$

$$= \frac{1}{j}\left[\int_0^{\infty}\left[\frac{\dot{a}(t)}{a(t)} + j\omega(t)\right]|\hat{x}(t)|^2 dt + \int_0^{\infty} x^*(t)\frac{d}{dt}\hat{x}(o^+)dt\right] \Big/ \int_{-\infty}^{\infty} a^2(t)dt$$

where the second integral is due to a step discontinuity at the origin.

$$\bar{\omega} = \left[\int_0^{\infty} a^2(t)\omega(t)dt + \frac{1}{2\pi j}\left[\int_0^{\infty}\dot{a}(t)a(t)dt + \frac{a^2(o^+)}{2}\right]\right] \Big/ \int_{-\infty}^{\infty} a^2(t)dt.$$

---

*Because the process is ergodic we will use time averages rather than expectations.

Since we assume that $\hat{x}$ is a well-behaved, finite energy function, $a^2(\infty) = 0$.

$$\bar{\omega} = \int_0^\infty a^2(t)\omega(t)dt \Big/ \int_{-\infty}^{\infty} a^2(t)dt \qquad \text{(II-D-2a)}$$

The effective bandwidth can be converted to a similar form (from II-A-7)).

$$BW^2 = \frac{\int_0^\infty |b_{x^s}(t)|^2 a^2(t)dt}{\int_0^\infty a^2(t)dt}$$

$$BW^2 = \frac{\int_0^\infty b^{r^2}(t)\ a^2(t)dt}{\int_0^\infty a^2(t)dt} + \frac{\int_0^\infty (\omega(t)-\bar{\omega})^2 a^2(t)dt}{\int_0^\infty a^2(t)dt} \qquad \text{(II-D-2b)}$$

Thus for constant coefficient operators we have weighted time average formulas for intuitive parameters. For time-varying operators, we are not so fortunate. In order to derive formulas we need an assumption that is often true for physical systems. We call a process locally ergodic if we may reasonably approximate ensemble averages by sliding time averages, i.e.

$$E\{c(t)\} \approx \frac{1}{\tau}\int_{t-\tau}^{t} c(\sigma)d\sigma \qquad \text{(II-D-3)}$$

Basically the assumption is that the time behavior of the parameter c(t) is "smooth" with respect to the statistical variations. This procedure is incorporated in many engineering systems, and we are merely recognizing this often-invoked assumption explicitly. The determination of τ is a key to this approach and depends on the nature of the processes. We will discuss its choice later.

Equation (II-D-3) can now be rewritten to give averaging equations for time-varying operators.

$$\bar{\omega}(t) \approx \int_{t-\tau}^{t} a^2(t)\omega(t)dt \Big/ \int_{t-\tau}^{t} a^2(t)dt \qquad \text{(II-D-4a)}$$

$$BW^2(t) \approx b^2_{x^S}(t) = \frac{\int_{t-\tau}^{t} b^{r^2}(t)a^2(t)dt}{\int_{t-\tau}^{t} a^2(t)dt} + \frac{\int_{t-\tau}^{t} (\omega(t)-\bar{\omega}(t))^2 a^2(t)dt}{\int_{t-\tau}^{t} a^2(t)dt} \qquad \text{(II-D-4b)}$$

Notice that (II-D-4b) gives a sliding time average of $b_{x^S}(t)$ and hence the time average BW(t) is denoted $b_{x^S}(t)$. In Appendix E, the relationship between sliding standard deviations and derivatives is shown. To summarize the arguments in Appendix E, the most estimator for the envelope is derived from the Hilbert transform. The absolute value estimator gives some distortion, primarily during epochs with changing frequency, but requires much less computation than the Wilbert transform estimator.

For real-time recognition of connected speech, the following estimation procedure (shown in Figure 14 and discussed in detail in Appendix E) is proposed. The output of a (wide) bandpass filter is passed to absolute value envelope and zero crossing frequency estimators. Lowpass filters then remove unwanted oscillations. In Appendix E, the best choice for the time constant of these filters (called subinterval length) is shown to be on the order of 1 to 2 ms.

A sliding mean and standard deviation is then computed on the output of each bandpass filter. This procedure has been chosen for its adaptability to real-time operation, its low-cost hardware implementation,

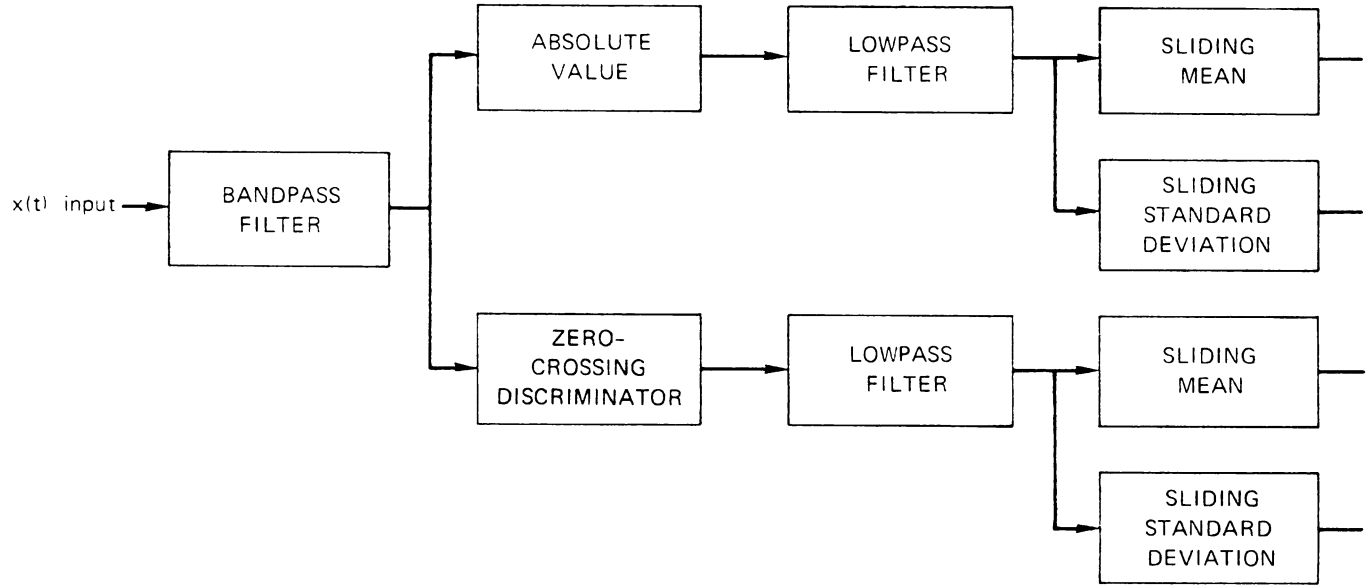


FIGURE 14 ESTIMATION PROCEDURE FOR TIME-VARYING PARAMETERS FOR BANDPASS FILTERED SPEECH SIGNALS

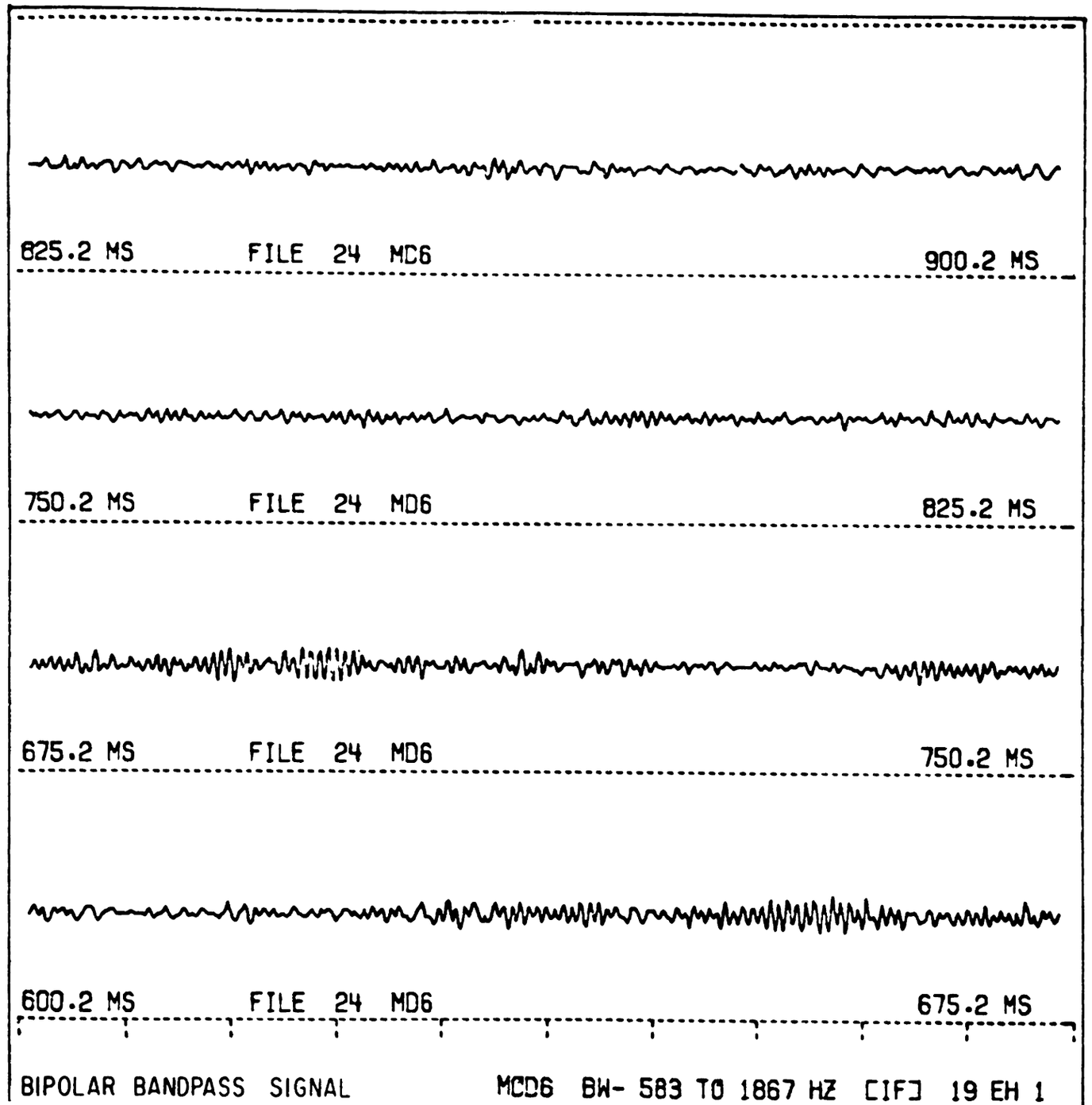
and the minimal degradation of any further processing that may be required.*

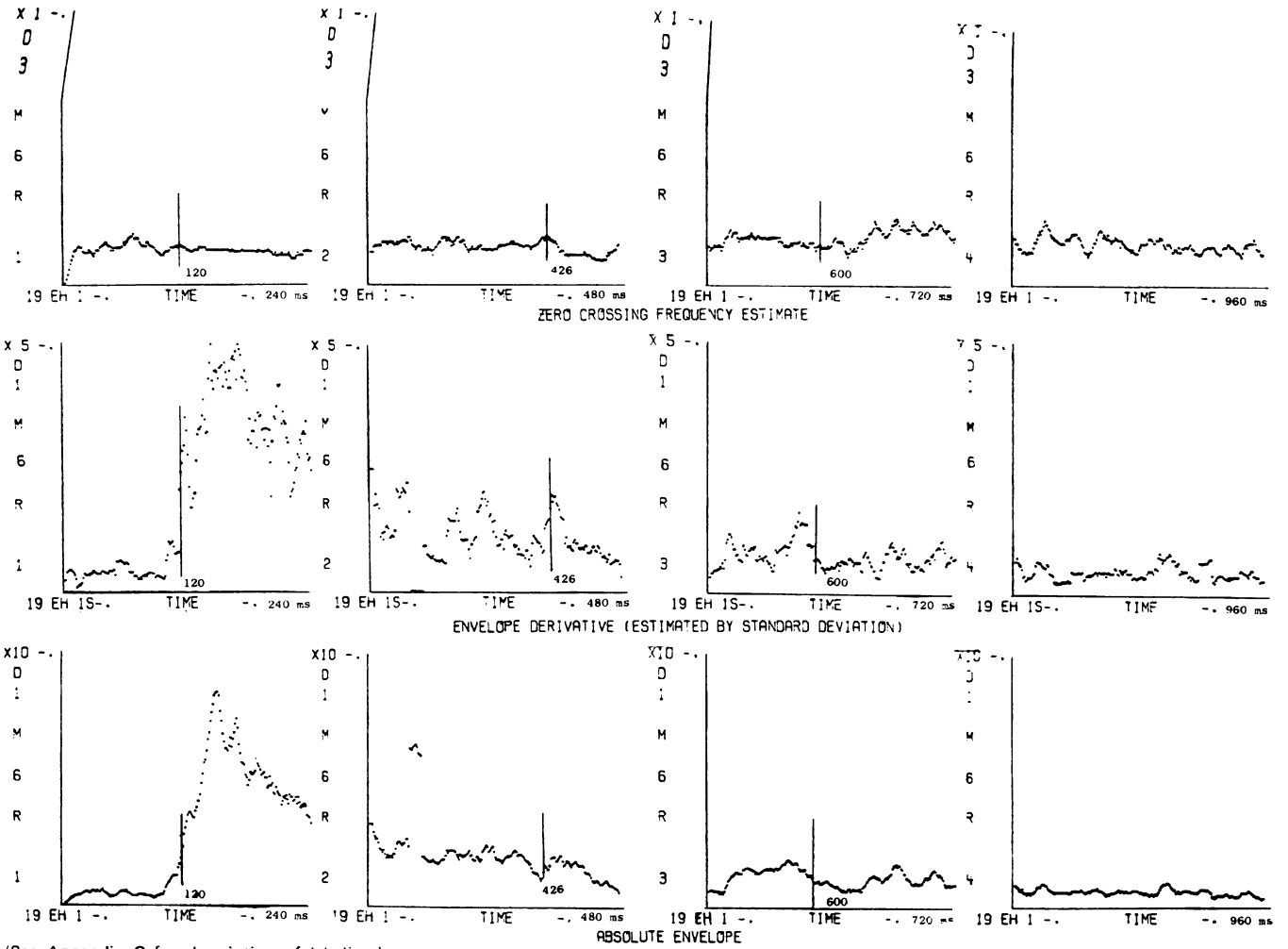
Pictures of bandpass-filtered speech acoustical signals indicate that fricatives such as [f] from [umbif] can be analyzed in a similar fashion. The instantaneous envelope and frequency estimators (both real-time and derivative) retain the stochastic nature of the bipolar signal (Figure 15). The subinterval length of 1.2 ms and sliding average length of 10 ms appears to be adequate for the frequency range shown (see Appendix C for filter bandwidths). Comparison of the bipolar signal (Figure 15a) and instantaneous estimators (Figure 15b) indicates that a narrowband assumption, which has been incorporated in the local ergodic assumption, is appropriate.

Consideration of many cases for different speakers and utterances indicates that the zero crossing-absolute magnitude envelope representation occasionally fails to represent bandpass-filtered signals adequately. The primary case where an ambiguous representation arises is when two energy peaks occur in the same filter (for example, the case discussed in Chapter I and illustrated in Appendix D). In the filter of bandpass 577-1867 Hz (Module 6), a (relatively) strong energy peak continues at 750-800 Hz from 370 ms to 700 ms. At 430 ms, a second energy peak begins to "move" away from 902 Hz toward 1445 Hz. Appropriate choice of filter band bandwidths could isolate these peaks; however, this approach would

---

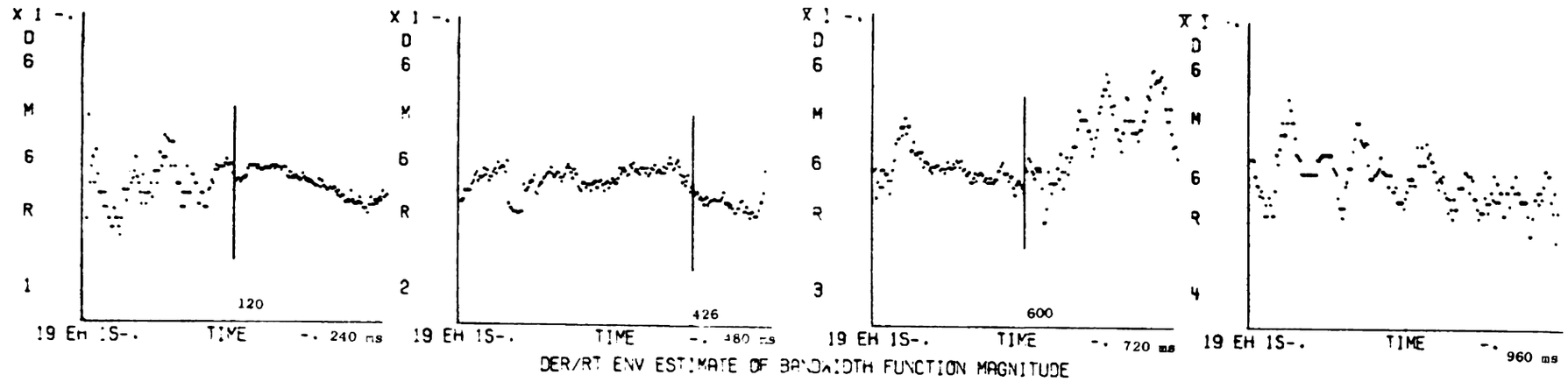
*For instance, if frequency resolution must be increased, the sliding mean length can easily be increased, averaging the previous time series again. However, more frequency resolution cannot be gained by further averaging of the output of a sliding Fourier series computation; a new transform must be computed.


825.2 MS FILE 24 MD6 900.2 MS
750.2 MS FILE 24 MD6 825.2 MS
675.2 MS FILE 24 MD6 750.2 MS
600.2 MS FILE 24 MD6 675.2 MS
BIPOLAR BANDPASS SIGNAL MCD6 BW- 583 TO 1867 HZ [IF] 19 EH 1

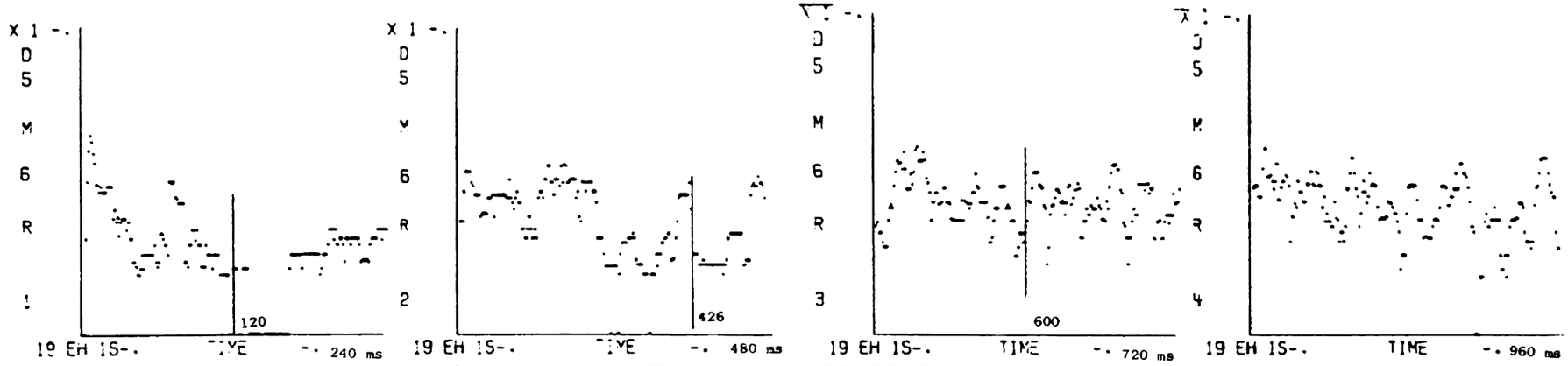


(See Appendix C for description of labeling.)

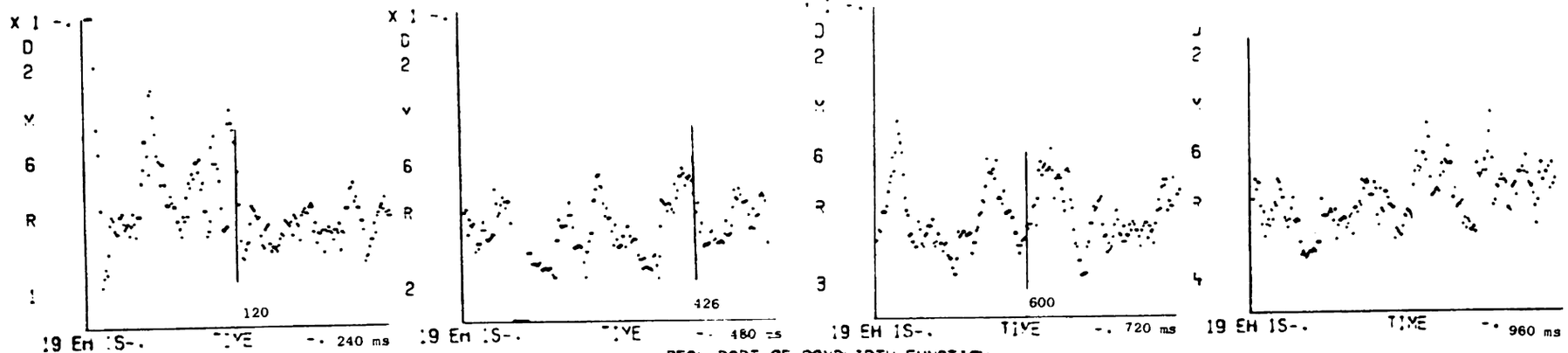
FIGURE 15 REPRESENTATIONS OF BANDPASS FILTERED SPEECH SIGNALS (Continued)



DER/RT ENV ESTIMATE OF BANDWIDTH FUNCTION MAGNITUDE



IMAGINARY PART OF BANDWIDTH FUNCTION



REAL PART OF BANDWIDTH FUNCTION

(See Appendix C for description of labeling.)

require a fixed bandwidth filter tailored to each speaker and utterance. For our particular choice of filters (described in Appendix B), the zero crossing-absolute magnitude representation (Figure 17) follows the stronger low-frequency energy peak.

One method of isolating the more interesting high-frequency energy peak is by first computing the time derivative of the bandpass-filtered acoustical signal and then the zero crossing-absolute magnitude representation. Several factors recommend this approach. Cherry and Phillips indicate an increase from 65 to 92 percent intelligibility by using the derivative (hardware derived) of the wideband acoustical signal for their zero-crossing intelligibility studies. Thomas, referring to this increase, states that the pre-processing accentuates the second formant, which (he proposes) contains the significant linguistic information.

For isolated formants, the increased intelligibility can be due to an emphasis of information-bearing parameters which are related to the prebandwith function (recall that $\frac{d\hat{x}(t)}{dt} = b_x(t)\hat{x}(t)$). In the wideband signal case, high frequencies are emphasized, as we see if we consider the transfer function of an ideal differentiator (linearly increases with frequency). Most physical differentiators* are necessarily approximations and incorporate smoothing to high-frequency variation. A typical transfer function of this approximation is depicted in Figure 16, where lower frequencies are deemphasized (with respect to higher frequencies). Thus, Cherry and Phillips' results can be

---

*In this study, a cubic interpolation is made between extrema of the bandpass filtered speech signal, and this interpolation equation is differentiated.

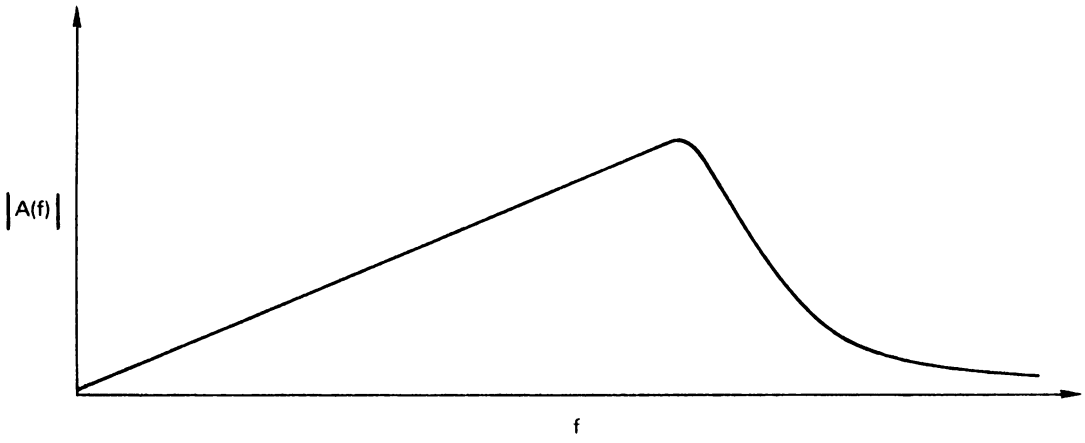


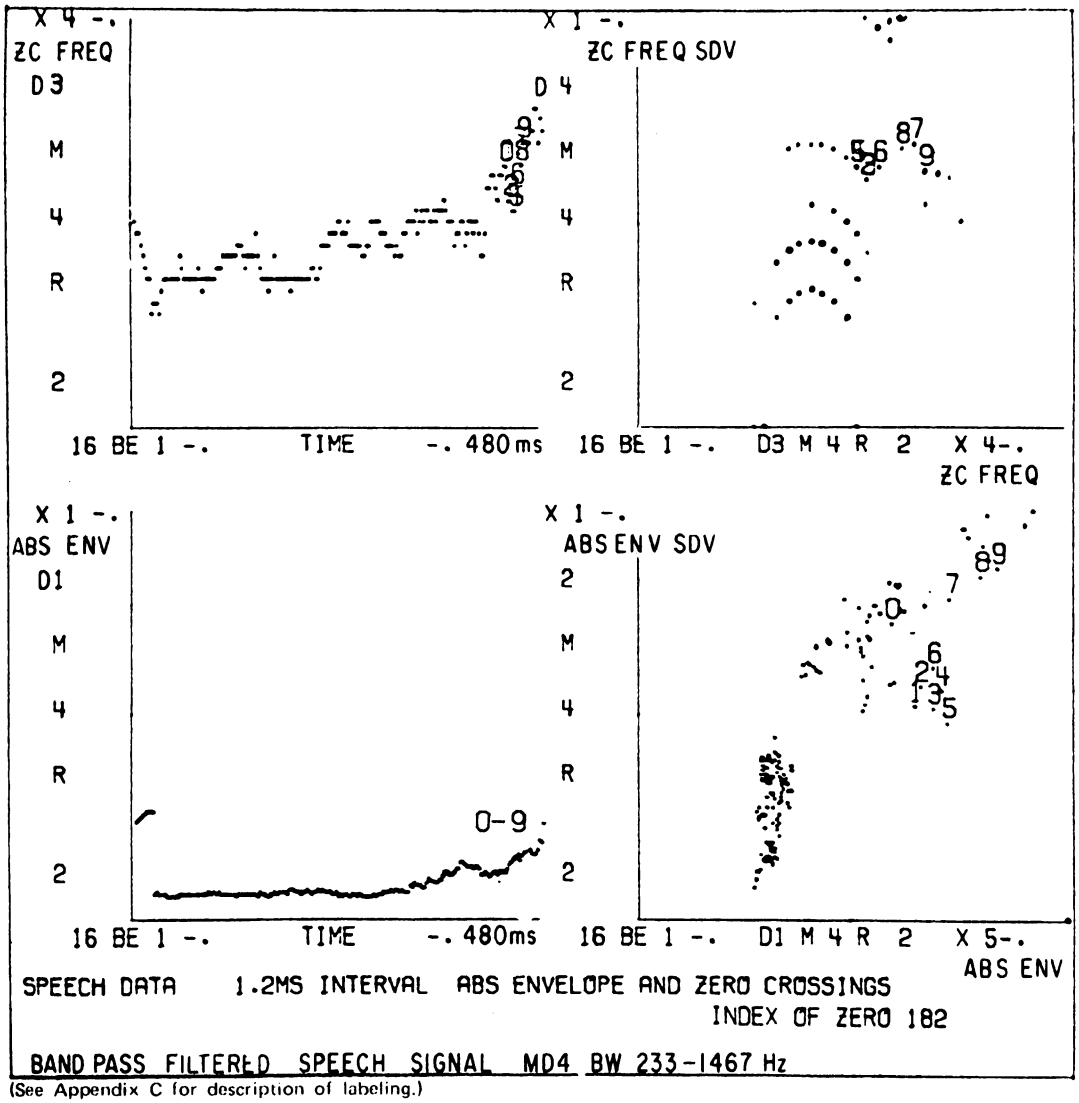
FIGURE 16 SMOOTHED DIFFERENTIATOR TRANSFER FUNCTION

explained if Thomas' hypothesis is true. Resulting zero-crossing-absolute amplitude representations are thus able to "capture" other energy peaks. The frequency estimate (upper left plot of Figure 17b) for Module 7 (derivative of Module 6 filter) and utterance 16 BE 1 clearly shows the frequency transition that was difficult to find in the Fourier series (Appendix D), or in the zero crossing frequency estimate for the undifferentiated signal (upper left plot of Figure 17a).* The resulting transition is depicted even in a situation where bandpass filter selection was not appropriate (for this particular case). The form of the transition is what one might assign by eye to the sliding Fourier series in Appendix D and also looks very similar to the dynamic articulator (tongue) trajectories depicted by Houde for vowel-to-vowel transitions.

Figure 17 shows another feature of the absolute amplitude-zero crossing representation. The sliding standard deviation is plotted against the sliding mean for the absolute amplitude (lower right) and zero crossings (upper right). In both cases (17a and 17b), the bivariate samples form a tight group during the first vowel segment (before 430 ms) and then cross a "bridge" toward a new group during the transition. The differentiated zero crossing case (Figure 17b--upper right) is the most dramatic. The two-dimensional plots can only approximate the actual four-dimensional situation, but it is still possible to recognize a coherent time behavior that is not apparent with standard preprocessing.

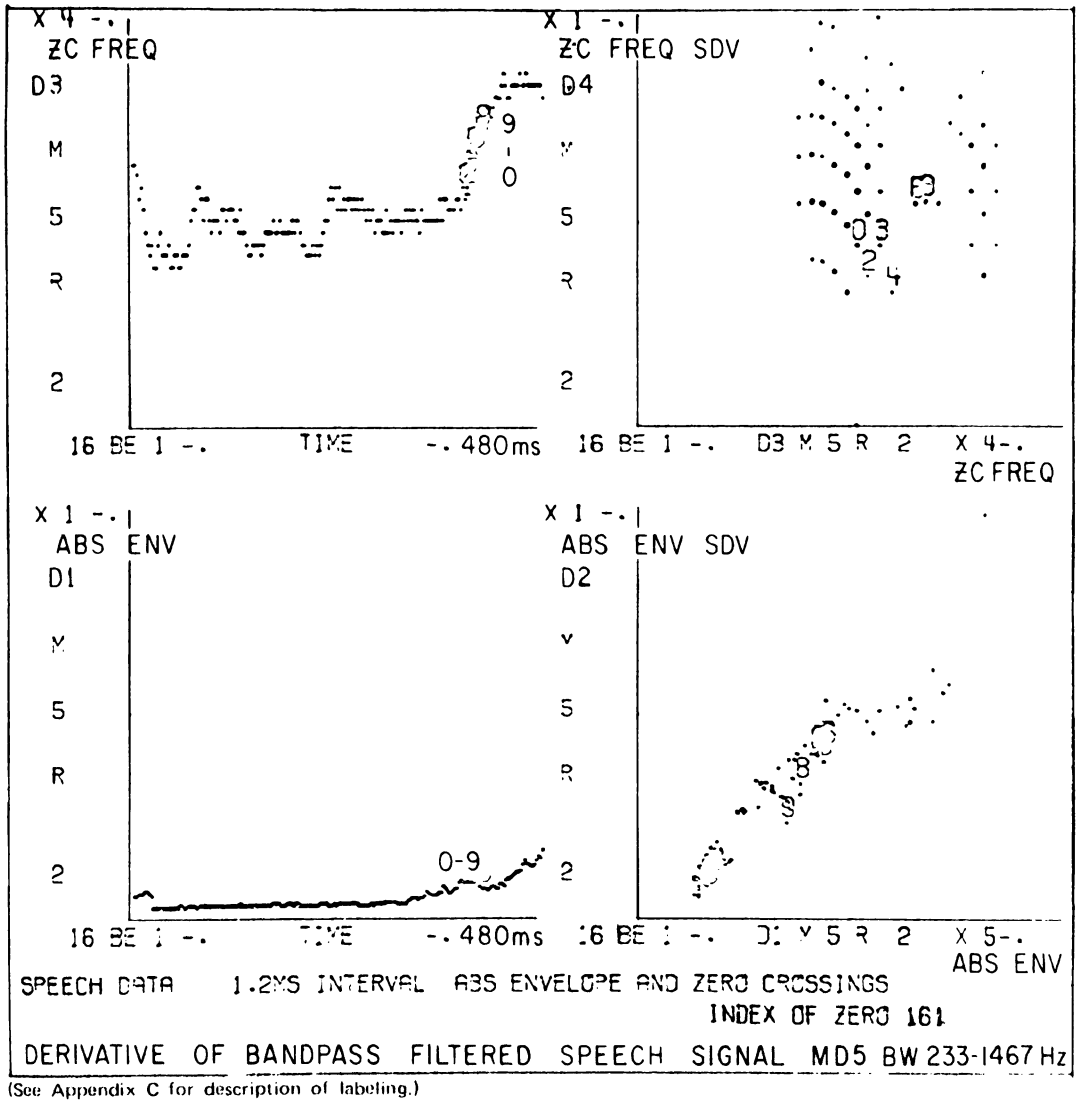
---

*The series of numbers, 0-9, indicates contiguous sample points simultaneously on all four plots. The "INDEX OF ZERO" gives the time of the starting zero in milliseconds.



(See Appendix C for description of labeling.)

FIGURE 17a STANDARD DEVIATION VERSUS MEAN FOR ENVELOPE (LOWER) AND FREQUENCY (UPPER)



(See Appendix C for description of labeling.)

FIGURE 17b STANDARD DEVIATION VERSUS MEAN FOR ENVELOPE (LOWER) AND FREQUENCY (UPPER) (Concluded)

This dynamic behavior is further displayed by various estimators related to the bandwidth function. Two utterances are considered, the u to a vowel transition from [ɖuat^{h}] [16 BE 1, Figure 17] and the utterance [umbif] [19 EH 1, Figure 15]. The following estimators are derived from Eqns. (D-2) and (D-4):

1. Real part--sliding standard deviation divided by sliding mean of envelope

   a. For the real-time bandpass signal

   b. For the derivative of the bandpass signal

2. Imaginary part--sliding standard deviation of the zero crossing frequency

3. DER/ENV--sliding mean of derivative envelope divided by sliding mean of envelope.

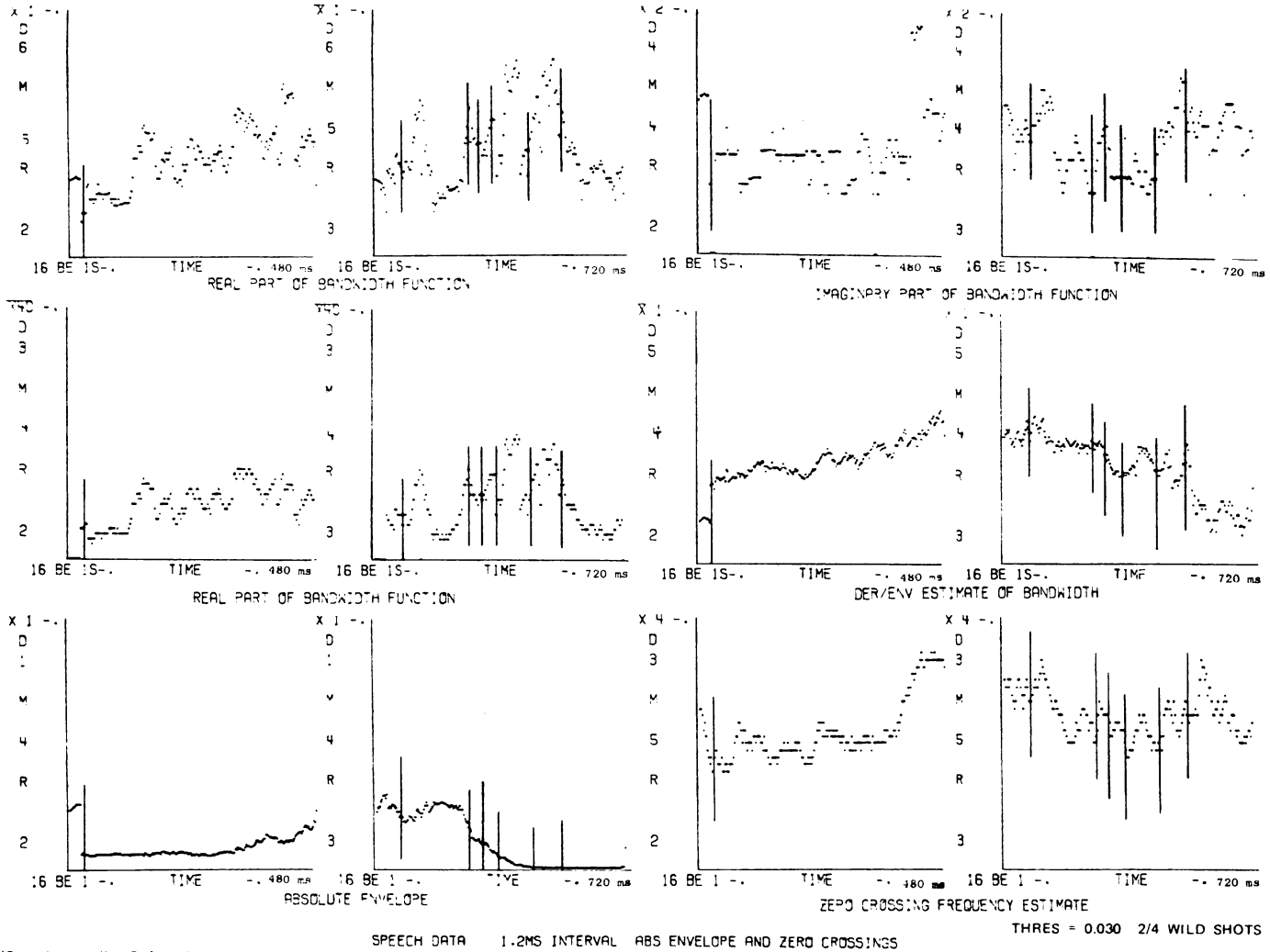
These estimators are shown in Figures 15c and 18. Several points are evident from the figures.

1. Bandwidth function estimates all have a stable "nature" for certain epochs with significant perturbations at the boundaries.

2. These epochs correspond (for some of the time series) to natural speech signal "groupings" (say, Reddy's phoneme classes).

3. The "nature" can be grossly defined in a consistent manner for strong deterministic signal groups (vowels) as opposed to weak stochastic (fricative) groups by the deviations of the bandwidth function about an epoch mean value.

4. The bandwidth function is relatively normalized across rather large amplitude variations while still showing variation for different groups.



(See Appendix C for description of labeling.)

FIGURE 18 SEGMENTATION RESULTS SHOWN WITH BANDWIDTH ESTIMATORS [dhuath] (male)

These results indicate that the first step in a procedure for segmenting connected speech is to identify the points in the (filtered speech) signal where "fundamental changes" occur in the "nature" of the signal. The precise definition of the terms "fundamental change" and "nature" involves specification of a real-time clustering* algorithm and the four time series which give a dynamic representation of the signal. The changes and nature are relative to information we can derive from the particular signal we have at this time (thus termed real time). Since we are dealing with signals that are heterogeneous in nature, any general assignment of functional models to simplify the representation or reduce computations would surely cause higher error rates, at least part of the time (for further inferences based on the functional models, for instance) or ambiguous interpretations of derived measurements. The clustering procedure is real time, operating on data as they arrive without requiring further passes through the data; self-normalizing and not dependent on a priori knowledge; conceptually simple (in terms of number of adjustable parameters); requires little storage and few computations; and gives a more revealing stabilized (in terms of stochastic variability) dynamic representation of the original output along with the marking of points of significant change.

---

*The procedure is termed "clustering" in order to relate a process for dynamic (differential equation generators) transient phenomena to the usual static data clustering techniques (ISODATA, Ball and Hall, 1967). A precise relationship between the static and dynamic clustering exists when one can choose a functional model for a set of differential equations and then estimate the parameters of this model. The set of all parameters would then, for a given time epoch, be one vector of the type that is discussed in static clustering procedures.

In the defined state space, the time trajectory of the differential equations varies about some mean value and the clustering would define limits about that mean value which expand and contract, depending on the time-varying parameters of the differential equations. For the single formant model (and for other higher ordered systems as well), a time-varying mean value and standard deviation can represent the time series state variable value and its first derivative. To show this variation, consider a normalized variable z at time n by the formulas

$$z_n^j = \frac{y_n^j - m_n^j}{\tilde{\sigma}_n^j} \qquad \begin{matrix} n = 1, 2, 3, \ldots \\ j = 1, 2 \end{matrix} \qquad \text{(II-D-5)}$$

for the two time series (envelope y_n^1 and frequency y_n^2) and describe the variations in terms of the distribution of this error term. Inspection of this quantity indicates that normalization is performed by the division by the sliding standard deviation. The segmentation procedure asks the question: Is the differential model, defined by our two time series for each state variable, adequate to describe the variations in the input signal. For that reason, we will consider predictive instead of synchronous normalized variables. That is, instead of using values of z at time n, we will look at distribution of the expected next value of z. If we write this out in a slightly different form, it becomes:

$$y_{n+1}^j = \tilde{m}_n^j + \tilde{\sigma}_n^j \, y_n^j \qquad \text{(II-D-6)}$$

Defining η_n^j as the difference between the mean value and the observed value at each time n,

$$y_{n+1}^j - y_n^j = \tilde{\sigma}_n^j z_n^j - \eta_n^j \qquad j = 1, 2 \qquad \text{(II-D-7)}$$

which is a discrete **version** of (II-D-1). The terms on the right side of (II-D-7) are functions of the state variables, excitation parameters, a stochastic term and time ($n \geq 1$).

These equations can occur in some classical estimation problem formulations:

(1) deterministic but unknown equation

$$M_{n+1} = M_n + f_s(M_n)$$

(2) observation equation (sample function generation)

$$y_n = x_{m+1} - x_n = g(m_n) + \xi_n f(m_n)$$

where

ξ_n are independent, identically distributed random variables for each n, independent of m_n and x_n at time n with moments $u_1, u_2, u_3, u_4, \ldots$

(3) observation equation (ensemble generation)

$$x_{n+1} = m_n = \sigma_n z_n \qquad m_n \overset{d}{=} E_z\{x_n\}$$

$$\sigma_n^2 \overset{d}{=} E_z\{(x_n - m_n)^2\}$$

where

z_n is a random variable with moments v_1, v_2, v_3, v_4, ...

The difference between ii) and iii) is primarily one's point of view (derivatives versus expected values of moments). The relation between these, which is empirically shown in Appendix E, can be derived by taking expected values of ii) and iii).

$$E_\xi\{x_{n+1} - x_n\} = g(m_n) + f(m_n)\, u_1$$

$$= E_z\{\sigma_n z_n\} = \sigma_n v_1 \qquad \text{(II-D-8)}$$

Thus, the four time series for envelope and frequency and their derivatives adequately represent the bipolar bandpass signals and the deviations from these time series can be exhibited in the normalized predictive variables defined by (II-D-6) where $y_n^j = z_n^j$ are the subinterval averages and $\tilde{m}_n^j$ is the sliding mean and $\tilde{\sigma}_n^j$ is the sliding standard deviation for envelope (j=1) and frequency (j=2). Then the points where these four time series no longer represent the input signal can be determined by a statistical test based on the distribution of the normalized predictive variables. This distribution can be estimated by use of the samples up to time n

$$z_r^j = \frac{y_r^j - \tilde{m}_{r-1}^j}{\tilde{\sigma}_{r-1}^j} \qquad j=1,2 \quad r=1,2,\ldots n-1 \qquad \text{(II-D-9)}$$

In general, these values will not be symmetrical about zero because of the nonlinearities. We need an estimation technique more powerful than

the currently popular procedures which are based on the normal distribution. Because of the local ergodic assumption, there will be continuous changes in the parameters rather than "jumps" between two or more ranges of values. Thus, the distribution for each epoch will be unimodal (bimodal distributions will yield two epochs), and a modified t-test with an Edgeworth approximation to the distribution is appropriate.

The segmentation procedure, then, uses normalized predictive samples derived from sliding mean and standard deviation time series to estimate four moments, the coefficients of an Edgeworth series. If the probability of occurrence (from the Edgeworth distribution) of the normalized values of envelope and frequency exceed a predetermined threshold, then that sample is included in the present epoch. If the probability falls below this threshold, then the sample is declared a wild shot. This procedure is useful in identifying (and eliminating) data values of questionable use (such as parity errors, computational errors, external impulse noise) which arise quite often in digital processing of acoustical signals.

The definition of a segment point is an extension of the concept of a wildshot. If the data continue to give low probability, it is quite natural to assume that their "nature" has changed and that a new epoch should be marked. This is controlled by two factors, the number of wild shots and the length of time within which this number of wild-shots must occur. (For example, two wild shots during four time units may define a segment point.) Examples of the segment points resulting from a computer algorithm based on this procedure are shown in Figures 18 and 19. Figure 18 demonstrates dramatically

that the procedure is most sensitive to changes in bandwidth estimators and not in the envelope or frequency estimators. This is not a limitation since simple thresholds can detect significant changes in these time series. Figure 19 shows how the fixing of locations for segment points depends on the choice of the criterion and threshold. There are several important observations related to the 7 regions depicted in Figure 19:

(1) Erroneous data (caused by a computation error) are detected and flagged (2)

(2) The values of threshold and segment criteria depend on the instantaneous nature of the signal (1, 3, 6, and 7)

(3) Immediately after a segment point, a higher threshold should be set to eliminate false alarms (e.g., the threshold might decay exponentially to the set value) (5, 6)

(4) The definition of homogeneity of the epochs is insensitive to amplitude variations even during highly transient behavior (4)

(5) Comparison with Figure 18 indicates that male/female differences do not affect the algorithm.

In summary, the acoustical speech signal is viewed as a composite nonstationary stochastic process and the mathematics of communication theory is used formally to describe and discuss its complicated nature. One isolated formant is modelled by a time-varying differential operator with stochastic or deterministic driving functions. The parameters of this model are related to steady-state concepts of envelope, frequency and bandwidth.

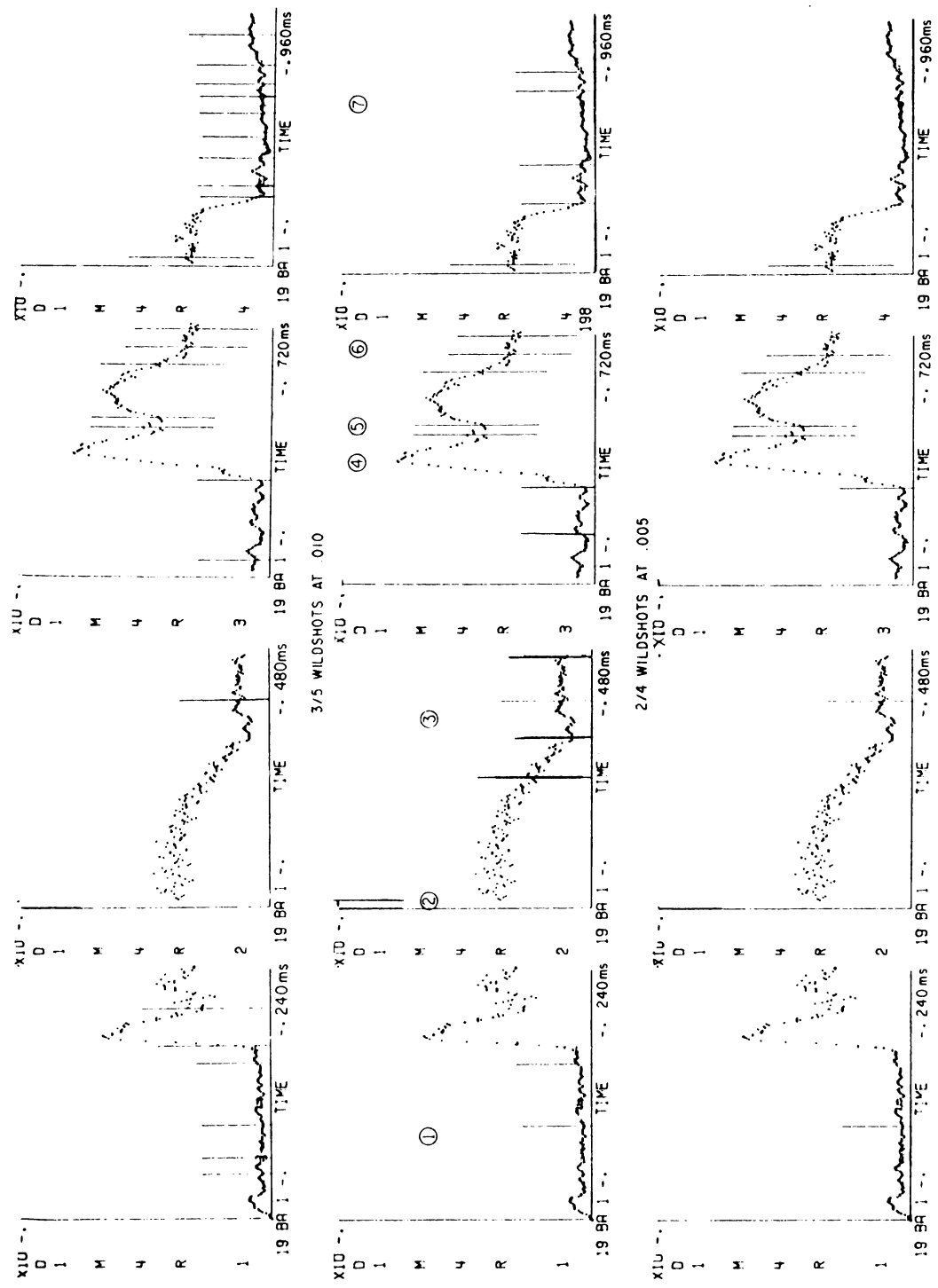


FIGURE 19 SEGMENTATION RESULTS FOR [umbif] (female)

(See Appendix C for description of labeling.)

A derivation of the transient response for linear filters has been used to show how fixed-frequency analysis (such as sliding Fourier transforms) are inadequate and misleading for representation of speech signals (especially the transient portions of these signals). This analysis also defines requirements for the preprocessing wideband filters and the sliding averages used for the pointwise estimators. Formulas for these pointwise estimators of envelope, frequency and bandwidth are derived, and a predictive differential equation segmentation procedure defines as an epoch those samples of the acoustical speech which have homogeneous characteristics.

The envelope and frequency estimators are standard absolute-value and zero-crossing counts averaged over a short interval to show the dynamic behavior of the parameters. Additional preprocessing (derivatives of the bipolar signal) increases the ability to isolate significant features. Using differential equation and statistical moment methods, we have defined time-varying bandwidth estimators which have a stable behavior during epochs corresponding to natural speech classes. They reveal information about the diriving function as well as the differential operator and are normalized with respect to large variations in envelope and frequency.

The segmentation procedure, defined on sliding mean and standard deviation time series for envelope and frequency, is most dependent on variations in the bandwidth function. It also eliminates erroneous (out-of-place) data, is invariant to scale changes (due to male-female and amplitude differences, etc.) but requires a sophisticated (plastic) threshold which depends on the recognition and use of the epochs. This is the subject of the following chapters.

CHAPTER III

THE USE OF LINGUISTIC THEORY FOR THE DECODING OF SPEECH ACOUSTICAL SIGNALS

III-A Introduction

The difficulties of making a correspondence between physical measurements and linguistic events (which we denote as code units), as discussed in Chapter I, lead us to consider the use of contextual constraints to reduce the ambiguities. Linguistic theory has been formulated in an attempt to characterize the complicated internal relationships within one language. In order to make any progress along this course, we must distinguish between characterizing the total knowledge that a language user may ideally acquire (all grammatical sentences, for example) and characterizing the production or perception of particular utterances (referred to as competence and performance, respectively). Chomsky and others have proposed finite models that generate all grammatical sentences of a language. An ASR system must be considered in the second sense, however, and one can justifiably ask whether formal language models merely cloud the issues. For communication between two humans this may be true, but the intended use of our ASR system is with a computer system incorporating a quasi-formal language (FORTRAN, ALGOL) to perform functions. The constraints imposed by dynamic real-time decoding of human language (even in a restricted context such as simple declarative sentences) demand a slightly different interpretation of the procedures suggested by linguistic theories.

In discussing the incorporation of contextual constraints into an ASR system to improve its operation we will consider the following topics:

(1) Error-types and correction

(2) Natural use of spoken language

(3) System memory and real-time operation requirements

(4) Flexibility of recognized vocabulary add-on and delete

(5) Junctures in connected speech

(6) Interface with existing computer operations

(7) Divers speakers.

(1) Error types and correction--Normal definitions of "error (caused by misclassification due to noise, etc.) do not adequately represent natural speech situations by postulating that individuals do not always produce the "correct" phoneme (or equivalent code unit) sequence because of variations in dialect, accent, speed of articulation, etc. Even "perfect" classification into code units will yield sequences that differ significantly from stored messages. A more general definition of "error" should include production as well as misclassification types of errors. Code unit production errors can be further broken down into:

1. Substitution (misspelling) errors
2. Omission errors
3. Insertion errors.

An ASR system which corrects production errors as well as misclassification errors could conceivably work on the principle of identifying in the "best" way the code units and then correcting (by the use of contextual constraints or other means) the code unit sequence (Alter[55], Reddy[56]). For limited vocabularies and one speaker, the correction can be performed by correlation with stored messages (Bobrow et al.[13]). However, such an approach requires much computation and can require meaningless measurements (computing pitch periods in a fricative segment, finding formants in a silent segment before stop-releases). Further, omission errors can be exceptionally disastrous because they can cause synchronization errors. (Alter attempts to protect against omission errors by requiring specialized names for FORTRAN variables.) A better approach would be to have a directed classification, where only the required and meaningful measurements are computed.

(2) Natural Use of Spoken Language--One can avoid some of the problems discussed above by an appropriate choice of vocabulary (throwing out problem words) or by requiring the speaker to articulate his words precisely. However, these alternatives are not compatible with a situation in which the normal use of language must be preserved (e.g., in storing a dialogue between two humans, even with limited vocabularies). In addition, it is known that considerable training is required for someone to be able to control his

articulations in a consistent manner. We are, then, forced to emulate the human's ability to decode speech acoustical signals and to determine the cues important to him (and communicated by him) rather than those convenient for a machine.

(3) System Memory and Real-Time Operation Requirements--
Modern computers with large random-access auxiliary memories permit the use of large vocabularies, but ASR systems for these vocabularies must provide a directed search to reduce computation time and computing requirements (e.g., by limiting the number of templates to be compared). It seems undesirable to tie up an entire computer to decode a speech input and then not to be able to do something with the decoded input. Engineering solutions for ASR systems are often ad hoc procedures which are "tailored" for memory economy and computation speed and which may work quite well for the situation and vocabulary for which they were specifically designed. However, extension of these procedures to larger vocabularies, more speakers, etc., is often a patchwork procedure resulting in a hodge-podge system which may or may not preserve the original economy.

(4) Flexibility of Recognized Vocabulary Add-on and Delete--
Changing the vocabulary by adding or deleting one word (incremental expansion) is not easy with ad hoc systems. Either a new section must be incorporated, or large amounts of reprogramming are required. The problem of code unit sequences with errors enters here too. A good closeness measure for code units must be defined in

order to use a system without adding special case analyses for each "perturbed" sequence. The search for an appropriate closeness measure for human perception of phonemes, for example, has not been fruitful.

Specific problem procedures have an additional difficulty with errors. Suppose the English word "blink" is decoded as "blik" in one case or "bnik" in another. The ASR system must find a best fit for these sequences from its past experience. In normal English, both "blik" and "bnik" are not found. Therefore, the only action taken would be to find the best fit. There is no way to determine that "bnik" can never occur in English (which says something about the "errors present") and that "blik" is a possible English word and possibly a new event. Ad hoc procedures have no way of deriving facts about the procedure operation. This is similar to storing a multiplication table when a multiplication rule would be more general and compact.

(5) Junctures in Connected Speech--Decoding of continuous speech texts requires the introduction of junctures (spaces, commas, periods) not needed in isolated word situations. The ambiguity resulting from lack of junctures can be seen in the following three examples:

1.	(space)	The phoneme sequence /ae ne^{i}m/ decoded as "a name" or "an aim"
2.	(part of speech)	"They are flying planes." ("They" refers to gliders or people?)
3.	(period)	"John is trying to understand this sentence is a problem" (the period could be placed either before "to" or after "understand").

Code-unit--code-word decoding when length of code word is not fixed has been discussed under the subject of uniquely decipherable (UD) error-free codes. The constraints on such a code (Ash[57]) are much too exacting for a natural speech situation. The constraints can be paraphrased for instantaneous codes, which are special types of UD codes. No code word can be a prefix of any other code word. Any subset of a natural language meeting this requirement would have little communicative power left. A UD code that is not instantaneous may be found for some situation, but real-time operation would suffer, as is seen in the following example from Ash.

Let $w_1 \ldots w_{n+1}$ be code words, and 0,1 be code units:

$$w_1 \sim 0$$

$$w_2 \sim 01$$

$$w_{n+1} \sim \underbrace{0 \ldots 01}_{n}$$

This is a UD code, but it is not instantaneous. The unique decoding of the first code unit of the sequence $\underbrace{0 \ldots 01}_{n+1}$ into w_1 must wait until n+1 elements are encountered.

Synchronization errors causing omission and insertion are disastrous for a UD code, and substitution errors could be disastrous without error correction before decoding into code words. This situation is not likely to be successful for correcting production errors.

(6) Interface with Existing Computer Operations--One of the useful applications of an ASR system is in conjunction with artificial intelligence tasks. We will denote these

tasks as text manipulation (TM) programs (information storage and retrieval, questions and answer programs). Presently existing TM programs have no flexibility in interpreting their input representation (typewriters, punched cards), i.e., using semantic information to resolve ambiguities. An ASR system could conceivably be independent of TM operations. However, the need for interaction for proper interpretation of natural language texts is apparent, and a system should not be limited by ignoring the possibilities. In addition, as mentioned previously, most computer operations make use of quasi-formal languages for communication with users as well as definition of tasks, making the interactive possibilities of ASR systems very attractive.

(7) Divers Speakers--Many of the problems of incrementally expandable vocabularies and production errors are caused by allowing many speakers to use the ASR system. Two different situations are possible: In the first, the speaker is known previously and, hence, specific adjustments can be made; in the second the speaker is not known previously and a speaker normalization period is necessary at the beginning of the use period. The variations induced by operating with several speakers (or with the same speaker in different emotional states, for example) constitute one of the major difficulties of applying idealized formal language models. No explicit account is given in the present theories for these variations. It is evident that these variations cannot be completely removed without interaction

between the grammatical system and the lower expressive system (describing the code units).

The seven interrelated topics discussed above give some indication of the problems involved with incorporating contextual constraints into ASR systems. After an exposition of some of the concepts of formal language modeling (from Chomsky[58]) we will proposed a different approach for operating with these requirements.

We will define a (specific) language L as a set (possibly infinite) of texts (each text may be one word, one sentence, one paragraph, etc.), finite in length, constructed from a finite set of elements (code units). The fundamental aim of a linguistic system is the separation of grammatical texts which are members of L from ungrammatical texts not in L. A grammar is a model of L in that it produces texts coming from L. The usefulness of the model is determined by how the texts it produces are related to the grammatical texts. A linguistic theory abstracts general principles about successful grammars and thus gives us a way to compare grammars for different languages. ASR systems perform a linguistic analysis and are thus able to be studied in linguistic theory. Ad hoc systems are not easy to fit into a general theory.

But why attempt to fit an ASR system into a theory of grammars that generate infinite sets of texts when the number of responses must be finite? A review of the preceding discussion should answer this question. First, impossible ("bnik") events are distinguished from improbable ("blik") events. Secondly, the problems of incremental expansion are simpler. We may be recognizing only a finite subset of any grammar's finite set of producible texts, but we can change the membership of the subset within

the larger set much more easily than we could with some bounded finite set.

We have already seen a dichotomization of a total system into TM and ASR systems. This is in accord with the arguments of linguistic theoreticians for such a division.* Linguistic theory does unify these these two seemingly independent tasks. The two parts (or Chomsky's components) are:

(1) The allowable sequences of words or grammatical sentences for a specific language

(2) The composition of the "words".

Rules are given to form sentences from ideas and sounds from "morpheme" representations. In addition to the organizational advantages we gain from such structuring, we can also give more precise definitions to such nebulous terms as "morpheme", "sentence", and "meaning". The vagueness results from the many uses in different theories in striving to overcome the inadequacy of "the written word" in representing spoken utterances.

A morpheme is generally accepted to be the smallest unit that conveys meaning. Its more precise specification is discussed within the linguistic theory presented later. These morphemes form the interface between the two levels in Figure 20 (corresponding to code words in Figure 1). The purpose of the grammatical system is the coding of ideas into morpheme sequences, and the purpose of the phonological system is the coding of morpheme sequences into speech signals. The primary weakness of this definition of morpheme is the use of "meaning". There has been a revival of interest in semantics, the study of meaning, but we still have

---

* "A language is a tremendously involved system and it is quite obvious that any attempt to present directly the set of grammatical phoneme sequences would lead to a grammar so complex it would be practically useless." Chomsky

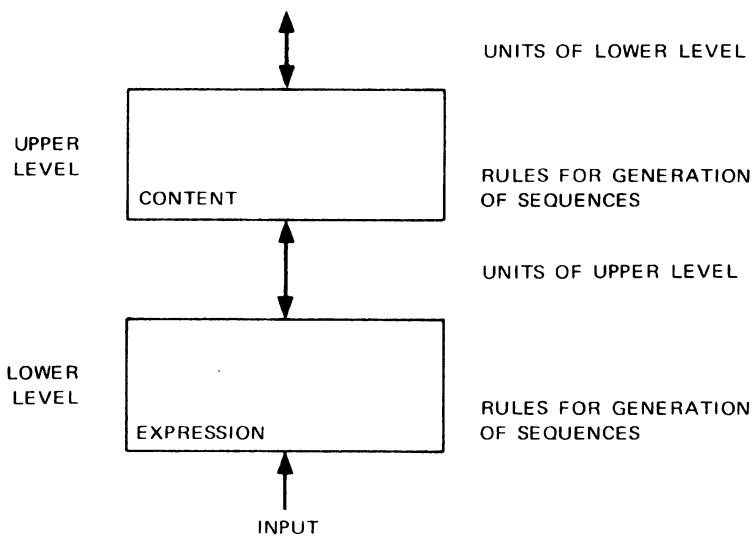


FIGURE 20 FORMAL LANGUAGE MODEL

no concrete guidelines for study of natural language use by humans. When we limit ourselves to communication with a machine, the situation can be a bit more promising. The restricted definition of "meaning" is, in this situation, in terms of machine response. Two morpheme sequences have different meanings if they give different responses. Machine responses could be different programs for compiling, different descriptors for indexing, etc. In restricted cases, such as the computer applications we are discussing here, an unambiguous workable definition of meaning can be given. It must be dependent on the actual linguistic theory used for its precise specification.

In the next two sections, we will make use of a linguistic theory (Lamb's stratificational theory[18]) in specifying an ASR system with the desired features. The language model proposed by Lamb must be augmented to give explicit formulation of "supra-segmental" features. We suggest that the difficult task of juncture identification requires these features. Most linguists agree and give implicit status to these features by calling them supplementary. The features we will use are:

(1) Juncture phon- modification

(2) Speaker expression

(3) Intonation and stress patterns.

The stratificational approach is used here in deference to transformational grammars because of the compatibility with ASR system requirements discussed previously and summarized here.

In the stratificational theory:

(1) Different strata reflect the distinction between representations of ideas where loops occur without regard to direction or to the linear time-ordered

representation closest to the speech acoustical signal (i.e. John hurt himself ~ John ←agent→ hurt. ←object→). We use this dichotomization to treat the predictable "error-like" phenomena discussed previously.

(2) Grammar rules are explicitly identified by the types of sequences on which they may operate, the number of units at which they must look, and the types of phenomena used to index them.

(3) Only those rules required for the particular text need be applied, and we need not apply all rules. The primary classification indexes the subset of rules.

Thus, the output of our ASR system will be sequences of morphemes with junctures. This input can be used with several TM programs now available.

III-B Stratification Model for Generative Phonology

In current linguistic theory, the two parts discussed above in the Introduction are called the grammatical system and the phonological system. (Lamb[18] refers to these levels as the upper and lower strata.) As we saw, our particular requirement for the interface between the two systems is a sequence of context-free morphemes plus junctures. It is not clear that there exists a complete set of rules to perform this task for unrestriced use of American English or even that finding such a set is feasible, but we wish to show a structure which will use such a hypothetical set of context rules for possibly a subset of English. Work with well-specified pseudo-languages indicates that this goal is a realistic one.

We go along with Lamb[18] in defining the following articulatory feature types:

(1) Universal for all human spoken language

(2) Automatically present when specific language is spoken (but not accounted for in (1))

(3) Distinctive presence not predicted by environment and expressing meaning to a listener

(4) Nondistinctive but automatically accompanying distinctive features.*

Thus, we can envision spoken communication as a neutral medium with distinctive variations which convey the desired information. This model leads us to rank the two representations on a given stratum vertically;

*Though not explicitly stated by Lamb, we believe that these types are sometimes quite variable and may be consistent only in one speaker's pronunciation (ideolect). This concept is discussed in the next section.

these ranks will be specified by the suffixes defined below:*

-on -- an objective (as possible) statement of the events

-eme -- the result of an analysis of the events with respect to a specific language for the purpose of identifying only distinctive events.

The property of distinctiveness is a very important one which, in one form or another, is fundamental to language usage and theoretical models to describe language. We cannot have any concrete understanding of the behavior of the two ranks without further specifying "events". The difficulties that arise in attempting to resolve this problem involve other fundamental concepts and lead to the primary contribution of Lamb, his separate strata. One of the most controversial topics in modern linguistics involves units and the segmentation of the speech signal that results from defining units.

It would seem natural to call the "s" on the end of "boys" a type of unit (since it indicates plural) different from that of the "s" in "stone", since the first is a unit of meaning and the second is only one element in a sequence of similar elements which have some meaning because they are grouped together in a particular order. The phonological system gives rules for constructing these sequences (for a given language) from distinctive units. (For example, words can be formed from the 26

---

*The original distinction was between "-etic" and "-emic" ("phonetic being an objective description of sounds that are perceived and "phonemic" being what is linguistically distinctive). The term "phonon" replaces "allophone" or "phone", and "morphon" replaces "morphophoneme" (thank goodness!). The new terminology is more symmetrical and less burdensome (although the term "phonon" has been used by physicists in another "context").

English graphemes.) However, the confusion that may result from a stiu-ation such as the one in our example, plus the requirements of "context-free" strings demands a separate code for meaning units (morphemes) and lower units which make up these meaning units.

One other example may motivate our discussion of Lamb's model. The spelling of English words is often effected by neighboring meaning units (context sensitive). The natural example is the way in which nouns are affected when they become plural ("knife" to "knives", "wife" to "wives"). This is a typical form of linguistic alternation. The pronunciation of these words presents another problem, because the final "s" is pronounced like a "z". This change is independent of the preceding alternation ("bed" to "beds") if performed in the correct order (change f → v; then z after voiced, s after unvoiced). Thus, there is a need to separate these two processes.

Lamb specified two strata in his model of phonology: the morph- stratum and the phon- stratum. It is very difficult to give precise definitions to these terms (at least in anything short of a full article). More often, a construction or generative algorithm is given which results in a precise quantity. Each of Lamb's strata has -emic units as upper interface and -onic units as lower interface (cf. Figure 21). The definition of the stratum label is best understood in terms of its upper distinctive unit. A morpheme is generally accepted as a minimal element, either a meaning unit or a grammatical unit. The following is a list of definitions for a phoneme; each of them contains a small grain of truth:

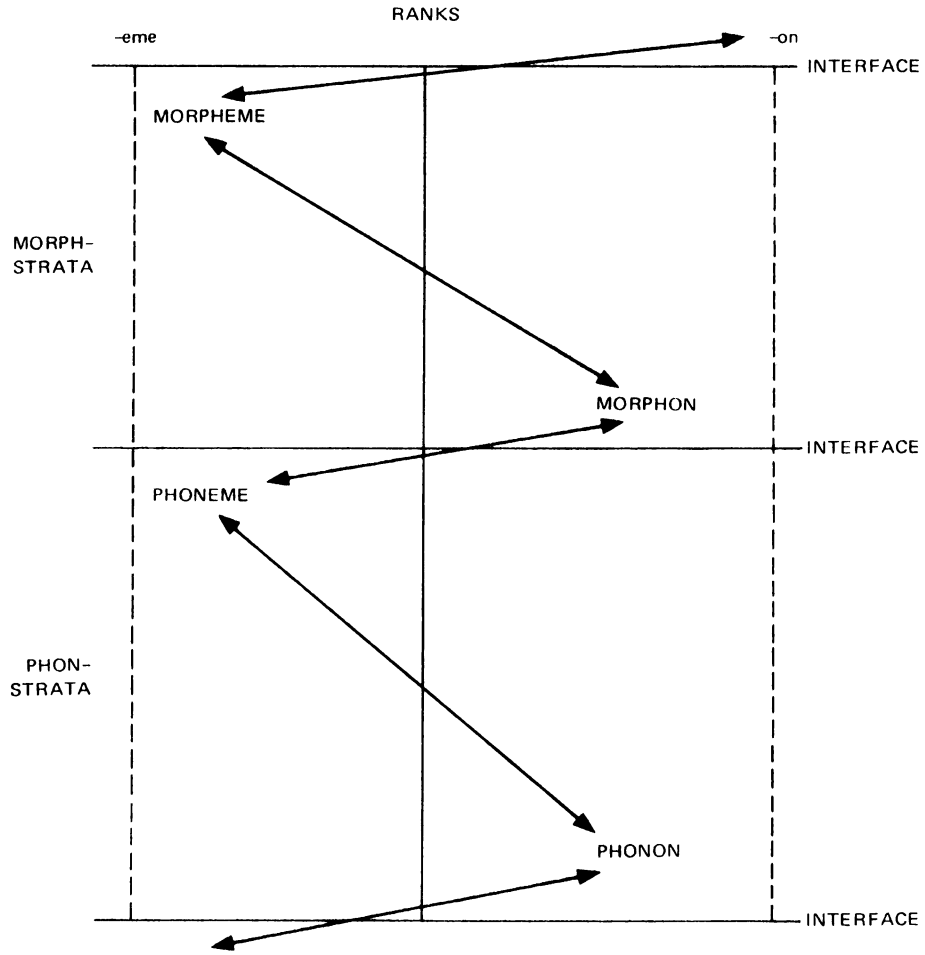


FIGURE 21 LEVEL AND RANKS OF A GENERATIVE PHONOLOGY

(1) "A bundle of distinctive features"

(2) "A class of phones in free variation or complementary distribution"

(3) "A minimal term in phonological opposition".

We would prefer a definition analogous to that for the morpheme: "A phoneme is a minimal element of distinctive expression". We hope that this is as truthful as any other short definition.

Each stratum, then, is simply a correspondence or set of interpretation rules for relating its upper and lower units. The purpose of the dichotomization (although a huge one-shot scheme is often suggested), just as in the discussion of the grammatical system and the phonological system, is to allow independent analysis in each stratum. Because the units at the interface must be context-free sequences (at least with the present state of the art), the operation within each stratum is to "unravel" the context dependencies. An example on the phon- stratum would be helpful. Many dialects of American English do not articulate the initial vowel in "before" precisely; it becomes /ɔ/ as in "buff" rather than /i/ as in "beef". The difference is distinctive for the pair "buff" and "beef" but evidently not in "before". At the upper interface of the phon- level the coding would be the same for both dialectal variations of "before" (i.e. /bifr/) but would distinguish "buff" /bəf/ from "beef" /bif/.

The actual mechanics of this context unravelling involves two types of rules and the unit epoch (duration in time) of each stratum. The length of the morpheme corresponds (approximately) to the syllable, and the length of the phoneme corresponds (even more approximately) to the

grapheme or Arabic letter of printed language. In line with our definition of the -on unit as an objective description of events, the morphon is the same length as the phoneme and the phonon is smaller than the phoneme.

The two types of rules, called realization and composition, establish relations between these units (cf. Figure 22).

(1) Realization rules are the code for the -eme unit in terms of the smaller -on units. Conditioning by neighboring -on units is accounted for here (as in our example above).

(2) Composition rules are the code for transforming the -on unit of a higher level into the -eme unit of the lower level. Conditioning by virtue of belonging to a unit of the higher level (i.e., stress on a morph-length unit affects vowel phonemes) is accounted for here. Alternation caused by linguistic constraints is also accounted for here.

We can return to our example. Suppose we have the morpheme string -- (wife), (Pl) -- to be encoded. The realization rules would code (wife) as /wa^{i}f/ and (Pl) as /s/, with the conditioning rules selecting the vowel glide after /w/. The composition rules would change /wa^{i}f/ to /wa^{i}v/ because of the alternation caused by the plural (cf. Figure 23). This example also shows another distinction between the morph- level and the phon- level mentioned above. Notice that the alternation of morphons occurred only within the syllable. This is the restriction imposed on this stratum. The alternation of the plural morphon /s/ to /z/ because of the/v/ ending of the previous syllable is performed in the phon- level, so the length of influence of each level's rules is

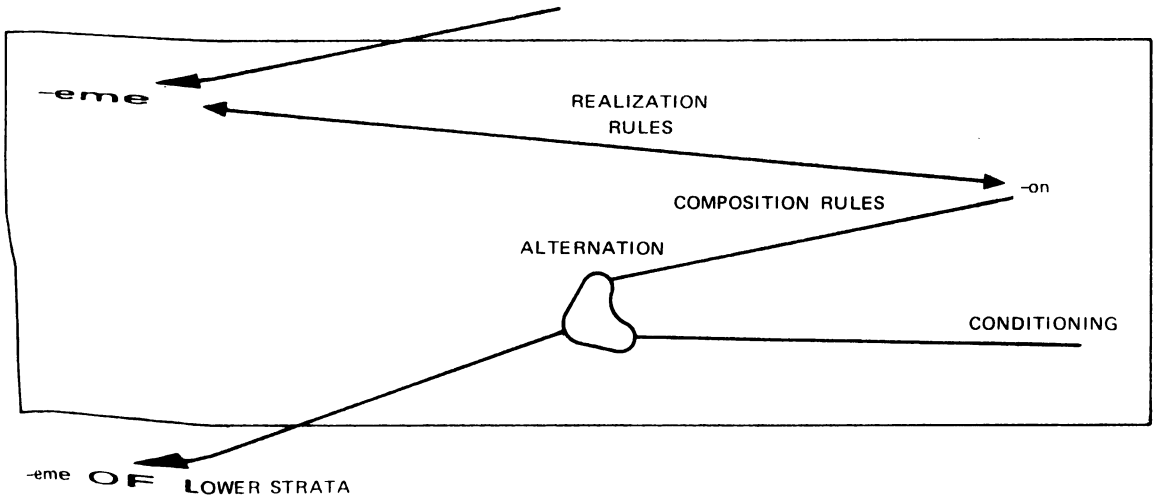
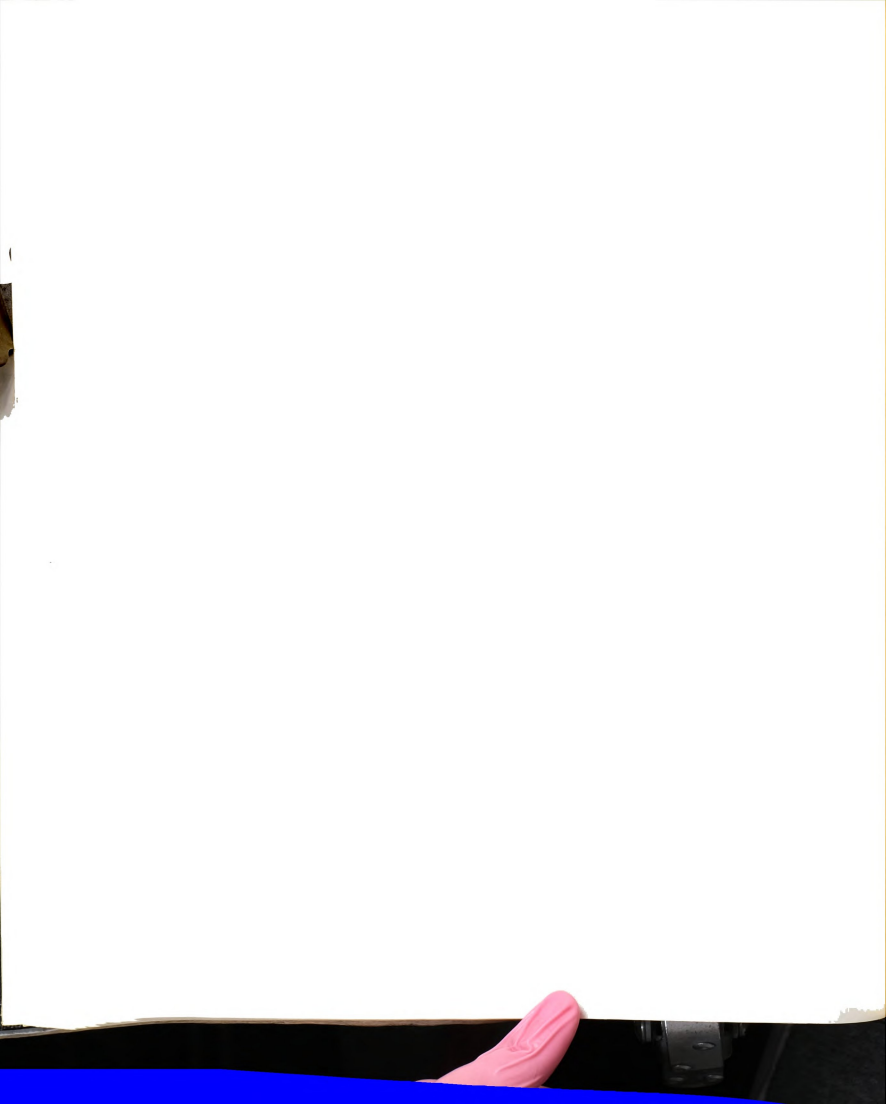


FIGURE 22 RELATIONSHIP OF UNITS WITHIN A STRATUM



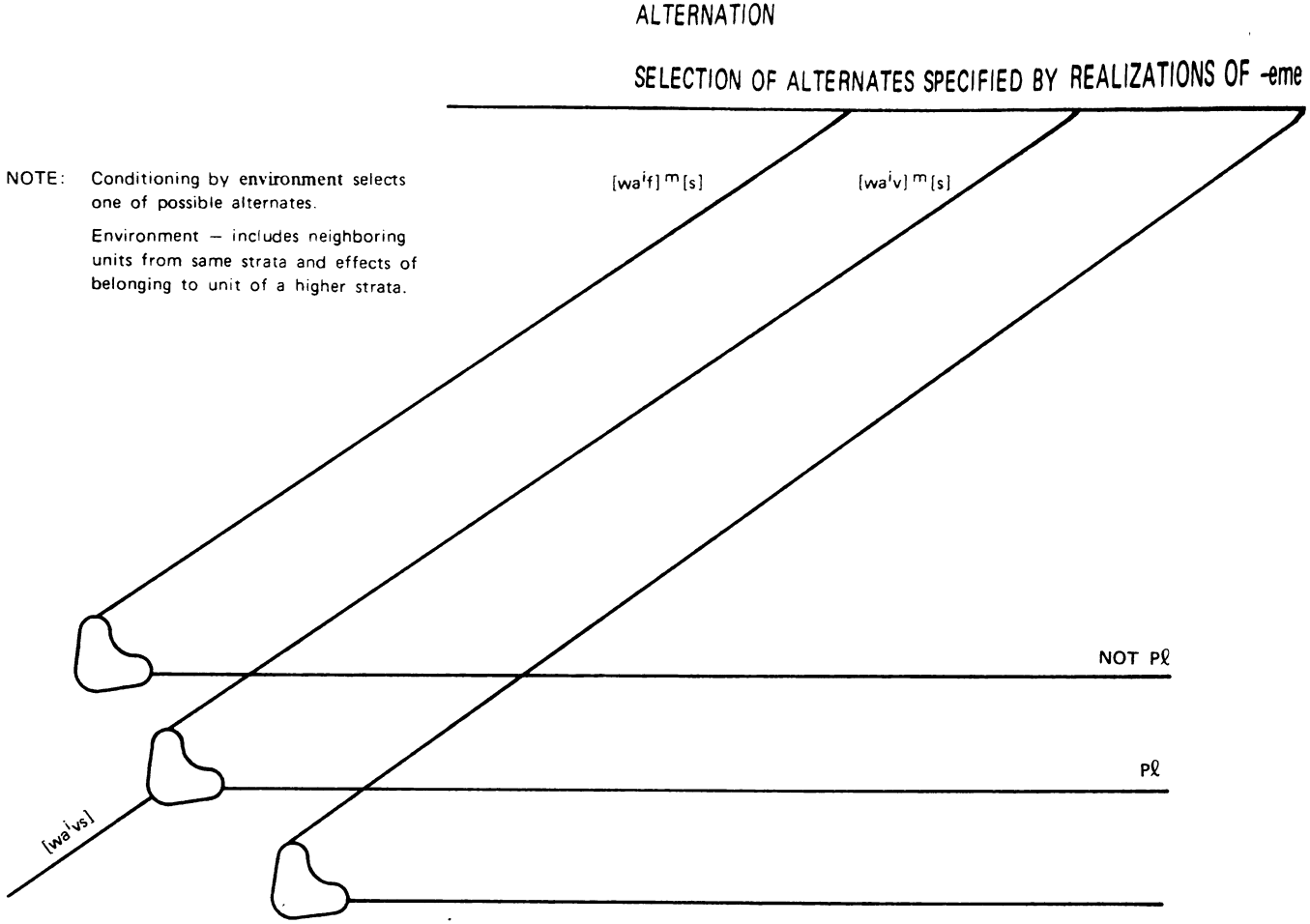


FIGURE 23 COMPOSITION RULES AND EXAMPLE OF THEIR APPLICATION ON MORPH-STRATA

defined as within syllables for the morph- and within clusters (vowel or consonant) for the phon-.

Several features of this model which are particularly attractive for automation can be summarized:

(1) Relation of units -- time length and types -- stresses the relation of a sequence to its members, to the sequence of separate time epochs, and to the combination of components (features) in one time epoch.

(2) The operations within a stratum occur nearly independent of the other strata. This is done by separating the objectives and operation on units within each upper unit. (Again the extent to which any natural language, especially English, can be so described, is not known at the present time.)

(3) The correspondence rules within a stratum are typed; that is, an algorithm which would implement the rule can be very specific with respect to inputs, outputs, and procedures. We have

   (a) Realization rules -- encoding of sequences into higher units

   (b) Composition rules

      1. Alternation -- alternative "spellings"

      2. Conditioning -- rewrite, depending on

         a. Neighboring units

         b. Membership in higher unit.

III-C Recognition Phonology

The stratificational model discussed in the previous section has many properties useful to a recognition system. The dichotomization into nearly independent strata with specification of interaction by means of several types of rules is extremely useful in specifying the training and use of an automated recognition system. The stratificational model is purported to be a two-way model (Chomsky[58]), for recognition as well as generation, but we find that this is not entirely true. Three problems arise:

(1) The lowest unit (closest to acoustical signals) of a generative phonology (Lamb's or any other) is still in terms of abstract quantized units that reflect economical encoding rather than good correspondence with features of the acoustical signal.

(2) This ideal sequence is still ambiguous (in general) unless the specific rule used at each point in the encoding is also known. In recognition situations we do not know the rules used until the correct sequence is known.

(3) Formal language representations only show redundant features in a secondary or "tacked on" fashion for the same reasons of economy mentioned in Item 1. The more realistic situation that apparently operates in human communication will be discussed below.

The highly redundant nature of the correspondence between acoustic features and perceived sounds suggests a slightly different approach than looking

for "primary and secondary features". Human speakers generally have individual language pronunciations, called ideolects; i.e., one person might find that a particular articulatory situation causes a noise burst of a specific center frequency with no modification of the following vowel, and his listener agrees that he "heard" a "b". Another speaker finds that precise modification of the following vowel with no specific noise burst elicits the same response. One cannot say that there is a primary feature here; the listener's responses to both speakers are equally positive. We might call this property of the listener a dialectal generalization. Each person may learn a particular set of features that must be controlled precisely in order to communicate. The remainder of the features (redundant in Lamb's terms for this speaker) are not precisely controlled; thus they may vary considerably with respect to many speakers. This variation will occur above that caused by lack of speaker normalization (suggested by Thomas,[19] Gerstman[3]). Perceptual experiments with repeated words corroborate this conclusion, and the work of Rupert[20] shows that this approach is needed for situations involving diverse speakers. In the light of this discussion, we propose a model which avoids the deficiencies mentioned above.

To overcome the first inadequacy, the same arguments that lead Lamb to a two-strata view of generative phonologies suggest a three-strata model of recognition phonology. This addition may also be useful in a generative phonology, as Lamb has suggested.

Note first that acoustical features can be considered as either absolute or relative with respect to speakers; i.e., schemes can be devised which measure stridency (to distinguish vowel-like and fricative-

like), checked (stop-release) silence and local envelope maxima without speaker normalization, whereas specific formant placement, duration, stress and intonation are very speaker dependent. We can then define a third stratum with upper units called acoustemes. An acousteme is a minimal distinctive unit corresponding to a homogeneous (with respect to both relative and absolute acoustic features) epoch of the acoustic signal. Some implications of this definition are:

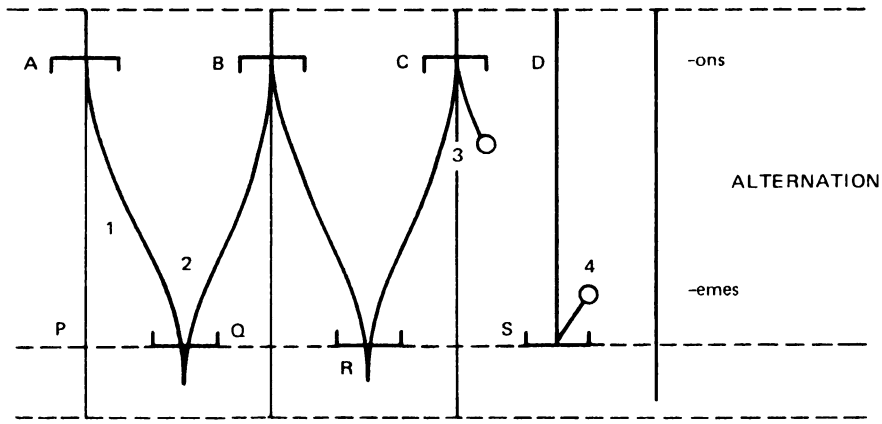
(1) It is a specific definition not only with respect to a particular language but also with respect to a particular speaker and utterance; i.e., different utterances of a given phrase, even from the same speaker, could give different sequences of acoustemes.

(2) The distinctiveness property requires that only the controlled features can be involved.

(3) The segmentation is performed with respect to controlled feature changes and hence induces a useful criterion.

The units thus defined, while more accurately representing the specific acoustical signal, also behave like units of higher strata and exhibit many of the same linguistic phenomena.

The four terms diagrammed in Figure 24 (diversification--A may become P or Q; neutralization--B and A may become Q; zero realization--C may not have a corresponding unit; empty realization--S may be filled in) can be used to describe the various ways in which a speaker actually performs the dynamic task of selecting which features are to be controlled and which are to "float". Examples of these are found in the work of Rupert[20] on isolated words. Several of the phenomena occur in "before".



1. Diversification
2. Neutralization
3. Zero realization
4. Empty realization

FIGURE 24 SEVERAL LINGUISTIC PHENOMENA DESCRIBED BY ALTERNATION RULES

Diversification is seen by the different types of formant structure in the diphthong on the end. Zero realization is almost always seen in initial "b" with the lack or prerelease voicing. Empty realization is exemplified in the extra state or fill-in after the release of "b". Neutralization, which is evident on higher levels ("bitter" becomes "bidder"), is an alternate explanation of the modification of the diphthong.

Lack of knowledge of the specific rules used to generate the acoustical signal and the resulting ambiguity in decoding requires another modification. This ambiguity primarily causes extreme difficulty in placement of junctures (word, phrase, sentence) in the morpheme sequence. An attempt at recovering this information can be made by attaching another rank to the model. The third type of information primarily affects lower strata, such as stress and intonation patterns. This information has not been included in any word recognition system known to date. It is well known that these patterns delineate phrases and sentences. Other types of information occur in smaller units; hence this rank operates on different strata also. For the present, we will label this rank the hyper- -on, which indicates how the information is abstracted at each stratum. The -onic units are the most objective description of the events. The -emic units are generalizations which show the distinctive events; the hyper--onic units are derived from the -onic units and show events which affect lower units. For instance, stress is a feature of a whole morpheme but affects (generally) only the vowel phonemes. Figure 25 shows the augmented model.

(1) Hyper-morphon features include stress and intonation patterns.

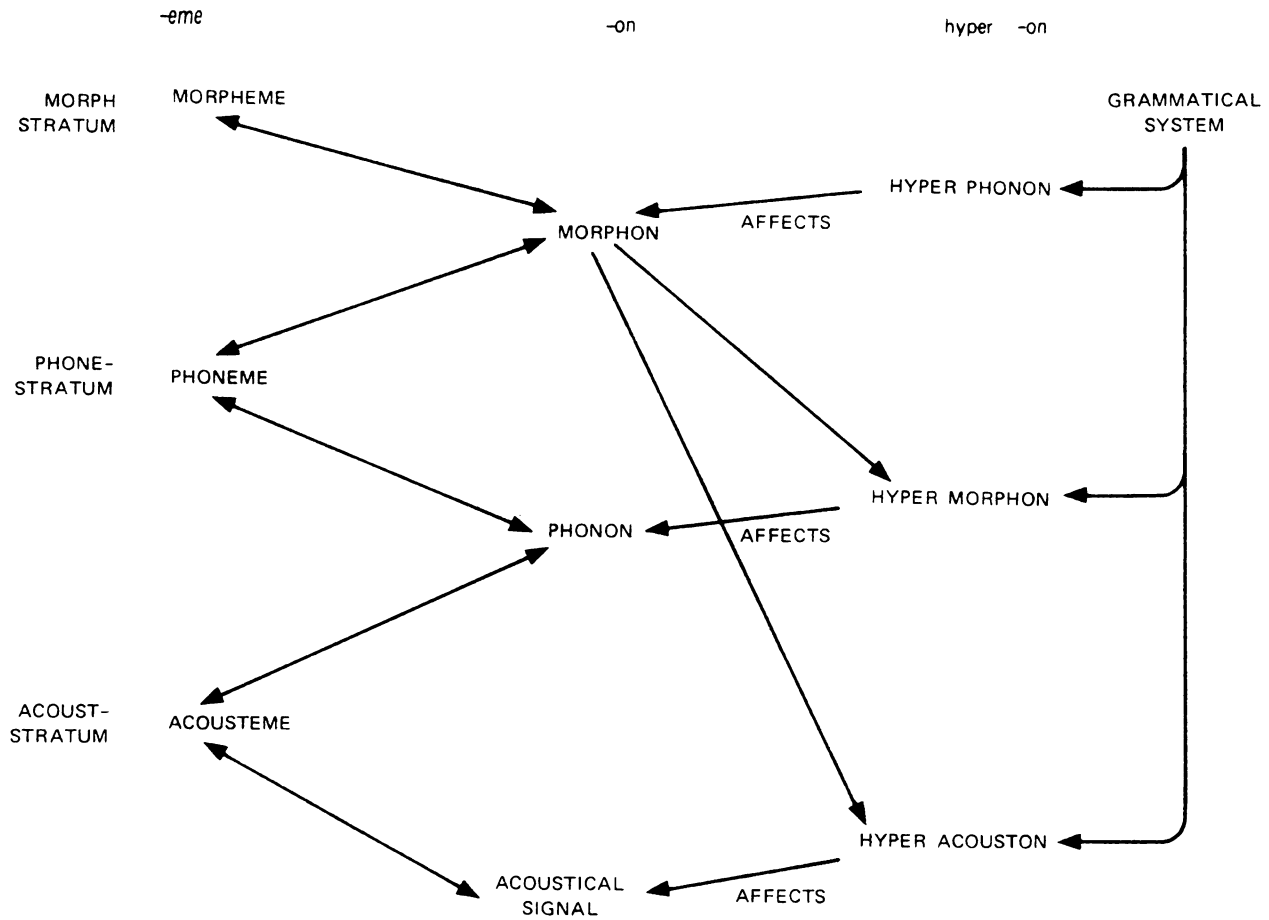


FIGURE 25 MODEL OF RECO-GENERATIVE PHONOLOGY

(2) Hyper-phonon features include juncture phonological variation.

(3) Hyper-acouston features include speaker identity and speaker emotional state.

What this implies in a recognition system is a different type of structure than that suggested by a generative approach, namely a directed algorithm in which predominant features control the search for more fuzzy features and decide which of them are actually needed. This can be accomplished in the generative model simply by attaching a priority to the set of rules so that we can dynamically select the proper rules to be applied. The priority would be a function of the high reliability features (in a degree of presence) filled in by long-term statistical expected events.

Figure 25 shows the augmented generative phonology model which may have the two-way property necessary for recognition. At the least, this figure points out the major problem in recognition of natural language type vocabularies, the number of "feedback" paths. That is, if one starts with the acoustical signal and tries to proceed upward through the model, every step is affected by a higher level. Decisions made on the first stratum will cause a certain interpretation of higher strata which will feed back a conditioning of the lower stratum causing a different decision, and so on.

We will propose a system based on this model without the feedback and discuss it in the next section. It will look roughly like Figure 26.

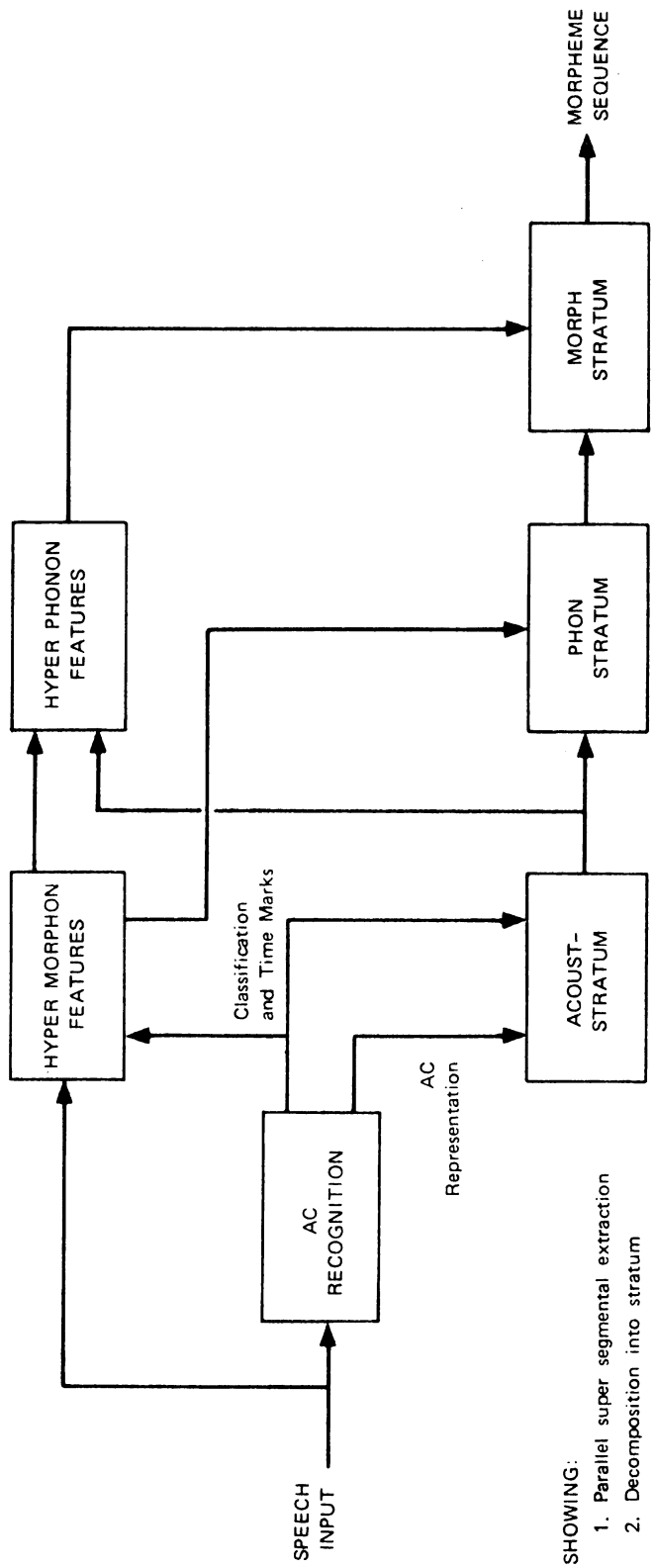


FIGURE 26 RECOGNITION SYSTEM WITHOUT FEEDBACK

IV RECOGNITION STRUCTURES FOR REAL-TIME SPEECH PROCESSING

Dynamic recognition of connected speech is much more difficult than most pattern-recognition problems. In the first place, the complex acoustical signal requires some modification or filtering to accentuate its significant characteristics. Secondly, the unknown nature of the precise generating model at each instant poses an identification problem. Thirdly, the dynamic changing nature of the information content of the acoustical signal requires a sequential decision structure. In this chapter, we will attempt to define an adequate structure for real-time recognition of the acoustical speech signal according to the models developed in Chapter II and the criteria developed in Chapter III.

IV A. Reduction of Dimensionality Using Bayes' Formulation

Speech-recognition algorithms that permit real-time computation require low-dimensional representations (input pattern vectors). We will use modern communication techniques to show how the representation/recognition schemes developed here can reduce the normal dimensionality of the input acoustical signals. The results of Section II-A indicate that we may represent each single formant present in the acoustical signal by a two-dimensional state vector with an associated differential equation (Eqn. II-A-13). Further, the results of Section II-D show that our segmentation procedure permits a time partition of the acoustical signal into epochs, each of which can then be classified in sequential fashion.

Lainiotis[59] considered signal detection and recognition in a recent paper and attempted to determine a "natural" dimensionality. His problem formulation, however, does not account for the complicated

relationships between features described in Section I-B. Thus, we have to restrict the conceptual model of information-bearing features to fit into this signal-detection model. First we must assume that one formant contains the significant linguistic information at any one time. This idea is proposed in the literature by Thomas[19] and implied in the theoretical linguistic work of Fant, Jakobsen, and Halle,[11] describing their distinctive feature matrices. Thus, one regards other formants present in the acoustical signal as correlated noise and therefore only a disturbing character for the true signal (dominant formant). In addition, we must restrict the type of driving function allowed in the state variable differential equations. Fricatives and nasals can be adequately modeled with white excitation processes, but vowel-driving functions (pitch pulses) are not easily accounted for. For this particular formulation, one can consider only whispered vowels. Yilmaz[6] and others have followed this approach in their studies, since there is some degree of intelligibility in whispered vowels.

We must emphasize at this point that these restrictions, although discussed in the literature, are made merely to permit the mathematical formulation of sequential detection theory. However, we can consider the implications of the theoretical results based on this restricted speech model and use these results to make further inferences about the more general speech signal case. Suppose we are given the following noisy observations:

$$\underline{y}(t, \underline{\theta}_\sigma) = \underline{H}(t, \underline{\theta}_\sigma)\, \underline{x}(t, \underline{\theta}_\sigma) + \underline{\nu}(t) \qquad \sigma \in \Gamma_\sigma$$

of the stochastic signal (one of the ensemble of possible speech "sounds") represented by a state vector $\underline{x}(t, \underline{\theta}_\sigma)$, where

$H(t,\theta_\sigma)$ is the transmission matrix

$\nu(t)$ is a white guassian noise process, independent of x, with zero mean and unit variance

$\underline{\theta}_\sigma$ is a parameter vector specifying the differential equations generating $\underline{x}$ (Eqn. II-A-13)

Γ_σ is the index set for σ. Each σ refers to a single discrete <u>value</u> of the parameter vector $\underline{\theta}_\sigma$ (i.e. the range of each component must be discrete)

and a set of signal switching times $\{t_i\}_{i=1}^{k}$ $t_1 < t_2 < t_3 < \ldots < t_{k-1} < t_k$. The natural dimensionality, or structure, is then determined by the parameter vector $\underline{\theta}_\sigma$. In addition to having different values for each component for different σ, $\underline{\theta}_\sigma$ may have fewer or more components or may reference different noise structures (one or two correlated formants).

Lainiotis shows that the pattern-recognition/detection problem of determining the presence of one of M signals from these noisy observations when the signals are generated by differential equations of unknown functional form has a Bayes minimum-risk solution of the following form:

(1) A bank of nonlinear Kalman-Bucy filters is derived based on <u>every</u> possible form of the differential equations, that is one for each $\sigma \in \Gamma_\sigma$. The output of each K-B filter is a conditional mean; i.e.

$\underset{\sim}{x}_\alpha(t/\nu_k^t, V_\alpha^{k-1}, \theta_\sigma^\alpha)$ is the expected value of the state vector for each possible hypothesis H_α (signal $y_\alpha(t)$ is present) $\alpha = 1, \ldots, m$, conditioned on the present signaling interval, $\nu_k^t = (t_{k-1}, t)$ $t \leq t_k$, and all past signaling intervals, $V_\alpha^{k-1} = (\nu_\ell / H_\alpha$ was active) and the value of the parameter vector $\underset{\sim}{\theta}_\sigma^\alpha$.

(2) The expected value of the signal process is derived from the conditional means from each K-B filter by a mixture probability formula

$$\tilde{y}_\alpha(t|\nu_k^t, V_\alpha^{k-1}) = \sum_{\sigma \epsilon \Gamma_\sigma} P(\underset{\sim}{\theta}_\sigma^\alpha | \nu_k^t, V_\alpha^{k-1}) \cdot H(t, \underset{\sim}{\theta}_\sigma^\alpha)\, \underset{\sim}{\tilde{x}}_\alpha(t|\nu_k^t, V_\alpha^{k-1}, \underset{\sim}{\theta}_\sigma^\alpha) \qquad \text{(IV-A-1)}$$

where $\underset{\sim}{\tilde{x}}_\alpha(t|\nu_k^t, V_\alpha^{k-1}, \underset{\sim}{\theta}_{\sigma i}^\alpha)$ is the conditional mean of the a posteriori distribution computed by means on a nonlinear* time-varying filter (Kushner,[60] Kalinapur[61]) $H(t, \underset{\sim}{\theta}_\sigma^\alpha)$ is a matrix which converts the state vector $\underset{\sim}{\tilde{x}}$ into the observed signal $\tilde{y}$ and $P(\underset{\sim}{\theta}_\sigma^\alpha | \nu_k^t, V_\alpha^{k-1})$ is the learned a posteriori probability of the parameter vector value $\underset{\sim}{\theta}_\sigma^\alpha$ conditioned on the present signal interval and all past signal intervals when H_α was active.

---

*Lainiotis specifies only linear differential equations, but the extension to the nonlinear case is obvious.

(3) The likelihood ratio for each hypothesis, H_α, is then computed using a correlator/estimator formula (Kailath[62]) with the input signal and the conditional mean $\tilde{y}_\alpha(t|\nu_k^t, V_\alpha^{k-1})$, and a standard Bayes criterion is used for the decision.

Even for the restricted speech model, and assuming the nonlinear estimation filters can be implemented, this is still an inadequate solution. The primary difficulty is the mixture formula (Eqn. IV-A-1). Although this formula is optimum in the Bayes sense for randomly selected generating models (Wainstein and Zubakov[63]),* one would not expect the converged a posteriori probabilities of each model parameter vector value $\underline{\theta}_{\sigma i}^{\alpha}$ to be either 1 or 0; i.e. the formula (Eqn. IV-A-1) reduces to selection of one filter. Then the conditional mean, $\tilde{y}_\alpha$, will be a sum of an output from the "correct" filter (given the particular signal and noise conditions) and others that are based on noise or other unwanted signals. For high signal-to-noise ratios (ratio of inner to outer distances for a pattern-recognition case), with appropriate models, 1-0 probabilities may be learned; in speech, however, this is very difficult, because of the large class of "signal-like" noise processes.

Implementation of the optimum solution involves many difficulties. Some of these problems are discussed below:

---

*The comparison for a parameter measurement/detection problem is between a mixture formula and choice of the maximum probability (likelihood ratio). For large m (number of filters), a factor of two minus signal-to-noise ratio is required to maintain the same probability of error.

1. Implementation of Filter Bank

The nonlinear estimation filters are physically unrealizable (they require an infinite number of components). Suboptimum techniques are available (Byrd[64]), but they sacrifice the ability of the optimum filters to match the transient response in order to have appropriate asymptotic behavior. As was pointed out in Chapter II, connected speech signals from divers speakers contain critical dynamic portions requiring good transient response in the preprocessing stages. Thus, the heuristic criteria developed in Chapter II are more appropriate.

The optimum formulation requires a model of the desired signal plus unwanted (correlated) signals and noise for each filter. For the suboptimum case, however, this is not always desirable, especially if one is not sure of the exact structure of the undesirable signals. Groner[65] has shown that under certain conditions the performance of linear threshold elements with adjustable weights can be decreased by increasing the number of inputs. In attempting to classify isolated words from one speaker using the zero-crossing counts and energy levels from eight bandpas filters, performance measures increased as a function of the number of inputs but then began to decrease. The posited explanations were:

(1) Assumptions about the pattern statistics were not correct.

(2) The number of sample patterns was insufficient.

(3) Incorrect structure or training algorithms were used.

(4) Training was not allowed to continue to convergence.

Determination of the proper number of inputs to give a maximum performance is difficult to accomplish except in special cases. One example from Groner may help.

When one is using Euclidean distance differences to classify pattern vectors, an additional measurement degrades performance when

$$2\sum_{i=1}^{n}\Psi_i + \Psi_o > \frac{2n+1}{n^2}\sum_{\iota=1}^{n}\sum_{j=1}^{n}\varphi_{ij}$$

where $[\varphi_{ij}]$ $i,j = 1, \ldots, n$ is the correlation matrix of the existing n measurements

Ψ_o is the variance of the new new measurement

Ψ_i $i = 1, \ldots, n$ is the correlation of the new measurement with each old measurement.

Hence the performance is degraded by the addition of a new measurement which is correlated with the others and which adds noise (proportioned to Ψ_o) to the recognition process.

2. Independence of Filter Bank Outputs*

The mixing formula (IV-A-1) is based on the assumption of randomly selected generating models and optimum least-mean-squared-

*Probabilities computed on two input sets X^k and X^ℓ are independent if $P(X^k, X^\ell) = P(X^k)P(X^\ell)$.

error estimation filters. The highly structured and situation-dependent interrelationships of acoustical features make the former assumption very suspect. Further, the choice of suboptimum filters again indicates a set of dependent probabilities. In order to achieve the superior performance of a mixture formula, Wainstein and Zubakov apply the central limit theorem for a sum of independent random variables. Thus it would be beneficial to the performance if the mixture probabilities were independent. A second observation about such sums is pertinent here. The study of robust estimators shows that convergence to a stable value is quicker for arbitrarily distributed random variables if "outliers" (events significantly removed from the mean) are not included. In this context, probabilities assigned to certain filter outputs can be "outliers" due to reasons cited in the discussion of the implementation of the filter bank. The long training period that may be required even for optimum classifiers (which is lengthened due to outliers) is especially detrimental in the speech situation. The plastic structure must be responsive to "drifts" and slow changes in the input's salient features.

Thus, for the suboptimum filter bank specified in Chapter II and Appendix B, we need to investigate recognition structures which form near independent probability estimates and mixture formulae which reduce the undesirable effects of outliers. We will show in Section B how the Lewis[42]-Brown[43] probability approximation technique attempts to compute independent probabilities and in Section C how the S-RETIC algorithm of Kilmer[41] operates to eliminate outliers.

IV B. Quasi-Independent Probability Distributions

The discussion in the last section indicated that the set of a posteriori probabilities computed on the outputs of the preprocessing filters should be independent in order to increase recognition performance. This will be difficult to achieve because of the overlap of the input sets--one for each probability computer--and also because of the correlated nature of the inputs.

The Lewis-Brown[42,43] iterative technique can be used to reduce the dependence between the probabilities. The notation follows that of Section I-E. Suppose we have (for a given class C_ℓ) a set of m low-order distributions $\{P_k\}_{k=1}^{m}$ such that

$$P_k(x) \geq 0 \qquad k = 1, \ldots, m \quad \text{for all } x$$

$$\int_{x \in X^k} P_k(x)\, dx = 1 \qquad k = 1, \ldots, m$$

where X_k^* is the set of n inputs for the k^{th} probability computer. Then, if we consider the entire m x n dimensional pattern vector for a given class and hypothesize a "true" distribution, each low-order probability distribution, $P_k(X^k)$, satisfies a marginal property; integration of the "true" probability distribution over all components not contained in X^k equals $P_k(X^k)$. Brown gives an iterative procedure for determining,

---

*For speech, one input, x_i, might be one component of a four-component state vector representing the output of one filter of an m-filter (overlapping) bank. X^k, then, is the state vector (n=4) for each filter.

among all products of low-order distributions that satisfy the marginal property, the one that minimizes an information measure of the closeness to the "true" probability distribution. Brown defines an iterative procedure as follows: Given an initial (a priori) m x n distribution $P^o(x)$, define the j^{th} iteration, j = 1, 2, 3, ..., probability distribution, P^j, from the set of low-order distributions $\{P_k\}_{k=1}^m$

$$\underline{P}^j(x) \stackrel{d}{=} \underline{P}^{j-1}(x)\left[P_k(x) \,/\, P_k^{j-1}(x)\right] \qquad \text{(IV-B-1)}$$

That is, multiply the $(j-1)^{th}$ probability distribution by the k^{th} low-order distribution, where k = j modulo m, and divide by the marginal distribution.

$$P_k^{j-1}(x) = \int_{x \,\epsilon\, X^k} P^{j-1}(x)dx \qquad \text{(IV-B-2)}$$

Brown shows that the distribution P^j does satisfy the marginal requirement for all j and does converge to a limiting distribution with the minimum information property.

At first it appears that P^j will contain low-order distributions raised to a power but if we rewrite the marginal distribution (B-2) we can see that this is not so.

$$P_k^{j-1}(x) = \frac{P_k(x)}{g_k^{j-2}(x)} \cdot g_k^{j-1}(x)$$

where the g's will be defined. Substitution of this into Eqn. (IV-B-1) gives (after m iterations)

$$P^j(x) \overset{d}{=} \prod_{k=1}^{m} P_k(x) \;/\; g_k^j(x) \qquad j = m, m+1, \ldots \qquad \text{(IV-B-3)}$$

where

$$g_k^j(x) = g_k^{j-1}(x) \qquad j \neq k \text{ modulo } m$$

$$g_k^j(x) = \int_{x \in X^k} \prod_{\substack{\ell=1 \\ \ell \neq k}}^{m} P_\ell(x) \;/\; g_\ell^j(x)\, dx \qquad j = k \text{ modulo } m$$

Note that $g_k^j(x)$ is a function of $x \in X^k$ and hence the g_k^j functions tend to make $\{P_k\}_{k=1}^m$ a set of independent probability distributions so that a product rule for recombination applies. The computation of g_k^j requires, for a given module, an integration over the set of measurements not contained in the input to that module. The g_k^j functions have the same limiting property as discussed in Brown. To see this, define the limiting probability distribution

$$P^r(x) = \lim_{j \to \infty} P^j(x) \qquad \text{(IV-B-4)}$$

and recall from Brown that $\underline{P}^r$ has the following marginal properties

$$\int_{x \in X^k} P^r(x)dx = P_k(x) \qquad k = 1, \ldots, m \qquad \text{(IV-B-5)}$$

Substituting for $\underline{P}^r$ from Eqn. (IV-B-3) and Eqn. (IV-B-4) (with proper assumptions to give interchange of limits, integrals, and products)

$$\int_{x \in X^k} P^r(x)dx = \lim_{j\to\infty} \int_{x \in X^k} P^j(x)dx$$

$$= \lim_{j\to\infty} \frac{P_k(x)}{g_k^j(x)} \; g_k^{j-1}(x) = P_{-k}(x) \qquad k = 1, \ldots, m$$

and

$$\lim_{j\to\infty} \frac{g_k^{j-1}(x)}{g_k^j(x)} = 1 \qquad \text{(IV-B-6)}$$

Thus, it is necessary only to compute the iterative definitions of g_k^j. An example may clarify the role that the g_k^j functions play. Let $P_o(x_1, x_2, x_3)$ be an a priori distribution over $X \subset R^3$. Let the two-dimensional (n=2) lower-order distributions (m-2) be (where P with no indices denotes the marginal distribution of the indicated arguments)

$$P_1(x) = P(x_1, x_2) \qquad X^1 = (x_1, x_2)$$

$$P_2(x) = P(x_2, x_3) \qquad X^2 = (x_2, x_3)$$

Then

$$g_1^j(x) = 1 \qquad j = 1, 3, \ldots$$

$$g_2^j(x) = P(x_2) \qquad j = 2, 4, \ldots$$

and

$$P^r = P^j = P(x_1, x_2)P(x_3/x_2) \qquad j = 2, 3, \ldots$$

Note that any a priori distribution is allowed and does not affect the final result. The effect of the g_k^j functions in this simple example is to change the marginal distribution into a conditional distribution. Chow's[45]

approximate scheme for learning conditional dependencies is analogous (allowing conditioning on one variable only) but the use of the g_k^j functions allows one to determine the structure of the problem for any set of lower-order distributions (possibly in a theoretical sense only, as the computations may become unwieldy especially if the lower-order distributions change).

In summary, we have shown how an iterative procedure for approximating probability distributions is a mathematical model for learning conditional dependencies such as those found between Kilmer's[44] STC-RETIC modules. The reduced formulae developed here require only integration and multiplication and no powers (as in the original scheme). These iterative formulae develop only the conditional dependencies and do not depend on measurements that are independent. That is, if X^k and X^q are nonoverlapping, independent input sets, then the integration set to compute g_k^j need not include X^q. The resulting approximation formula is a product which implies independent measurement sets. Thus, they form an appropriate set of mixture probabilities discussed in the last section.

IV C. Specification of First-Level Decision Structure

At this point, we have specified a set of m state vector representations of the input signal, each state vector having dimension n, and a set of a posteriori distributions for the probability of each state vector, given one classification C_ℓ, $\ell = 1, \ldots, r$. We wish to decide, on the basis of this information (and possibly other information which needs to be specified), the appropriate subset of state vectors that best represents the pertinent features in the input signal. We can write a general formula to compute r numbers to decide between the different classifications (hypotheses), including the Bayesian approach developed in the last two sections.

$$S_\ell = \sum_k f_k(p_{k\ell}) \qquad \ell = 1, \ldots, r \qquad \text{(IV-C-1)}$$

where f_k is a monotonic, nondecreasing, continuous function and $P_{k\ell}$ is $P_k(C_\ell/X^k)$, the a posteriori probability of class C_ℓ given the input set X^k. This formula includes a large number of likelihood functions. We will discuss these different formulations and relate them to the specific problem of speech recognition.

The usual choice of monotonic functions of the probabilities is the natural logarithm, which converts a product of independent probability distributions, as discussed in the previous section, into a summation. Since the function is monotonically nondecreasing, a decision test based on the probabilities alone will have similar results for a function of those probabilities. Another function of this type is discussed

in Kilmer.[41] There, the purpose was to emphasize probabilities that were different from a uniform value, 1/r. Thus, if a given $P_{\ell k}$ was significantly greater than or less than $\frac{1}{r}$, the f function would tend to emphasize this particular probability.

The formulation (Eqn. IV-C-1) also allows several types of cost factors to be included in the decision quantities. Various cost factors are discussed in the literature. One of the most pertinent to the study is an information measure that is related to the amount of information in the a posteriori distribution $p(C_\ell)$, given X^k, $\ell = 1, \ldots, r$. Here the implication is that a module input should be considered very strongly in the decision if there is a significant peaked distribution among the various categories. Another possible interpretation of a cost function of the input X^k, especially for suboptimal systems operating in noisy environments, is a quality measure which could be determined in two ways: first, in terms of the distance from the cluster centers of the input variable. This would indicate whether the input were quite far from the majority of inputs seen previously with respect to the given set of learned categories (where we are not concerned with unknown or new input classes). This type of cost factor would indicate that low values of a posteriori probability have less influence, especially in the case of an insufficient or small number of training patterns. This is so because the majority of known probability distribution estimation techniques give much worse estimates of the tails of a distribution (events with low probability of recurrence) than they do of more densely populated modes. Another type of quality measure based on the physical characteristics or measurements (low signal-to-noise ratio of input, extremely high background interference, etc.)

would lessen the effect of noisy inputs. These types of cost factors can easily be incorporated into the formulation (Eqn. IV-C-1).

The third thing that may be included in this formulation is prior distributions. Lainiotis only used the prior distributions as thresholds for comparison of likelihood ratios that he generated. In speech it is well known that successive speech segments are highly dependent (redundancy of about 33 percent); hence there is much information in the probability of a given segment, given the last decision or classification of the preceding segment. Thus, we must augment Bayes' formula that was stated in Section I-E to include this conditional probability.

$$P(C_\ell/X^k, X^{k-1}) = \left[P(X^k/C_\ell)\right] \Big/ \left[(P(C_\ell/C_{\ell-1})P(C_{\ell-1}/X^{k-1})\right] \qquad \text{(IV-C-2)}$$

This should be incorporated in such a way that when the a posteriori probabilities computed on the present input do not contain sufficient information to give a reliable estimate of the present category, the conditional distributions should be used. Even with the different interpretation given to cost factors and prior distributions it is possible to formulate a recognition problem for a restrictive speech signal within the Bayesian framework, as discussed in the previous two sections. However, there are several events, especially for suboptimal systems operating in noisy environments, that will have a probability matrix $P = (P_{\ell k})$, $\ell = 1, \ldots, r$; $k = 1, \ldots, m$, which does not give acceptable decisions using Eqn. IV-C-1. This can be due to conflict between modules having high probabilities for one class and other modules having high probabilities for another class. This situation involving

"outliers," as discussed previously, can occur because of: (1) inappropriate assumptions; (2) presence of noise in the input that is very much signal-like (white noise that looks like a fricative, sinusoidal inputs that look like vowels or nasals, etc.), or (3) a dichotomization of inputs; that is one module may have the same input for two completely different classes of the input signal. An example of this would be during a nasal, when a high-frequency-bandpass filter might have a strong formant that looks vowel-like, whereas the presence of no signal in other filters and low energy of pitch frequency component would indicate that this signal interval is not a vowel. This type of correlation, of course, should be incorporated by the Lewis-Brown[42,43] approximation technique, but it may not be a sufficient mechanism.

Wainstein and Zubakov[63] have used the central limit theorem with respect to likelihood ratio formulae, such as Eqn. IV-C-1 for the following reasons: given that the individual terms (probabilities, likelihood ratios) are independent events and given certain restrictions on the tails of the distribution of these events (that they are well behaved and go to zero sufficiently fast as the value of the event goes to infinity), the distribution of the sum tends toward the normal distribution. As is well known, this is an asymptotic property, but it illustrates two things: (1) convergence to a definite value, and (2) this value is not a local minimum, since the asymptotic distribution is unimodal (these theorems have been proven for a larger class of distributions than the normal, but with similar convergence and unimodal properties). Thus one can expect a stable rule for finding a maximum value with a guaranteed convergence property. The problem that occurs with low probability outliers is that it will take a large number of terms in the summation

to counteract its effects. In our situation, where there is such a mixture of probability distributions and a large number of possibilities for generating such outliers, one cannot sit back and hope that they will only occur with a small probability. Several theorems have been proven (Hertz[70]) where the tail behavior of the even distributions has been relaxed by eliminating outliers and still obtaining the central limit theorem. This of course is intuitively the correct thing to do in order to maintain the convergence and unimodel properties.

The discussion of causes for outliers' occurrence leads one to consider two approaches: One is to use a Bayes decision formulation but compute a larger dimension probability distribution, possibly over the entire m X n dimensional space. The discussion of the first chapter has indicated empirical objections to this approach. With respect to the discussion in the first two sections of this chapter, the module concept can be justified by stating that the filtering representation scheme presented in Chapter Two is better matched to the natural dimensionality of each feature and thus is better able to eliminate unwanted signals and noise and thus to isolate individual features. Second, as pointed out by Groner, too many inputs to a suboptimal design Bayes decision network very often add noise and thus degrade the overall classification performance. Thus, it would seem very natural that the first level of decision logic would be to extract the features as separate entities and then, on the basis of this extraction, look for the interrelationships and more detailed properties of the features.

The other possible decision structure is to allow interconnections between the modules that allow lateral passage of gross information. Kilmer[41] has considered this problem and has related the S-RETIC modal computations to nonlinear summative schemes such as Eqn. IV-B-1. He shows that, based on three symmetries that are assumed for such systems, "S-RETIC computes a mode [detection/classification] function, F, that no S-RETIC net without α and δ [lateral] connections but with nonlinear summative output scheme could compute even though it is allowed more equipment." The three symmetries that are assumed for these systems are as follows (note that the first two are typical for Bayesian schemes of the type discussed):

(1) We must be able to compute the same classification decision regardless of which module has the proper information. This is especially necessary in speech, as is evident in Figure 1, since the same module will not always have the appropriate classification information, especially when different speakers are expected to be using the system. Further, Figure 17 indicates how different processing schemes will isolate the pertinent information, dependent on the surrounding feature environment.

(2) The evaluation scheme must be the same for any classification decision (the computation of S_ℓ is independent of ℓ, $\ell = 1, \ldots, r$).

(3) Strength-of-effect symmetry. Given prior distributions, an average (summative) decision across the net and conflicting decisions, any two can overcome the third or any one can overcome the other two, whether the other two are in favor of the same classification or conflicting classifications. This symmetry states that the decision rule must give equal weight and operate in an equal fashion in judging the effects of these three possible situations that can occur simultaneously.

[The last symmetry requires the lateral communication, since Bayesian schemes have the first two symmetries (any of the formulae from Wainstein and Zubakov discussed previously) but when faced with the type of situation depicted in (3) will not operate in a consistent or appropriate manner.] To paraphrase Kilmer's statement, in light of the outlier situation, it is seen that the lateral communication is necessary to decide among the probability matrix and the prior distribution matrix, which modules should work in conjunction and be averaged together to determine the output and which should be considered outliers and eliminated. The S-RETIC algorithm is iterative and thus the intuitive arguments we are presenting here are intended for understanding rather than analysis, but the importance of Kilmer's statement about decision algorithms of this nature is that the structure must be implemented in this way or else the performance will suffer.

In the next section, we will specify a first-level recognition system that operates according to these principles and incorporates the S-RETIC type of decision logic. We will see how the state variable representation presented here can be incorporated in a dynamic real-time asynchronous decision network.

IV D. Proposed First-Level Recognition Block Diagram

The purpose of this study is to specify a mathematical model and a system block diagram that are tailored to the acoustical speech signal, rather than the converse. The deficiencies of state-of-the-art solutions derived by means of a Bayes minimum-risk criterion have been discussed: It is necessary to use a restricted speech model; it is very difficult to implement the nonlinear estimation filters that are required; a high dimensionality is required because of the complicated interrelationships of the speech signals. Even the optimal filters' outputs will be dependent, in a probabilistic sense; the mixture probability formula allows the possibility of adding together nonsimilar waveforms, based solely on the learned probability of presence. Adding to these difficulties those that have been discussed for suboptimal solutions, which can give rise to outlier probabilities for particular classes, one is left with a very negative picture. There are several other requirements of a recognition system that are difficult to include in a Bayes formulation, which will be mentioned here to help specify the recognition system;

(1) Significance of the marked change--The segmentation marks that are derived from the inherent signal characteristics must be monitored with respect to past occurrences of the speech signal to determine whether the marked change is due to noise (parity error...), another energy peak entering the filter bandwidth, the actual start of a new feature, a change from one feature to another, or the finish of a feature.

(2) Correlation with overall system behavior--Each module decision must be compared with all other modules to determine if this is a new feature, whether an energy peak has moved from one filter to another, or whether (one of) the predominant feature(s) has finished.

(3) Precisely controlled features--The main criterion for classifying a pertinent feature is whether it is repeatable for different speakers and contexts, whether it is a transition of prescribed form, and whether the terminal state of the transition is predictable before the end in case the segment is terminated.

The Bayesian formulation of course has a different philosophy toward marked changes, in that they are assumed to be a true segment and the mixture probability formula is used to decide on the actual significance. Since in a speech recognition system these marked changes also have linguistic meaning in higher levels (determining the consonant/vowel relationships, directing higher-level analysis, ...), there must also be a decision on their validity. As is well known, the overall system behavior must not be degraded in allowing individual modules to make classification decisions. It has been demonstrated that the S-RETIC will work in a correlated fashion as a total system rather than m individual systems, each screaming for its own way (as in pandemonium machines). This is a very serious requirement which will not necessarily be satisfied by using a simple mixture formula. The particular method of training the

classification network is well specified in our Section I-C and the work of Rupert[20]. We can see that the requirements of precisely controlled features are very pertinent to determining a consistent system performance. The work of Houde[17] also indicates that recognizing transients can be performed because of the consistent and precise form of articular transitions. Since it appears from the preprocessing pictures that these transitions also exist in the acoustical waveform, we can see that this is a desirable and necessary requirement for efficient recognition.

As was mentioned previously in the discussion of the filter bank, the fixed-frequency filters that were used are not tailored to actual speech characteristics, especially during frequency-transition epochs. As can be seen from the three requirements stated above for the recognition system, there might also be difficulties for specific systems that have set filter bandwidths, in that energy peaks can move across filter boundaries. Depending on the skirt response of the filter it may be very difficult, for a particular filter, to distinguish an energy peak which moves into a filter from one that simply begins in that filter. For this reason and to avoid a very complicated classification system which must make these additional decisions, we should make use of the time-varying tracking filters discussed in Section II-D. We will outline a procedure for their use in conjunction with the classification system. First we make the assumption that at any one given instant of time the filter bank is constructed such that there is at least one filter that isolates the pertinent feature information (here the use of the word "filter" indicates the derivative calculations as well as the actual bandpass filter operation, since the combination of filtering and differentiation is sometimes

required to isolate the desired feature). Given that assumption, we can then specify a tracking filter, shown in Figure 12, in the following way. At a marked change, we make a classification of the overall system input (i.e., each module decision is calculated and then an overall global decision is arrived at from these local decisions). Then, based on this decision, selected filters which have the pertinent features are activated to start tracking. The estimates of frequency and bandwidth are used, as indicated in Figure 12, to modify the input signal further to emphasize the particular pertinent features. Thus, other formants entering this particular filter will not affect the tracking filter output. Also, it will be possible to allow the tracking filter to operate across the filter bank boundaries. The combination of this tracking filter with the fixed-frequency filter bank will then lock on certain features and follow them throughout their duration, emphasizing the chracteristics which may be needed for higher-level classification.

These requirements allow us to specify a recognition and preprocessing structure which matches the nature of the speech signal and allows higher-level linguistic classification. This structure is shown in Figure 27. The wideband speech signal is processed by the overlapping filter bank. Each filter output is operated on by a measurement device similar to that described in Section II-E. The inherent signal changes are detected to give derivative segmentation indicators. The measurement outputs from $A1_i$ go to $A2_i$, which is the acousteme class selection. Here the stored precisely controlled feature information is compared to the input and local class decisions are made. Based on these local class decisions, the outputs of $A2_i$ corresponding to degree of presence (DOP) vectors shown in Figure 6 are compared with the derivative segmentation

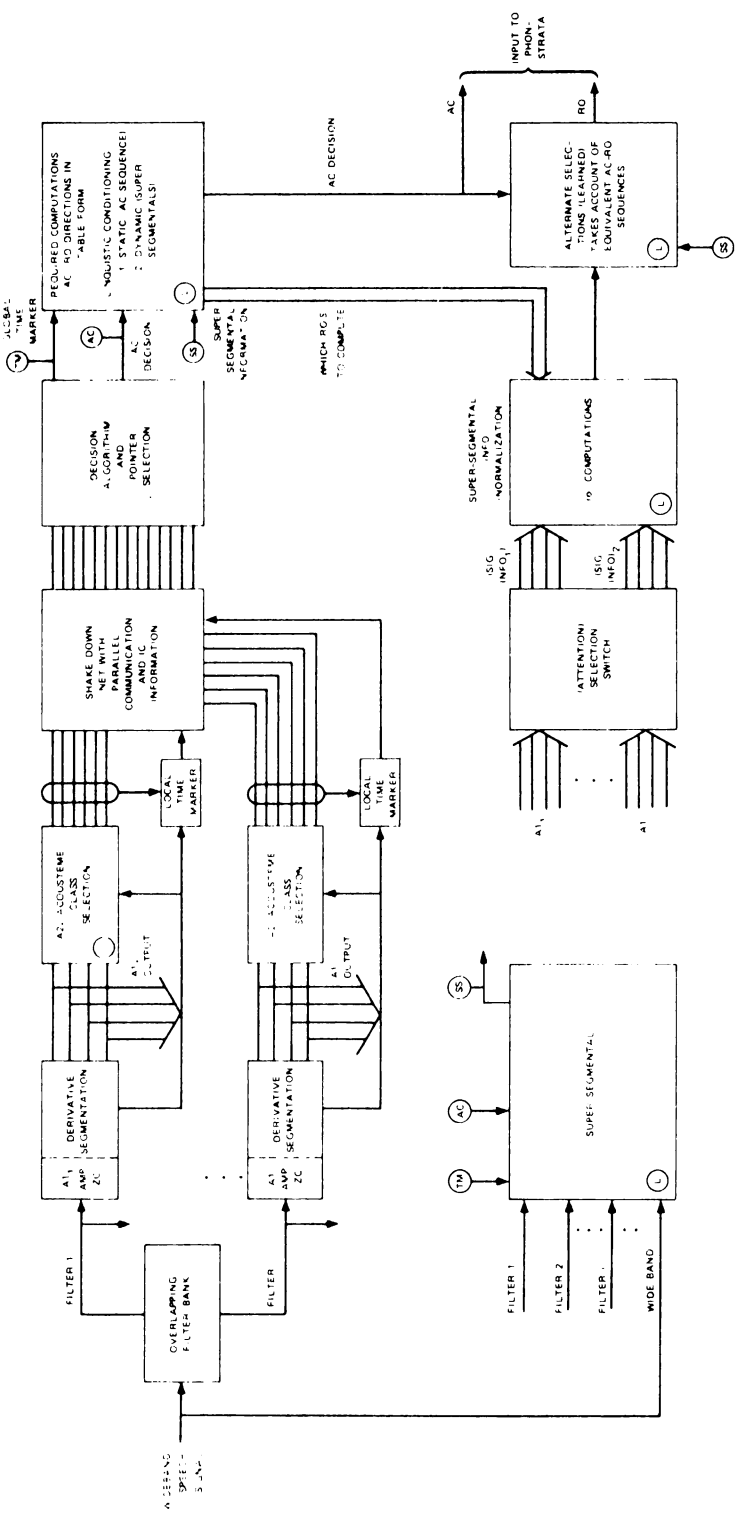


FIGURE 27 BLOCK DIAGRAM FOR FIRST LEVEL OF ASYNCHRONOUS REAL TIME SPEECH RECOGNITION SYSTEM

information to give a local time marker. This local time marker incorporates the predictive segmentation on the state vectors, which correspond to physical clusters, and also probability changes in the DOP vectors, corresponding to linguistic clusters. The two types of time markers are integrated to give a local time marker for the i^{th} module. Based on the table of DOP vectors and the local time marker information, a shakedown net with parallel communication and initial condition information determines an overall global decision and also decides which modules contain the pertinent acoustical features. This decision network also computes a global time marker, which is a segmentation point in the input acoustical signal and also an acoustical class decision. Based on this information, a library is searched for the required computations to determine further classification. The decision and pointer information is also used with an attention-focusing device (called the attention selection switch) which selects pertinent filters and performs the tracking operation on these filters, based on the bandwidth and frequency information . The lower box, which is discussed in Chapter III, computes suprasegmental features such as pitch contours and stress placement, based on the filter outputs, the wideband signal, global time markers and ac decisions. The suprasegmental information is used in higher-level decisions shown here in the terminology of Rupert's Relative Oppositions, to give a first-level output which is an acousteme class label and a list of RO's to completely determine the first-level output and input to the next level.

The block diagram gives an integrated philosophy of speech recognition, rather than a particular implementaton. The requirements that were specified throughout this report have led to this type of structure. The particular names and implementation, although specific, indicate the

necessary sort of structure that is required. We have tried to indicate from as much experimental evidence as is available the feasibility and practicality of this particular approach. However, nothing short of a full-scale implementation and testing will actually prove its worth.

V CONCLUSIONS AND RECOMMENDATIONS FOR FURTHER STUDY

This study has specified a segmentation procedure which also is the first level of analysis of a total system intended to give a commanded response based on recognition of the speech input. Section I-A contains a summary of the arguments presented which lead to this specification. The deficiencies of current segmentation procedures in trying to mark signal epochs in the speech acoustical signal exist because they are not adequately tailored to the complex speech acoustical signal. The complicated interrelationships of pertinent linguistic features in the acoustical signal necessitates a more sophisticated procedure; that is, one which first isolates individual, primary features and then further processes them to determine secondary features, defined as perturbations of the primary features. The general purpose of this study is to specify a first level recognition system which isolates individual formants, then, guided by a syntactic and semantic structure, computes the necessary measurements. The purpose of this thesis is to describe a segmentation procedure within this general specification which not only specifies basic units for recognition but also gives an adequate description of the complicated speech signal. Further, the segmentation procedure that identifies lower units will direct the higher levels of decoding so that the search space is kept within practical bounds.

By formant, we mean the resulting time waveform for one cavity of the vocal tract excited by glottal pulses, or frication noise. Thus, vowels and continuants may have three predominant formants; some fricatives have one formant, and silent portions have none. Each formant is

assumed to be generated by a second order time-varying differential equation driven by either a pitch pulse or white noise source. First order differential equations are derived for four state variables. In terms of these state variables, a procedure for isolation and accentuation of the formants present in the acoustical signal at any one time is given which:

(1) Uses a wideband overlapping fixed-frequency filter bank and real-time processing to attempt to isolate individual formants (the criterion is that at least one filter processing combination sufficiently isolates each formant).

(2) Derives gross measurements of the state variables from the observed acoustical signal.

(3) From these measurements, identifies a model for the current state of the acoustical signal and thus specifies how many formants are present and which estimation formulae are appropriate.

(4) Selects the filter/processing combination which best isolates each formant.

(5) Tailors an estimation procedure to track each formant and give reliable estimates of the state variables to be used for furhter analysis of the speech signal.

(6) Uses a predictive comparison to determine when a given model is no longer valid.

Experimental evidence shows that this procedure can achieve the
ation of the pertinent features from the nonessential (possibly
elated) portions of the speech acoustical signal and also isolation
background noise. Also, the processing can reduce speaker depen-
e (examples given in Chapter 2).

This study also specifies a real-time procedure, which is a combin-
n of analog and digital processing, to accomplish the feature rep-
ntation. It involves:

(1) Estimation of the parameters of a dynamic time-varying differential equation model for single formants; and

(2) Use of these estimates to isolate and accentuate the single formants.

representation shows the varying parameters of frequency, bandwidth,
mplitude and gives a compact (low memory requirement) representation
nly the information necessary for further linguistic processing.
levelopment of this real-time procedure is partially mathematical
d in part on an attempt to relate the results to theoretical studies
timation of time-varying parameters), but it is mainly empirical,
se of the inadequacy of any mathematically tractable model to show
omplete, complicated nature of the speech acoustical signal.
Linguistic theory is needed in addition to this acoustical signal
ssing technique to give a proper learning criterion for the pattern-
nition portions of the algorithm. The inadequacies of existing
etical linguistic studies to properly account for complicated acoustical
l properties are pointed out, and an alternative theoretical
work is described to incorporate these properties.

This study has resulted in the block diagram specification of a first-level recognition system. It is evident that many of the results can be checked out by implementation and testing of the proposed system. Existing computer analysis programs can be used to design and check out the various parts of the recognition system. Further theoretical work is suggested by much of the discussion in this study. The adequate analysis of the transient response of nonlinear time-varying estimation filters (extensions of Kalman-Bucy filters), can possibly be achieved through use of the quasi-steady-state formulation and prebandwidth function definitions given in Chapter II. The detection/estimation problem formulated in Chapter IV may possibly be treated through a combination of these techniques for the particular models developed here to give significant theoretical results in the relatively new areas of nonlinear filtering and time-varying signal recognition. Also the formulae for deriving bandwidth estimates can be very useful for investigation of time-varying systems, and the linguistic studies that have been specified can be continued further for natural spoken American English to define and evaluate linguistic elements that are more closely related to the acoustical signal. The effective use of suprasegmental features such as pitch and intonation has only been suggested in this study but will surely be useful for further progress in recognition of connected speech. The techniques developed here for representing the acoustical signal give a very practical method for computer analysis of (suprasegmental) features such as stress or intonation patterns. The integrated approach, through linguistic and communication theories, gives methods to attack the complicated problem of specifying the interactions and effects of these features on all levels of linguistic element recognition.

BIBLIOGRAPHY

BIBLIOGRAPHY

1. C. F. Hockett, "A Manual of Phonology, Memoir 11," Internatl. J. Am. Linguistics (Waverly Press, Inc., Baltimore, Maryland, 1955).

2. G. E. Peterson and H. L. Barney, "Control Methods Used in a Study of the Vowels," J. Acoust. Soc. Am., Vol. 24, pp. 175-184 (1952).

3. Louis J. Gerstman, "Classification of Self-Normalized Vowels," IEEE Trans. Audio Electroacoust., Vol. AU-16, No. 1, pp. 78-80 (March 1968).

4. R. C. Potter, G. A. Kopp, and H. C. Green, Visible Speech (D. Van Nostrand Co., Inc., Princeton, New Jersey, 1947).

5. F. S. Cooper, P. C. Delattre, A. M. Liberman, J. M. Borst, and L. J. Gerstman, "Some Experiments on the Perception of Synthetic Speech Sounds," J. Acoust. Soc. Am., Vol. 24, pp. 597-606 (1952).

6. H. Yilmaz, "A Program of Research Directed Toward the Efficient and Accurate Machine Recognition of Human Speech: A Theory of Speech Perception," Final Report No. 2, Contract NAS 12-129, Arthur D. Little, Inc., Cambridge, Massachusetts (November 1967).

7. D. R. Reddy, "Phoneme Grouping for Speech Recognition," J. Acoust. Soc. Am., Vol. 41, No. 5, pp. 1295-1300 (1967).

8. C. C. Tappert, N. R. Dixon, D. H. Beetle, Jr., and W. D. Chapman, "A Dynamic-Segment Approach to the Recognition of Continuous Speech: an Exploratory Program," Final Report, Technical Report RADC-TR-68-177, Contract F30602-67-C-0123, Project 4027, International Business Machines Corp., Systems Development Division, Research Triangle Park, North Carolina (June 1968).

J. Gazdag, "A Method of Decoding Speech," Technical Report 9, AF Grant 7-66, University of Illinois, Urbana, Illinois (June 1966).

Z. S. Harris, Structural Linguistics (University of Chicago Press, Chicago, Illinois, 1961).

R. Jakobson, C. Bunnar M. Fant, and M. Halle, Preliminaries to Speech Analysis (MIT Press, Cambridge, Massachusetts, 1965).

N. Chomsky and M. Halle, The Sound Pattern of English (Harper and Row, New York, New York, 1968).

D. G. Bobrow, A. K. Hartley, D. H. Klatt, "A Limited Speech Recognition System II," BBN Report No. 1819, Final Report, Contract NAS 12-138, Bolt Beranek and Newman, Inc., Cambridge, Massachusetts (April 1969).

14. D. R. Hill, "An ESOTerIC Approach to Some Problems in Automatic Speech Recognition," Int. J. for Man-Machine Studies, Vol. 1, No. 1, (1969).

15. L. R. Focht, "Single Equivalent Formant Extractor System," Contract NAS-12-582, Philco-Ford Corp., Blue Bell, Pennsylvania (November 1967).

16. S. E. G. Öhman, "Coarticulation in VCV Utterances: Spectrographic Measurements," J. Acoust. Soc. Am., Vol. 39, No. 1, pp. 151-168 (January 1966).

17. R. A. Houde, "A Study of Tongue Body Motion During Selected Speech Sounds," SCRL Monograph Number 2, Air Force Office of Scientific Research Grant No. AF-AFOSR-1252-67, Speech Communications Research Laboratory, Inc., Santa Barbara, California (August 1968).

18. S. M. Lamb, "Prolegomena to a Theory of Phonology," Language, Vol. 42, No. 2, pp. 536-573 (1966).

19. I. B. Thomas, "The Significance of the Second Formant in Speech Intelligibility," Technical Report 10, Contract AF-33(615)-3890, University of Illinois, Urbana, Illinois (July 1966).

20. W. P. Rupert, "A Representation for the Information-Carrying Units of Natural Speech" (Ph.D. Thesis, Montana State University, Bozeman, Montana, April 1969).

21. J. F. Hemdal and G. W. Hughes, "A Feature Based Computer Recognition Program for the Modeling of Vowel Perception," Proc. Symp. on Models for the Perception of Speech and Visual Form, AFCRL (November 1964).

22. R. W. Shafer and L. R. Rabiner, "A System for Automatic Formant Analysis of Voiced Speech," Bell Telephone Laboratories, Inc., Murray Hill, New Jersey.

3. J. L. Flanagan, Speech Analysis Synthesis and Perception (Academic Press, Inc., New York, New York, 1965).

J. L. Flanagan, "A Difference Limen for Vowel Formant Frequency," J. Acoust. Soc. Am., Vol. 27, No. 3, pp. 613-617 (May 1955).

W. B. Davenport, Jr., and W. L. Root, An Introduction to the Theory of Random Signals and Noise (McGraw-Hill Book Company, Inc., New York, New York, 1958).

R. Kalman and R. S. Bucy, "New Results in Linear Filtering and Prediction Theory," Trans. ASME Ser. D, J. Basic Eng., Vol. 83, pp. 95-108 (March 1961).

M. Halle and K. N. Stevens, "Speech Recognition: a Model and a Program for Research," IEEE Trans. Info. Thy., Vol. IT-8, pp. 155-159 (1962).

R. M. Lerner, "Representation of Signals," Chapter 10 in Lectures on Communication System Theory, E. J. Baghdady, ed., pp. 203-242 (McGraw-Hill Book Co., Inc., New York, New York, 1961).

J. R. Hanne, "Formant Analysis," Report 12, NR 049-122, Contract Nonr 1224(22), Communication Sciences Laboratory, University of Michigan, Ann Arbor, Michigan (March 1965).

M. Lecours and J. J. Sharkes, "Adaptive Spectral Analysis for Speech Sound Recognition," IEEE Trans. Audio and Electroacoust., Vol. AU-16, No. 4, p. 523 (December 1968).

E. J. Baghdady, "Analog Modulation Systems," Chapter 19 in Lectures on Communication System Theory, E. J. Baghdady, ed., pp. 439-555 (McGraw-Hill Book Co., Inc., New York, New York, 1961).

B. J. Leon and D. D. Weiner, "The Quasi-Stationary Response of Linear Systems to Modulated Waveforms," National Science Foundation Contract GP-581, Purdue University, Lafayette, Indiana (May 1964).

J. D. Duncan and L. E. Cannon, "Investigation of Pre-Detection Filtering Techniques," Quarterly Report, Contract DA28-043AMC-01548(E), Electronics Research Lab, Montana State University, Bozeman, Montana (December 1965).

J. W. Cooley and J. W. Tukey, "An Algorithm for the Machine Calculation of Complex Fourier Series," Mathematics of Computation, Vol. 19, No. 90, pp. 297-301 (April 1965).

H. L. Resnikoff and G. A. Sitton, "Linguistic Segmentation of Acoustic Speech Waveforms," paper presented to the Seventy-Fifth Meeting of the Acoustical Society of America, Ottawa, Canada, 21-24 May 1968.

E. C. Cherry and V. J. Phillips, "Some Possible Uses of Single Sideband Signals in Formant-Tracking Systems," J. Acoust. Soc. Am., Vol. 33, No. 8, pp. 1067-1077 (August 1961).

E. Peterson, "Frequency Detection and Speech Formants," J. Acoust. Soc. Am., Vol. 23, pp. 668-674 (1951).

R. W. A. Scarr, "Zero Crossings as a Means of Obtaining Spectral Information in Speech Analysis," IEEE Trans. Audio and Electroacoust., Vol. AU-16, No. 2, pp. 247-255 (June 1968).

G. H. Ball and D. J. Hall, "A Clustering Technique for Summarizing Multivariate Data," Beh. Sci., Vol. 12, No. 2, pp. 153-155 (March 1967).

H. P. Zeiger, "Cascade Synthesis of Finite-State Machines," IEEE Conf. Record on Switching Oct. Thy. and Log. Design, pp. 45-51 (October 1965).

41. W. L. Kilmer, W. S. McCulloch, and J. Blum, "Some Mechanisms for a Theory of the Reticular Formation," Final Scientific Report, Air Force Office of Scientific Research Grant AF-AFOSR-1023-66, Division of Engineering Research, Michigan State University, East Lansing Michigan (February 1967).

42. P. M. Lewis, "Approximating Probability Distributions to Reduce Storage Requirements," Information and Control, Vol. 2, pp. 214-225 (1959).

43. D. T. Brown, "A Note on Approximations to Discrete Probability Distributions," Information and Control, Vol. 2, pp. 386-392 (1959).

44. W. Kilmer, W. McCulloch, and J. Blum, "Embodiment of a Plastic Concept of the Reticular Formation," in Proc. Symposium on Biocybernetics of the Central Nervous System (Little and Brown, Boston, Massachusetts, 1968).

45. C. K. Chow and C. N. Liu, "An Approach to Structure Adaptation in Pattern Recognition," IEEE Trans. Sys. Sci. and Cyber., Vol. SSC-2, No. 2, pp. 73-80 (December 1966).

46. D. Gabor, "Theory of Information," J. Inst. Elec. Engrs., Part III, Vol. 93, pp. 429-457 (November 1946).

47. E. C. Titchmarsh, Introduction to the Theory of Fourier Integrals (Clarendon Press, Oxford, England, 1948).

48. R. Deutsch, Nonlinear Transformations of Random Processes (Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1962).

49. E. J. Kelly and I. S. Reed, "Some Properties of Stationary Gaussian Processes," Technical Report No. 157, Massachusetts Institute of Technology, Lexington, Massachusetts, 1957.

50. N. Abramson, "Nonlinear Transformations of Random Processes," IEEE Trans. Info. Thy., Vol. IT-13, No. 3, pp. 502-505 (July 1967).

51. E. A. Guillemin, Theory of Linear Physical Systems (John Wiley & Sons, Inc., New York, New York, 1963).

52. J. D. Bruce, "Discrete Fourier Transforms, Linear Filters, and Spectrum Weighting," IEEE Trans. Audio Electroacoust., Vol. AU-16, No. 4, pp. 495-499 (December 1968).

53. A. A. Kharkevich, Spectra and Analysis (Consultants Bureau, New York, New York, 1960).

54. W. R. Kincheloe, Jr., "The Measurement of Frequency with Scanning Spectrum Analyzers," Technical Report No. 557-2, Contract AF30(602)-2398, Systems Techniques Laboratory, Stanford Electronics Laboratories, Stanford, California (October 1962).

55. R. Alter, "Utilization of Contextual Constraints in Automatic Speech Recognition," IEEE Trans. on Audio and Electroacoustics, Vol. AU-16, No. 1, p. 6 (March 1968).

56. D. R. Reddy and A. E. Robbinson, "Phoneme to Grapheme Translation of English," IEEE Trans. on Audio and Electroacoustics, Vol. AU-16, No. 2, p. 6 (June 1968).

57. R. Ash, Information Theory (Intersciences Publishers, New York, New York, 1967).

58. N. Chomsky, Syntactic Structures (Janua Linguarum, Series Minor, No. 4, Mouton Publishing, The Hague, Netherlands, 1957).

59. D. Lainiotis, "Sequential Structure and Parameter-Adaptive Pattern Recognition--Part One: Supervised Learning," IEEE Trans. on Information Theory, Vol. IT-16, No. 5, pp. 548-556 (September 1970)

60. H. J. Kushner, "Dynamical Equations for Optimal Nonlinear Filtering," J. Differential Equations, Vol. 3, pp. 179-190 (1967).

61. G. Kalinapur and C. Striebel, "Stochastic Differential Equations Occurring in the Estimation of Continuous Parameter Stochastic Processes," Technical Report No. 103, University of Minnesota, Minneapolis, Minnesota (September 1967).

62. T. Kailath, "A General Likelihood-Ratio Formula for Random Signals in Gaussian Noise," IEEE Trans. on Information Theory, Vol. IT-15, No. 3, pp. 350-361 (May 1969).

63. L. A. Wainstein and V. D. Zubakov, Extraction of Signals from Noise (translated by R. A. Silverman) (Prentice-Hall Book Company, New Jersey, 1962).

64. M. W. Byrd, "Asymptotic Convergence of Nonlinear Continuous Time Filters, Ph.D. dissertation for Michigan State University, East Lansing, Michigan (1969).

65. G. F. Groner, "Statistical Analysis of Adaptive Linear Classifiers," SEL Report No. SEL-64-026 (TR No. 6761-1), Stanford Electronics Lab, Stanford University, Stanford, California (April 1969).

66. E. Sapir, "Sound Patterns of Speech," Language, Vol. 1, pp. 37-51 (June 1925).

67. K. L. Pike, Phonemics: A Technique for Reducing Languages to Writing (University of Michigan Press, Ann Arbor, Michigan, 1947).

68. T. H. Crystal and L. Ehrman, "Design and Applications of Digital Filters with Complex Coefficients," IEEE Trans. on Audio and Electroacoustics, Vol. AU-16, No. 3, pp. 315-320 (September 1968).

69. E. Parzen, Stochastic Processes (Holden-Day, Inc., San Francisco, California, 1962).

70. E. S. Hertz, "On Convergence Rates in the Central Limit Theorem," Ann. Math. Stat., Vol. 40, No. 2, pp. 475-479 (April 1969).

71. P. J. Donoghue, "System Identification by Bayesian Learning," Ph.D. thesis, Department of Electrical Engineering, Michigan State University, East Lansing, Michigan (1968).

APPENDICES

APPENDIX A

Description of Sapir's Pseudo-Language

The choice of a data base is governed by two conflicting criterion. In order to allow computer analysis, it must be of a limited size, but also, the data base must be representative; i.e., it must show:

(1) phonetic patterns determined by some consistent rules as are normally found in natural languages

(2) phonetic "slurrings" by native speakers

(3) several speakers pronunciation, including male and female and various accents

(4) long phrase environments with stress and intonation patterns.

It is very difficult to find such a data base among common American English because of the assimilation of words from other languages and the many dialects that exist. For that reason, we have chosen a pseudo-language from which we will draw our experimental utterances. These linguistic forms were constructed by Donald Stark from data suggested in Edward Sapir[68] "Sound Patterns for Speech", (June, 1925) (See Pike[67] p. 156). This data was intended to serve as an illustration of a classical phonetic analysis and, as such, points out several of the more difficult problems that would be encountered by an ASR system operating on a natural language. The phonetic chart for this pseudo-language is shown below.

Phonetic Chart for Sapir's Language B

	a	ɛ	e	i	u	o	ɔ
Vocoids	(a·)	(ɛ·)	(e·)	(i·)	(u·)	(o·)	(ɔ·)
Semi-Vocoids	(?)	h	(w)	(y)	(l)	m	n
Non-Vocoids	p	t	k				
Stops	(p^h)	(t^h)	(k^h)				
	b	d	g				
Fricatives	(f)	(θ)	s	(x)			
	v	d	z	g			

Several rules, with regard to conditional variants for this language, were proposed by Sapir:

(1) long vowels (denoted by v·) can arise only when the syllable is opened and stressed

(2) The glotal stop (?) is not an organic constant, but, as in North German, an attack of initial vowels. This rapid onset is lost in mid-utterance position

(3) W and y are merely semi-vocalic developments of u and i that correspond to a glide between adjoining vowels

(4) L arises merely as a dissimilated variant of n

(5) Aspirated (denoted by c^h) p, t, and k are characteristic of this pattering at the end of the word. It is a reverse of the American-English habit

(6) F, Θ, and x similarly arise from the unvoicing of final v, d, and g. Z and s also alternate in this way, but there is a true s besides.

The linguistic forms are given below. The English "translations" are merely to give an indication of how the utterances are composed from the shorter forms by typical natural language rules.

Table A-2. Experimental Linguistic Forms from Sapir's Language B

1.	['hoo.sek^{n}]	'she is tired'	14.	['næ æ .ɔm]	'smoke'
2.	['hɔɔ.sex]	'bear'	15.	[' æ æ .om]	'working'
3.	['hɔɔ.zex]	'onion'	16.	[ɖw.'ath]	'horse'
4.	[bo.'ɡif]	'to answer'	17.	['?eg.n kh]	'bloody'
5.	[po.'ɡin]	'dish'	18.	['voo.eθ]	'man'
6.	['gaa.yph]	'round'	19.	[?um.'bif]	'four'
7.	['ɡaa.na]	'tarantula'	20.	['?e .go]	'fire'
8.	['?al.ba]	'white'	21.	[hil.'duu]	'cloudy'
9.	[?al.baa]	'knife'	22.	['taa.ha]	'square'
10.	['?æ l.bas]	'radish'	23.	['daa.os]	'water'
11.	['duu.e]	'two'	24.	['kaa. ɔph]	'acrid'
12.	['?el.bas]	'three'	25.	['haa]	'you'
13.	['?i .gɔ]	'even though'	26.	[pɔ.'ɡin]	'I wash'

27. [mæ k.'soth. 'al.ba] 'white stones'

28. ['zɔl.gi. um.'bif] 'four houses'

29. ['daa.oz. ɔ.'keθ] 'she carries water'

30. ['hɔɔ.zeɡ. 'duu.e] 'two onions'

31. ['voo.eɖ. ' æ æ .om] 'the man is working'

32. [ɖw. 'ath. um. 'biv.am] 'his four horses'

33. [po.'ɡil. 'gaa.yph] 'round dish'

All symbols that are used are taken from Pike and are given with their English equivalents in that book. The non-English sound, ɡ, is a voiced x or lambda, as it is commonly called by phoneticians. It is a sound common to several African languages. The glotal stop, ?, appears in some German dialects and also, may be a part of American-English pronunciation that has been neglected. The choice of speakers were four: one American male

(EH), one African male (BE), one African female (BA), one American female (MJ). These people had previous training in linguistics, but were not accomplished phoneticians, so that their pronunciations of these words were more natural, not academically stilted. Their linguistic background allowed them to comment on the exact nature of those pronunciations. This data does not have any r's or l's, as commonly found in American English. The result is that this is a slightly easier case because of the lack of complicated vowel glides. They do exist, however, between adjacent vowels so that this is not a completely academic data base. All utterances that are referenced in this paper are indexed by the following method: a seven-character label is assigned, the first two characters correspond to the utterance number given in the list above, a blank separates them from the next two characters, which are the initials assigned to the individual making the utterance, and the last two characters are assigned to the repetition number of the utterance (i.e., if the speaker has said the same utterance three times, then the last number will reflect this as 1, 2, or 3 -- for example, 16 EH 2 is the sixteenth utterance in the list, ɖwat^{h}, said by Earl Herrick (EH) and it is his second repetition.

This data base is sufficient for the study of very many questions arising in automatic recognition of natural speech.

(1) Automatic machine determination of existing phonetic patterns. There is a sufficient data set available to perform that experiment.

(2) The actual phonetic variations caused by different speakers of this basic data. This is the primary question that we are investigating in this research.

(3) Measurement of stress and intonation and other supra-segmental features.

(4) Determination of stress variation of vowels.

(5) Determination of syllable and word boundaries.

There are a total of 258 utterances on the analog tape from the four speakers with repetitions. A subset of these were chosen for an initial recognition experiment and these were subsequently filtered and digitized as described in Appendix B. That subset is:

4. [bo.' gif]
16. [ɖw at^{h}]
18. ['voo.eɵ]
19. [ʔum.'bif]
27. [mæ k.'sot^{h} . 'al.ba]
28. ['ʒɔl.gi um.'bif]
31. ['voo.eɖ .' æ æ . om]
32. [ɖw.'at^{h} . um.'biv .am]

APPENDIX B

Recording Apparatus Used to Collect Experimental Data

The recording of the experimental data on analog tape was performed on the apparatus shown schematically in Fig. B-1.

The subjects were seated in a quiet office and arranged around a microphone so that they were talking in a conversational-type atmosphere. The microphone used was an EV Model 654 Dynamic Non-Directional Microphone and it was input to a Type 122 Tektronix Pre-Amplifier, then to Channel 2 of an FR 1100 Ampex Recorder with 1/2-inch instrumentation tape at a speed of thirty inches per second. In order to minimize timing variations in recording/reproducing due to wow and flutter and possibly the use of different recorders with possibly different speed adjustments, a timing waveform was also recorded on another channel simultaneously. The timing circuit was provided by a Tektronix Type 114 Pulse Generator fed directly into Channel 5 of the Ampex FR 1100 Tape Recorder. The input was a 20-microsecond pulse repeated every 50 microseconds for a timing frequency of 20 kHz.* This analog tape was then processed using the equipment shown schematically in Fig. B-2 and B-3. First, the analog tape was marked with start pulses on Channel 7 indicating beginning points for the words that were to be processed and converted to digital samples. Six separate passes of the analog tape were used to get the various filtered combinations that were required. The use of the start pulses assured a uniform beginning point on all the

*The output variation in a 10-KC square wave, with reference to the input for a record/reproduce situation, was as much as 80 to 100 microseconds.

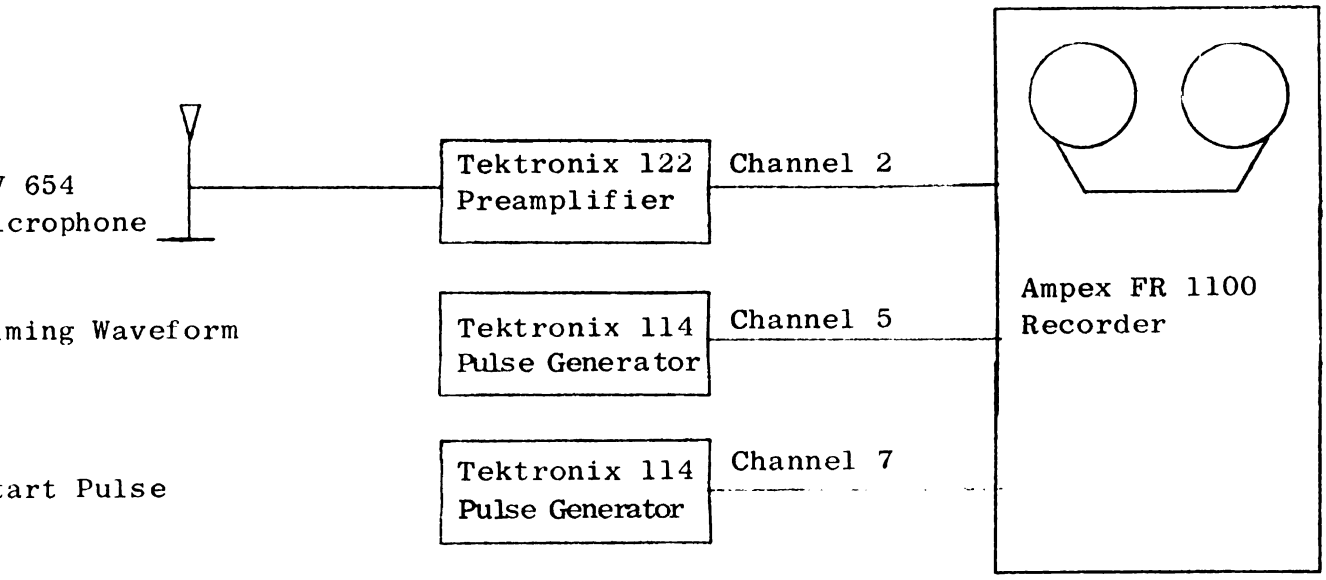


Fig. B-1. Apparatus for Recording Speech Signals on Analog Tape

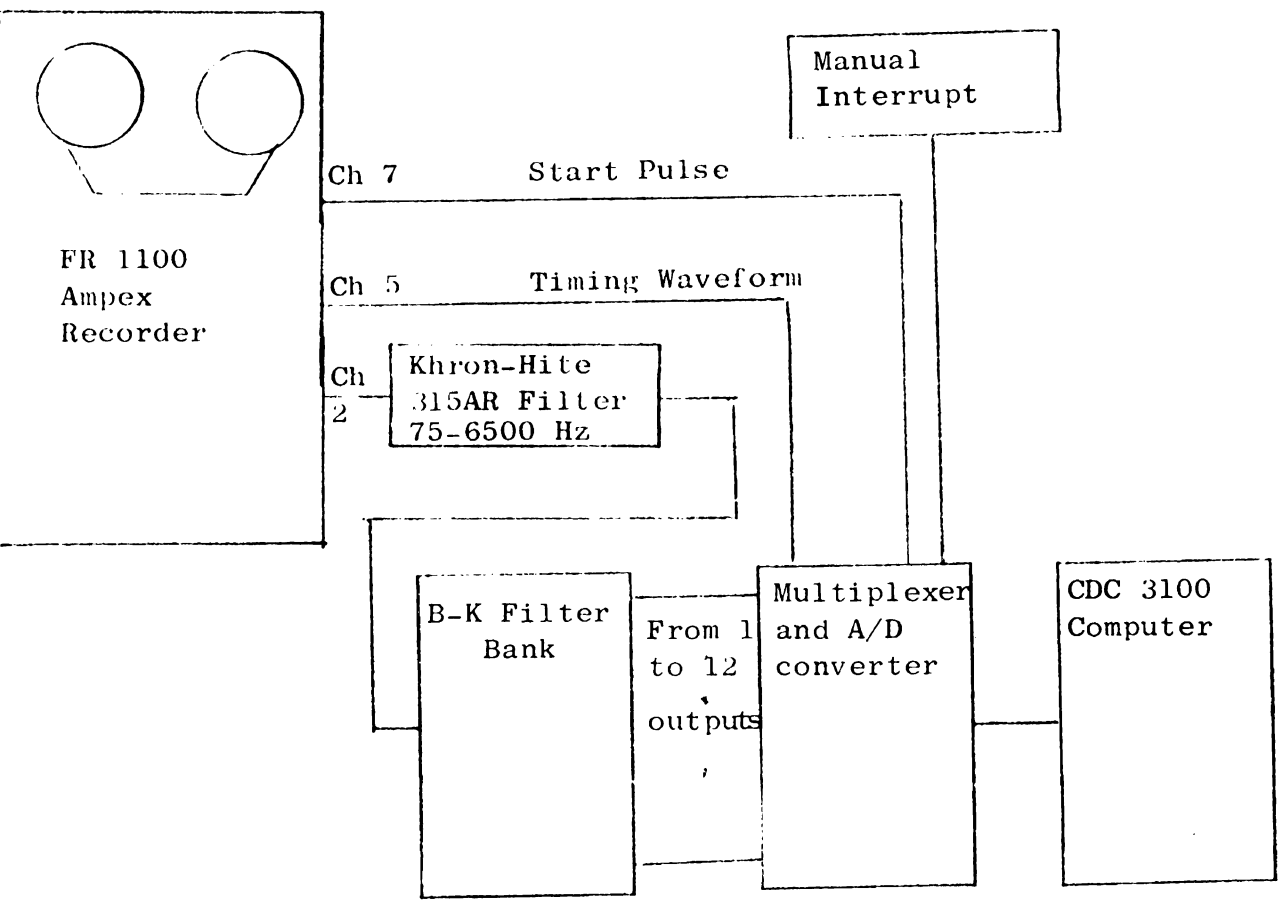


Fig. B-2. Apparatus for Multiplexing and Digitizing Data from Analog Tape

In Fig. B-2, the following operations are shown:

1) The analog tape was mounted on another FR 1100 Ampex Recorder. The reproduced output was passed through a Khron-Hite 315AR variable bandpass filter set to lower limits of 75 Hz and upper limit of 6500 Hz, then to a Brüel-Krüjer bank of third octave filters. This bank is shown in Fig. B-3 with further hardware and software recombination of these filter outputs to achieve the bandpass overlapping filter outputs as desired. The B-K filter bank was chosen because of its linear-phase characteristics, 50 dB per octave skirts for adequate bandpass filtering and ability to sum adjacent filters to increase bandwidth. Sampling was done on a CDC 3100 computer giving continuous A/D operation on an analog tape using a 12-channel multiplexer and a 10-bit A/D converter. The data were then sorted on digital magnetic tapes and used for further processing.

2) The A/D sampling rate was controlled by the timing waveform reproduced from the analog tape, minimizing variations from the desired 50 microsecond sample interval.

) Start pulses from Channel 7 initiated A/D operations (via 30-interrupt in the A/D converter). A manual interrupt was used for termination.

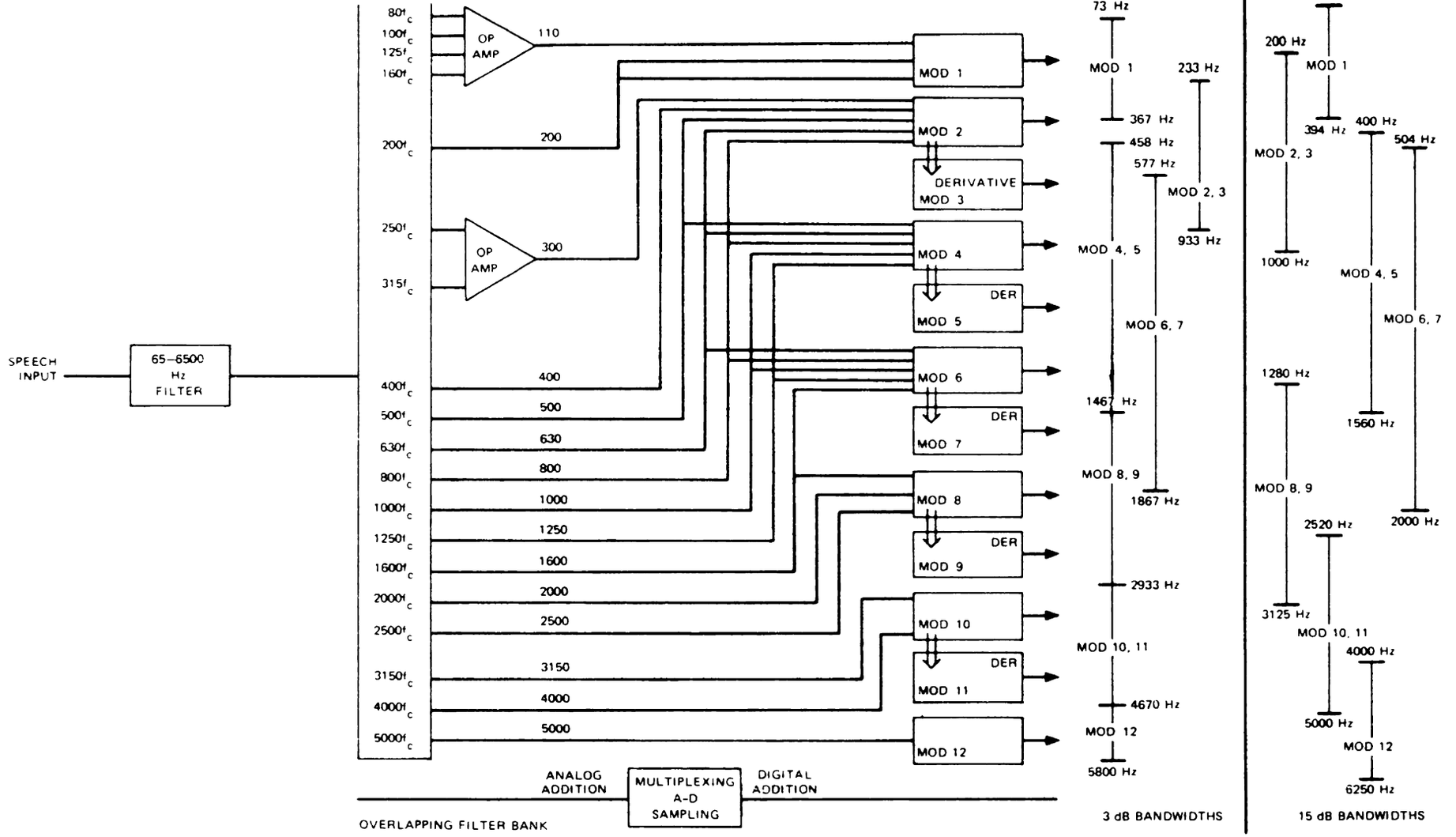


FIGURE B-3 OVERLAPPING FILTER BANK WITH 3 dB AND 15 dB BANDWIDTHS

APPENDIX C

TIMSER

A Program for Interactive Analysis of Time Series

MSER--Techniques for Interactive Manipulation of SEquential
ns--is a program ensemble that runs on the CDC 3300 at SRI.
developed to allow a user to edit and transform time series
tively, observing the results on a CRT display. The time
of primary interest are bipolar, one-dimensional time series
s the result of A/D operation on an analog voltage)and unipolar
riate time series (such as the envelope and zero-crossing count
ries derived from the bipolar sampled analog signal). Figure 1
n overview of the operations available to perform these two
f analysis.

ata

WTMSR--One-Dimensional Bipolar Time Series; i.e., the output
an A/D conversion of an analog waveform from magnetic tape.

e system allows inputting data from up to three tapes, which may
fferent sampling time intervals, accuracy (range of data), and
length. No capability is available now to unpack multiplexed data;
, it could be implemented by modifying a single subroutine. The
ers that are needed to read each magnetic tape are read in from
card. These tape parameters are fixed for each magnetic tape;
e:

(1) Tape ID (4 BCD characters);
(2) Length of record;
(3) Sample time interval;
(4) Logical unit; and
(5) Range of the data, 8 bits (± 256) to 12 bits (± 2048).

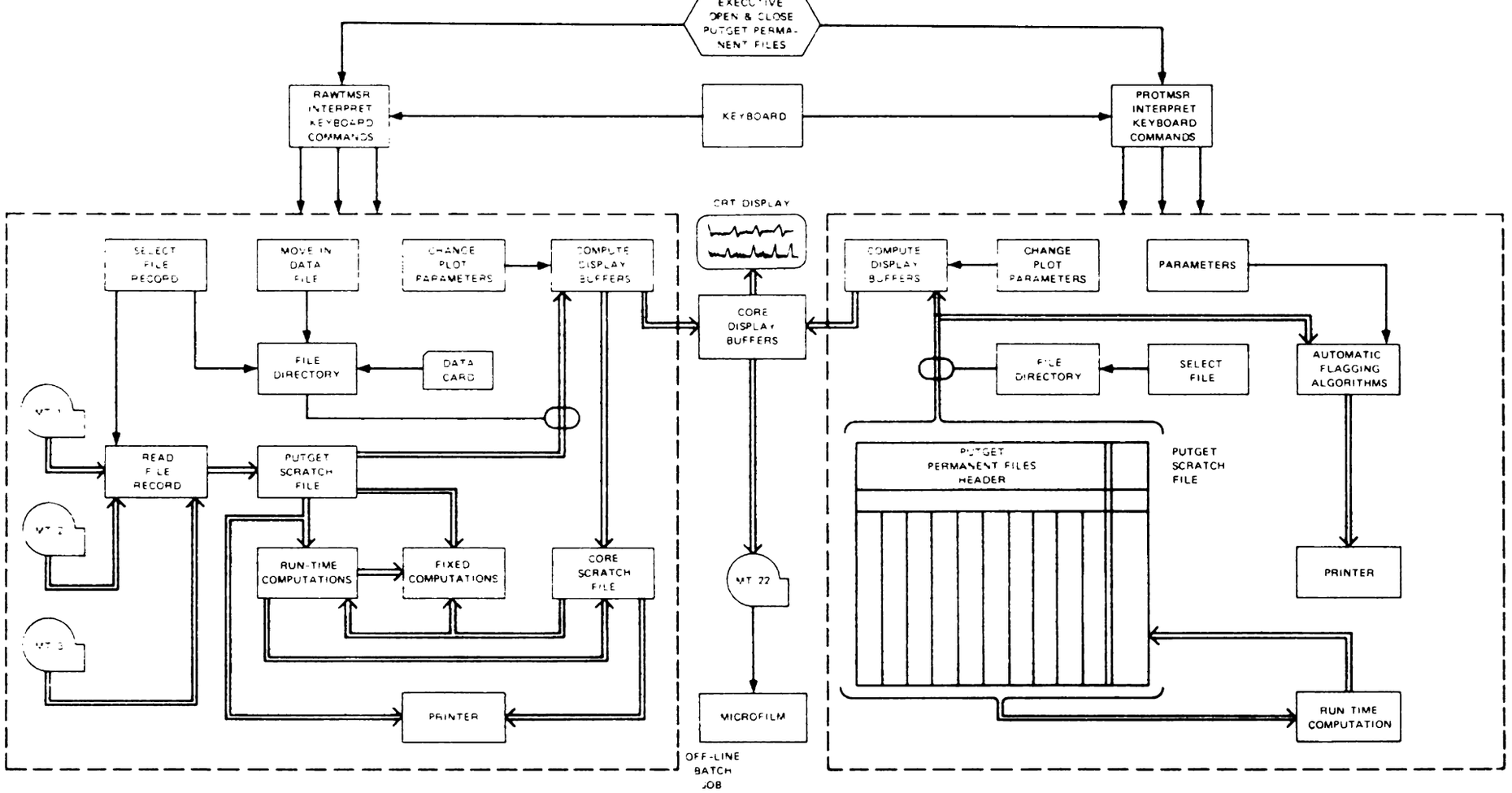


FIGURE C-1 TIMSER-INTERACTIVE DISPLAY PROGRAM ENSEMBLE FOR ANALYSIS AND DISPLAY OF MULTIVARIATE TIME SERIES

ucture of the magnetic tapes is typical for A/D operations; namely,
ry number of records per file and data blocked by end-of-file marks,
arbitrary number of files. A compass subroutine enables quick
ng for files.

OTMSR--Unipolar Multiple Time Series; i.e., results of sampling
terrelated analog waveforms or results of processing one-dimensional
gital time series.

ese time series are stored in random-access files that have
ble structure, depending on four parameters:

(1) Length of record
(2) Number of records
(3) Number of modules
(4) Number of dimensions.

four parameters are contained in a header at the start of each file
e assumed to be constant for that file. Use of a virtual-core storage
(PUTGET) permits storage of up to 200 files. Maximum record length
ited to 200; however, the rest of the parameters are bounded only by
ble disk storage. The number of dimensions is the number of different
eries in this file. The number of modules is a sub-file structure
llows a within-file breakdown of data. For instance, there may be
l processing schemes for one multiple time series which the user
to compare. The number of records parameter refers to each module.

tions Available to Both Programs

crofilm Hard Copy

is possible to obtain hard copy of the actual picture on the
splay. This is done by dumping the octal display buffers onto
ic tape and then converting them to microfilm pictures on the
) display, later, as a batch job.

omments and Titles

he user can type in a title or a comment on the CRT display; comments are transmitted to the hard copy. This is useful for ting where you are in your analysis and for titling microfilm es.

un-Time Computations

limited incremental compiler, allowing the user to manipulate, e, scale, or transform the time series, will soon be added. The ill be able to type up to 10 algebraic equations (which may perform ear transformation, lead-lag averaging, magnitude, absolute value, of squares computations, normalizing by maximum value, etc.)

g and Transformation Capabilities

AWTMSR

Data on magnetic tapes is transferred to a circular virtual-core r, that is, a buffer with a fixed length; when this length is led the first data introduced is overwritten. The user can select ne of three magnetic tapes from the keyboard; at this time, a data s read in with the tape parameters discussed above. Arbitrary tion of files and records from that tape can be made so that the lar buffer may contain an allocation of data not necessarily the as the original tape. From the circular buffer the time series are rted to octal display buffers for the CRT display. A pointer is to determine the origin of display in the circular buffer. It is ble to change plot parameters, including the number of points in urve, the number of curves on the screen, and the scale of the . The pointer can automatically increment through the buffer allow- apid editing. The user may also reference (save) one or more curves e screen and edit others for comparison. Once a curve is saved, er modification of the plot parameters will not affect it. Various ns are available for rapid editing of data from several tapes.

In addition to the editing capability, RAWTMSR allows some computations e performed on the time series. There are two types:

(1) Fixed Computations--including Fourier series and a smoothed time derivative computation. These computations can be made on any selected portion of the data in the circular buffer by setting limit pointers. The resulting time series are stored in a temporary scratch buffer, located in core, and immediately displayed above the last reference curve. These curves are automatically referenced so that further computations or moving of time series will not remove them.

(2) Run-Time Computations--These computations will be performed by the incremental compiler. They can be performed either on data from the circular buffer, again indicated by beginning and ending pointers, or on data in the scratch buffer in core.

The result of any of the above computations will be written over ything in the scratch buffer and immediately displayed. Permanent cords of the computations can be made either by the hard copy option by printing the scratch buffer contents.

For example, the combination of these fixed run-time computations an result in the following display: First, pointers are set in the cratch file and a Fourier transform is computed. Then the incremental ompiler is called, and a logarithmic transformation of the magnitude of he Fourier series is computed and normalized. Then another Fourier ransform, on the resulting time series, is computed and displayed. The esulting waveform, called a cepstrum, is useful in speech analysis. The procedure of introducing the data in the scratch file, assigning beginning and ending pointers, and calling a subroutine to do the computing is common to many forms of time series analysis; namely, autocorrelation computations, convolution operations with matched filters, etc. and allows a general structure for incorporation of additional operations.

A typical hard copy (microfilm) picture (Fig. 4 in the text, repeated) indicates some of the editing capabilities. Wave forms from four erent files (4, 5, 6, 24) and three different tapes (MD4, MD6, MD8) are layed simultaneously. Each waveform is labeled with the beginning and ng time references (from the start of each file) computed from the le interval parameter for each tape. File names (19 BA 1, 19 MJ 1, BE 1, 19 EH 1) and a general comment (ØUTPUTS ØF ØVERLAPPING BAND PASS TERS [M] TØ [B1]) are also displayed.

PROTMSR

The PROTMSR program allows a study of the interrelationships of time ies selected from random access files. Four two-dimensional tter plots are displayed. The selection of the plot parameters for ch of these four plots allows the plotting of an individual time series rsus its index, the scatter plots of two time series (from the same le, but not necessarily the same record or module), and comparisons of atter plots from four different files. The structure of the files on the sk is completely determined by the parameters (discussed above) in the eader; thus, it may vary from one file to another. An index function nvolving four parameters (instead of the three commonly available in ortran) is used in a separate subroutine called INDEX so that he usual Fortran requirements of predetermined dimensions and maximum alue of each dimension are not necessary. A file directory showing the various parameters and file identification data from the header is available at user option for selecting the files. Various options are available to facilitate the comparison of the four scatter plots:

(1) A time sequence option, which allows ten points on each scatter plot to be labeled 0 through 9 according to their sequential index. These ten labels are then incremented through the scatter plot, showing the sequential relations.

(2) An overlay option, which plots all four scatter plots on common axes.

(3) Automatic incrementing of either dimensions, modules, or records for rapid editing.

un-time compiler can be used in this program to generate a new file and ew set of derived time series from any of the existing files. Similar es of transformations, as discussed before, are available here.

A typical picture (Fig. 17b in the text, repeated here) shows four nultaneous time series plots, the two on the left are univariate plots d the two on the right are bivariate (scatter) plots. Labels on the ttom are added by the user. Each plot is labeled with a file name 6 BE 1) three index parameters (D1 M5 R2) or TIME (indicating the dex) with maximum time shown (480 ms). Names of the D index are also nown (ABS ENV). Scale for each axis is shown by a factor (X 4) which ultiplies the original data.

APPENDIX D

SLIDING POWER SPECTRA

Sliding power spectra are computed from the A-D tapes described in Appendix B by means of the TIMSER display program ensemble. Each curve displays the square root of the power spectra computed over a fixed time interval (25 milliseconds for all curves in this appendix).

Spectrum smoothing is done by multiplication of the time waveform by the following taper function:

$$(1-x^2)^2 , \quad -1 \leq x \leq 1$$

This smoothing is performed to minimize the effects of the pitch frequency and give better side lobe response (see Blackman and Tukey).

The labels on each picture give fundamental frequency (40 hz), file number and tape ID (corresponding to module), linear or log magnitude plot, maximum frequency, utterance and speaker label, and filter bandwidth. The label for each curve shows the start time and the square root of power (all curves in this appendix are stepped 15 milliseconds).

Each curve is normalized to the maximum frequency component. The linear magnitude plots show percentage of the maximum component. The log magnitude plots show dB relative to the maximum component (50 dB range).

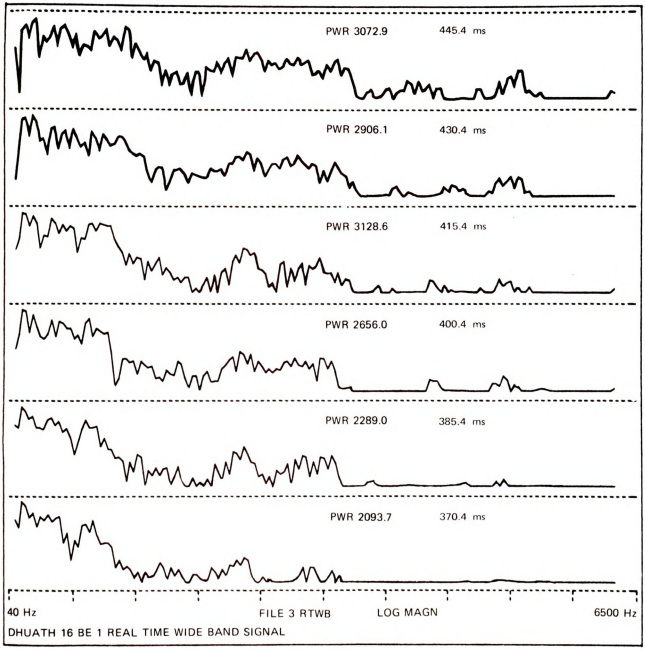


FIGURE D-1 DHUATH 16 BE 1 REAL–TIME WIDEBAND SIGNAL

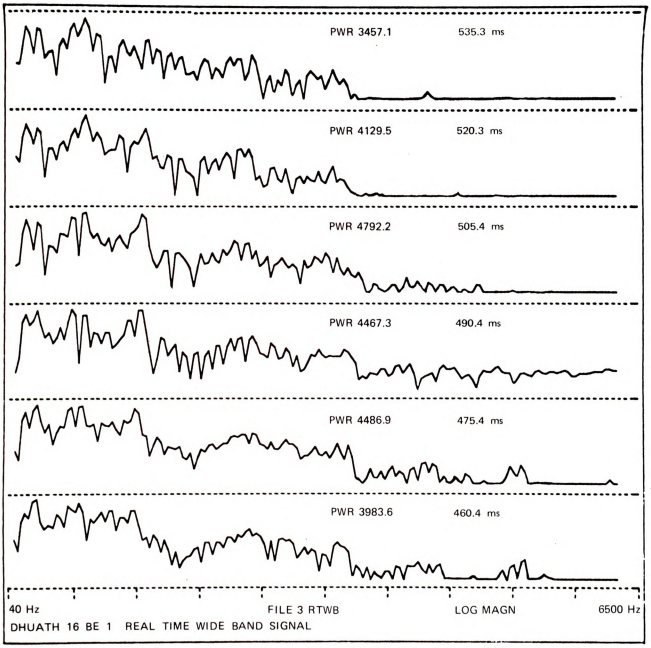


FIGURE D-1 DHUATH 16 BE 1 REAL-TIME WIDEBAND SIGNAL (Continued)

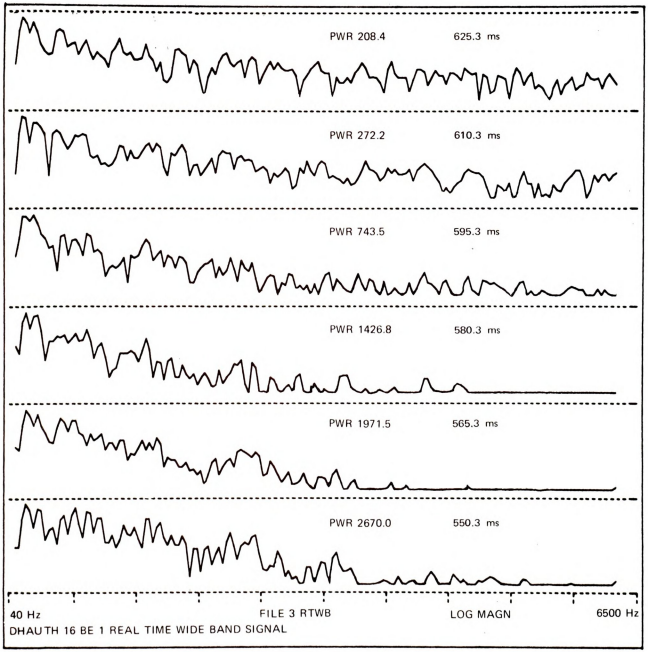


FIGURE D-1 DHUATH 16 BE 1 REAL-TIME WIDEBAND SIGNAL (Continued)

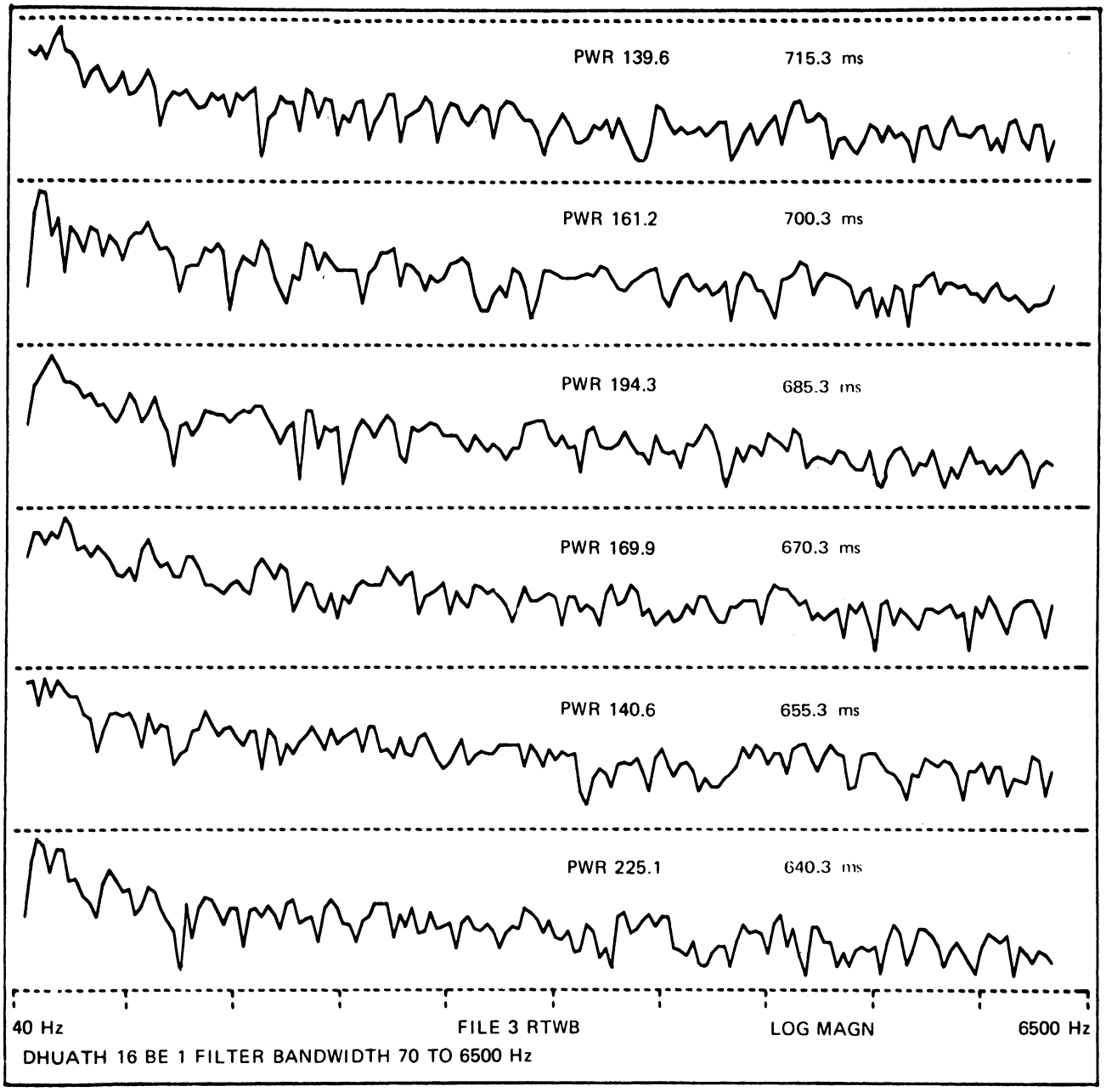


FIGURE D-1 DHUATH 16 BE 1 REAL-TIME WIDEBAND SIGNAL (Concluded)

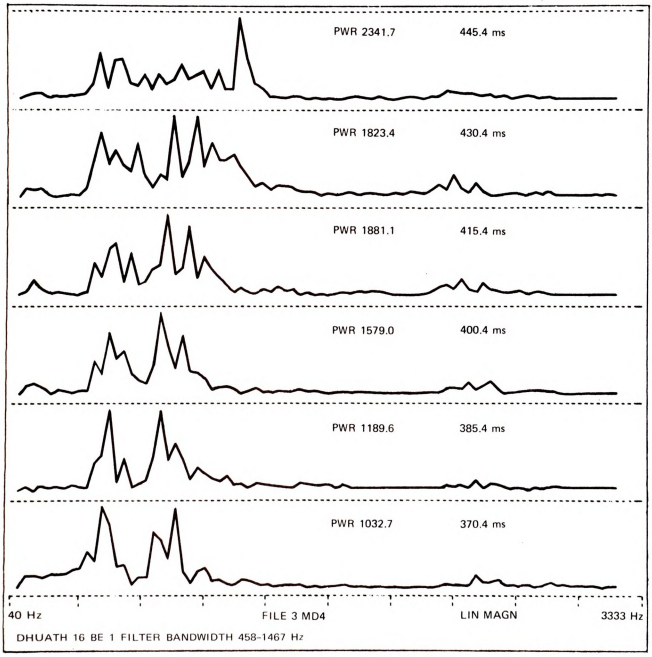


FIGURE D-2 DHUATH 16 BE 1 FILTER BANDWIDTH 458-1467 Hz

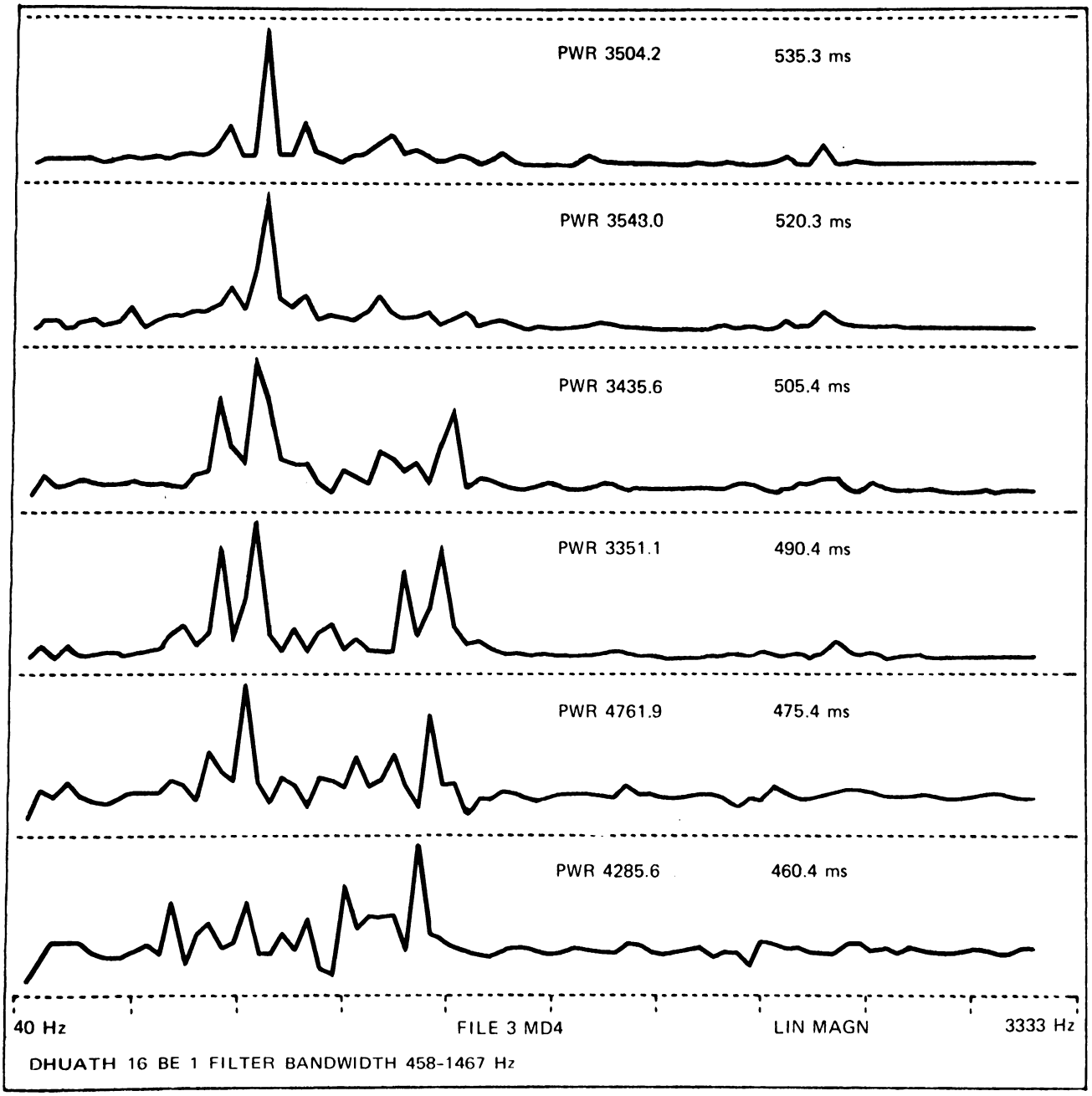


FIGURE D-2 DHUATH 16 BE 1 FILTER BANDWIDTH 458-1467 Hz
(Continued)

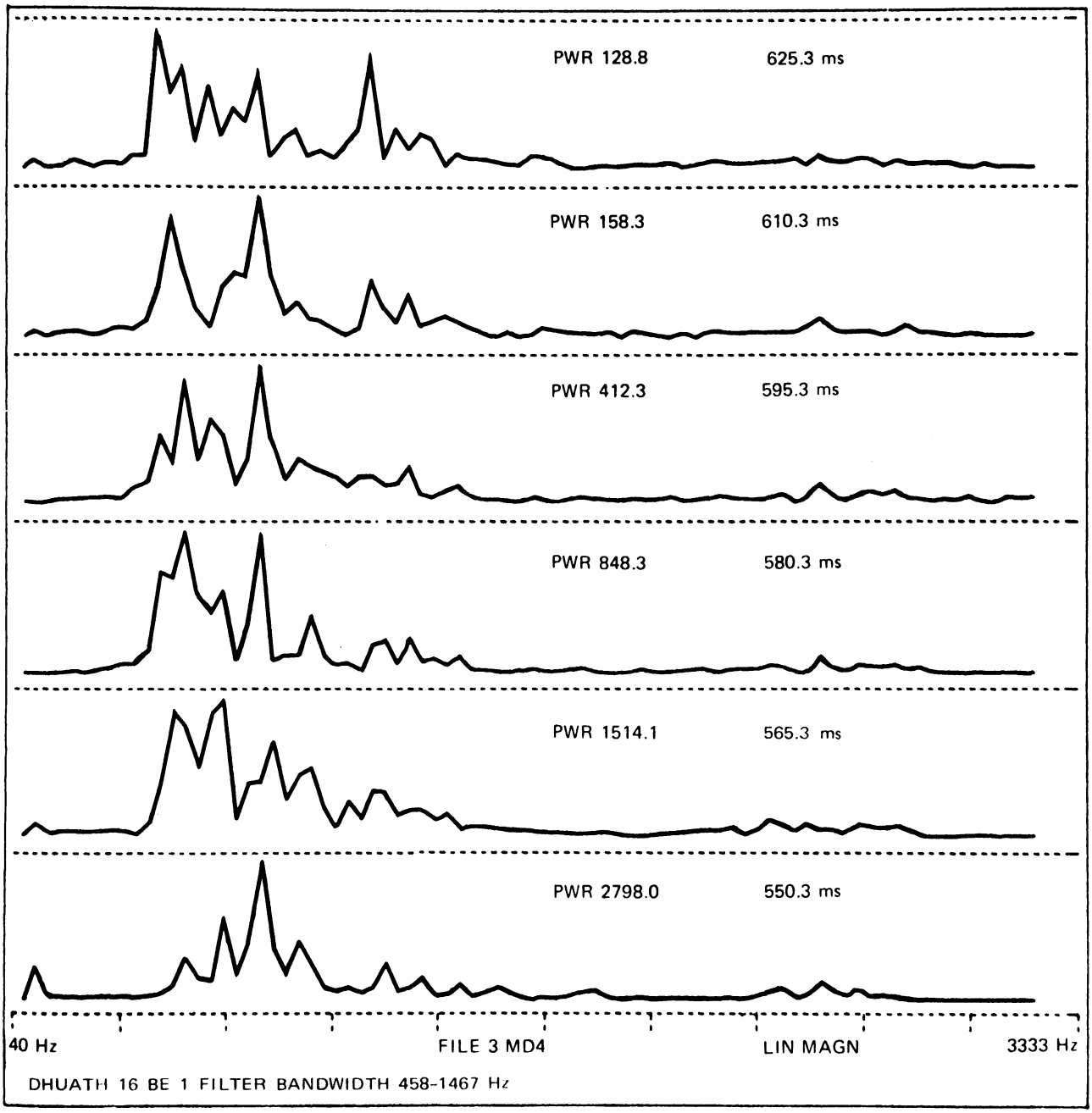


FIGURE D-2 DHUATH 16 BE 1 FILTER BANDWIDTH 458-1467 Hz
(Continued)

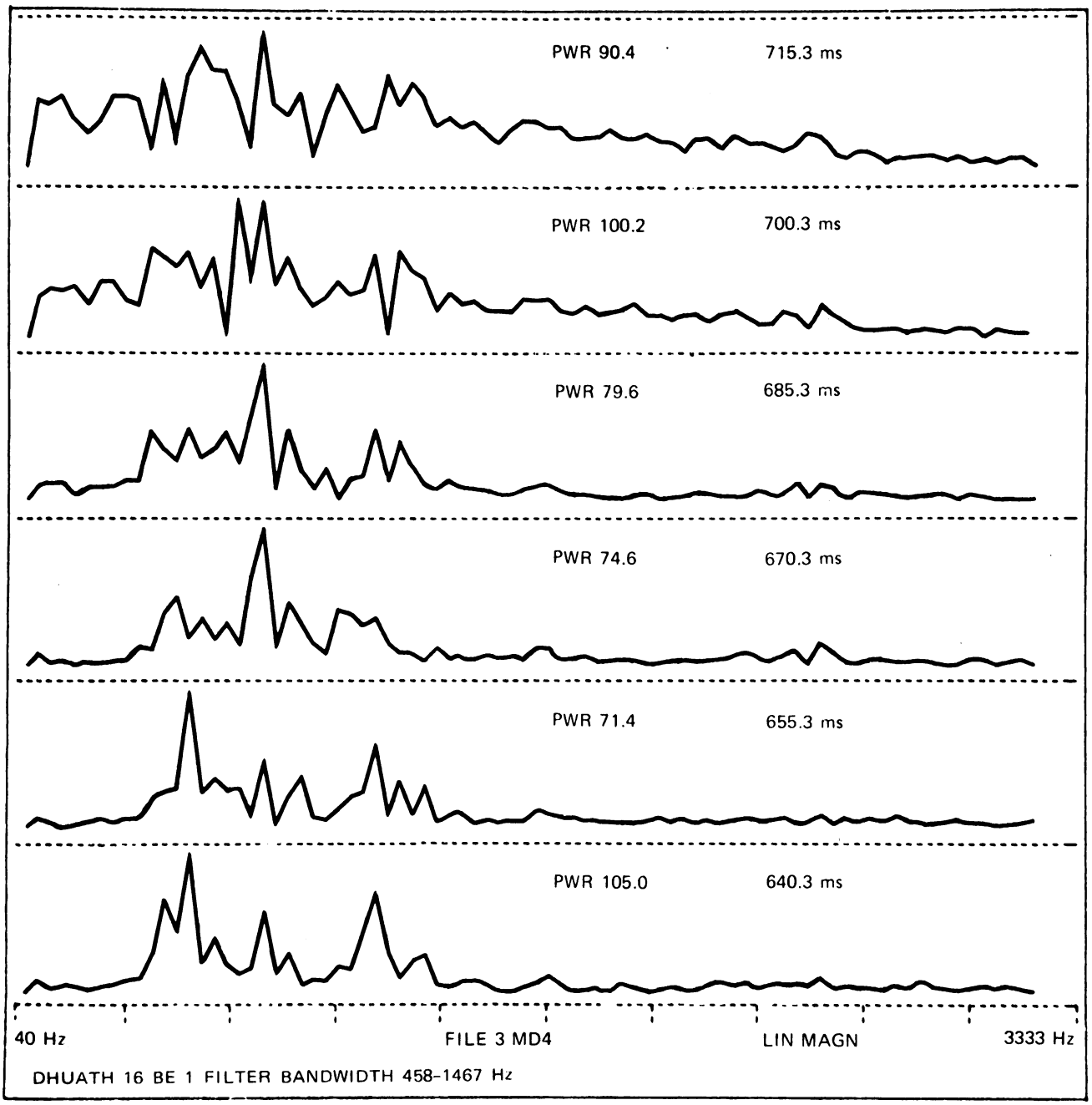


FIGURE D-2 DHUATH 16 BE 1 FILTER BANDWIDTH 458-1467 Hz
(Concluded)

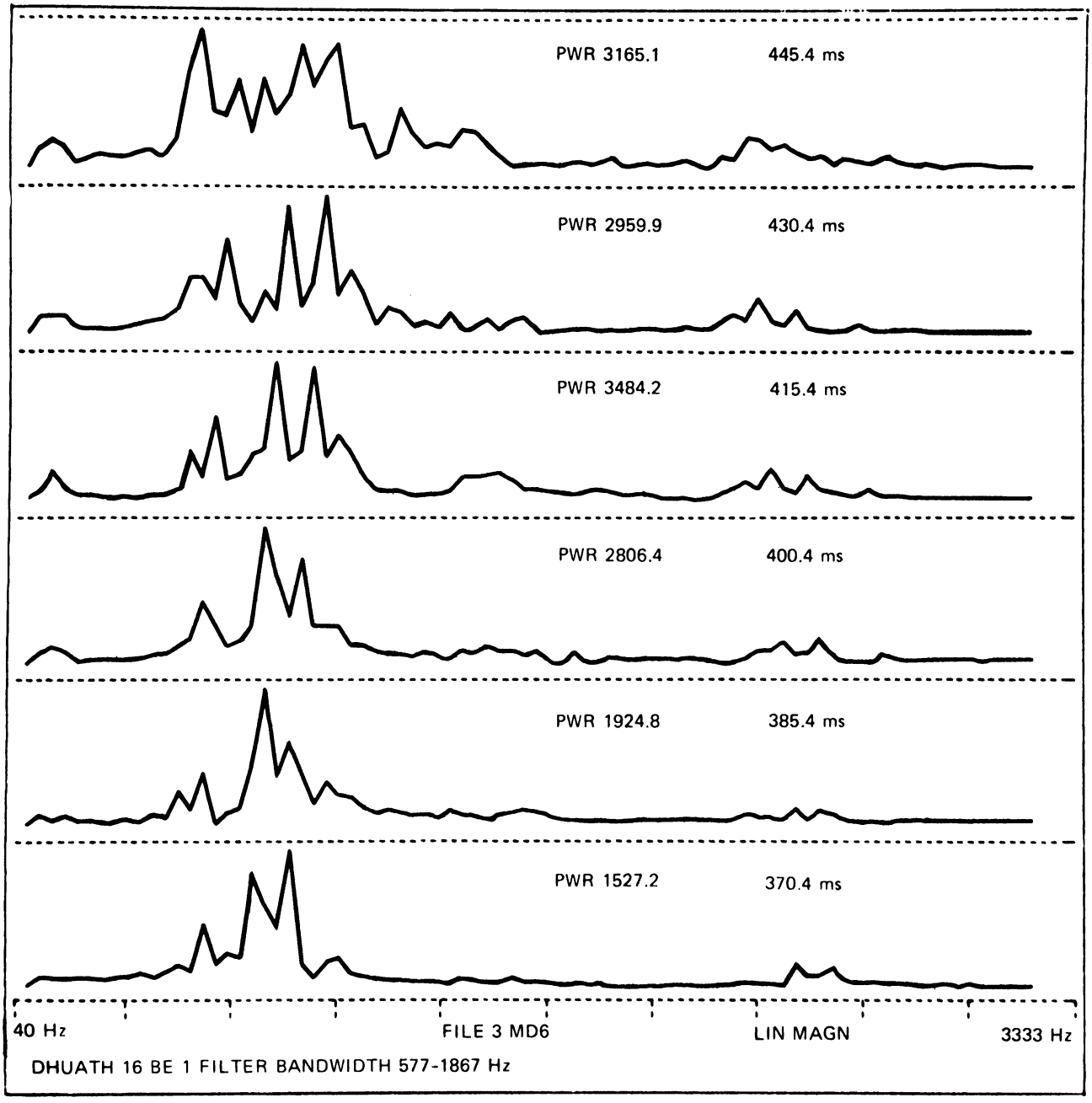


FIGURE D-3 DHUATH 16 BE 1 FILTER BANDWIDTH 577-1867 Hz

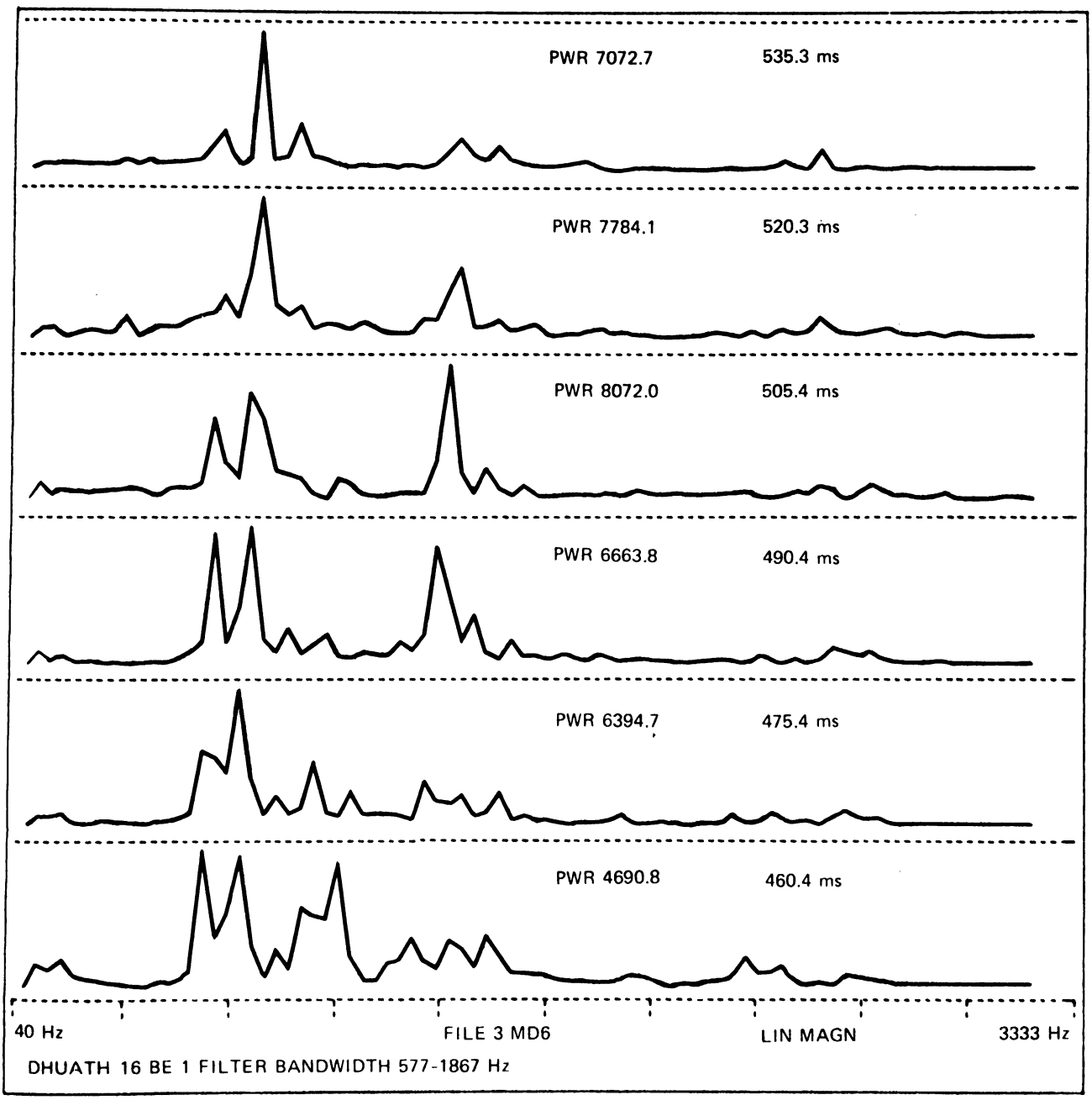
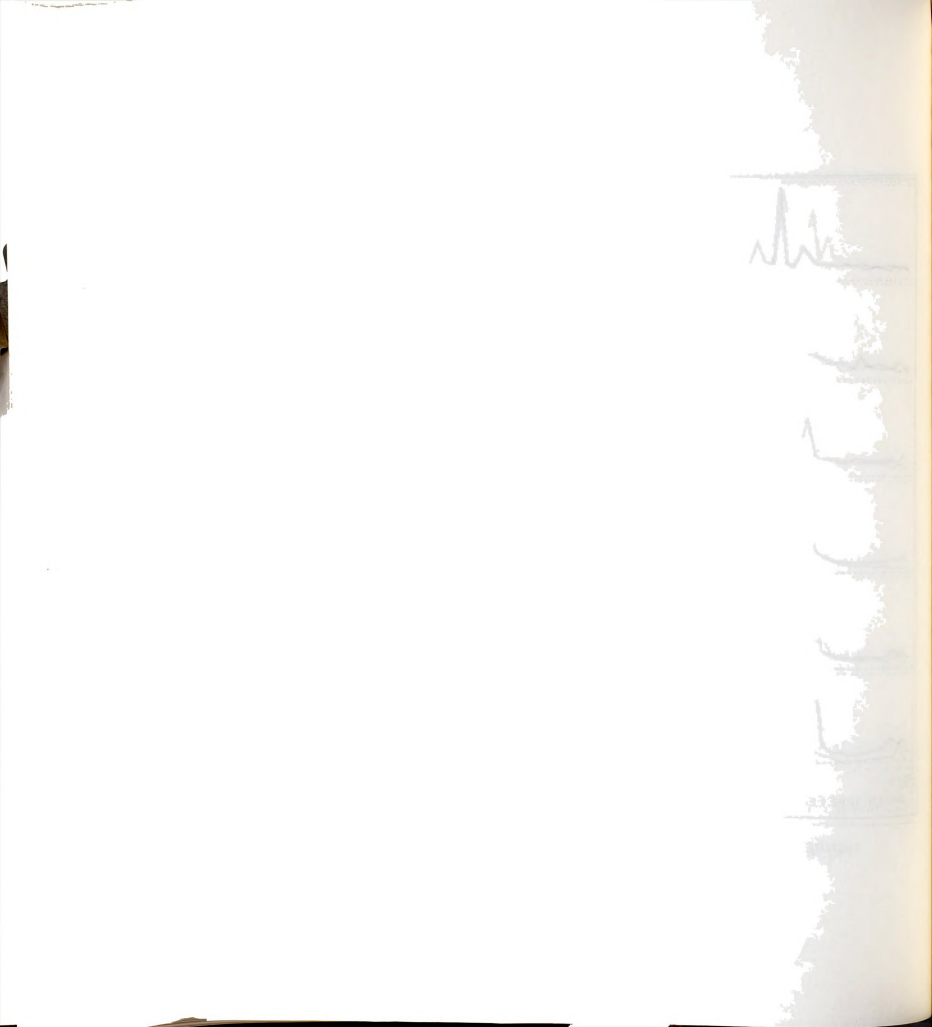


FIGURE D-3 DHUATH 16 BE 1 FILTER BANDWIDTH 577-1867 Hz
(Continued)



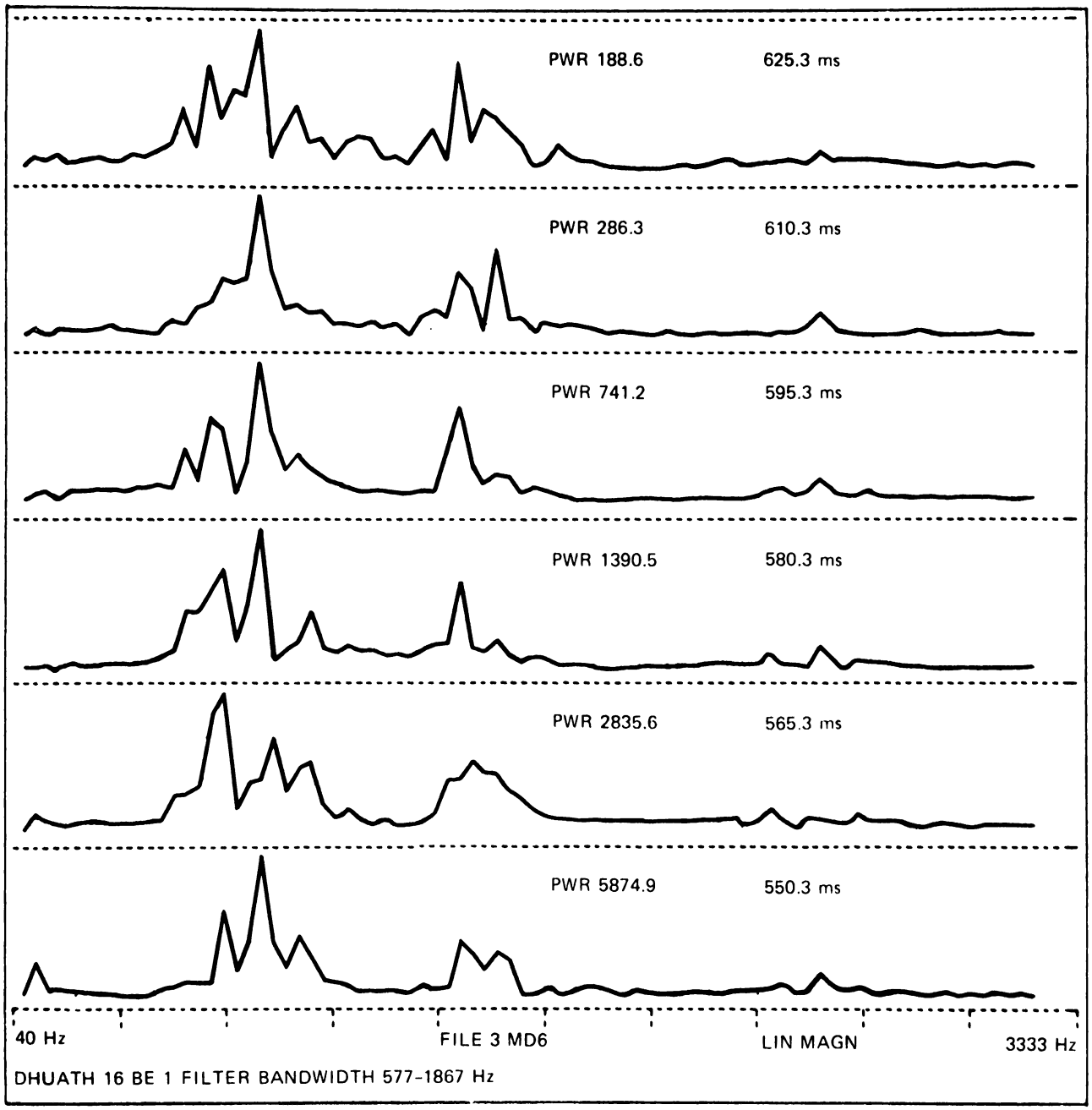


FIGURE D-3 DHUATH 16 BE 1 FILTER BANDWIDTH 577-1867 Hz
(Continued)

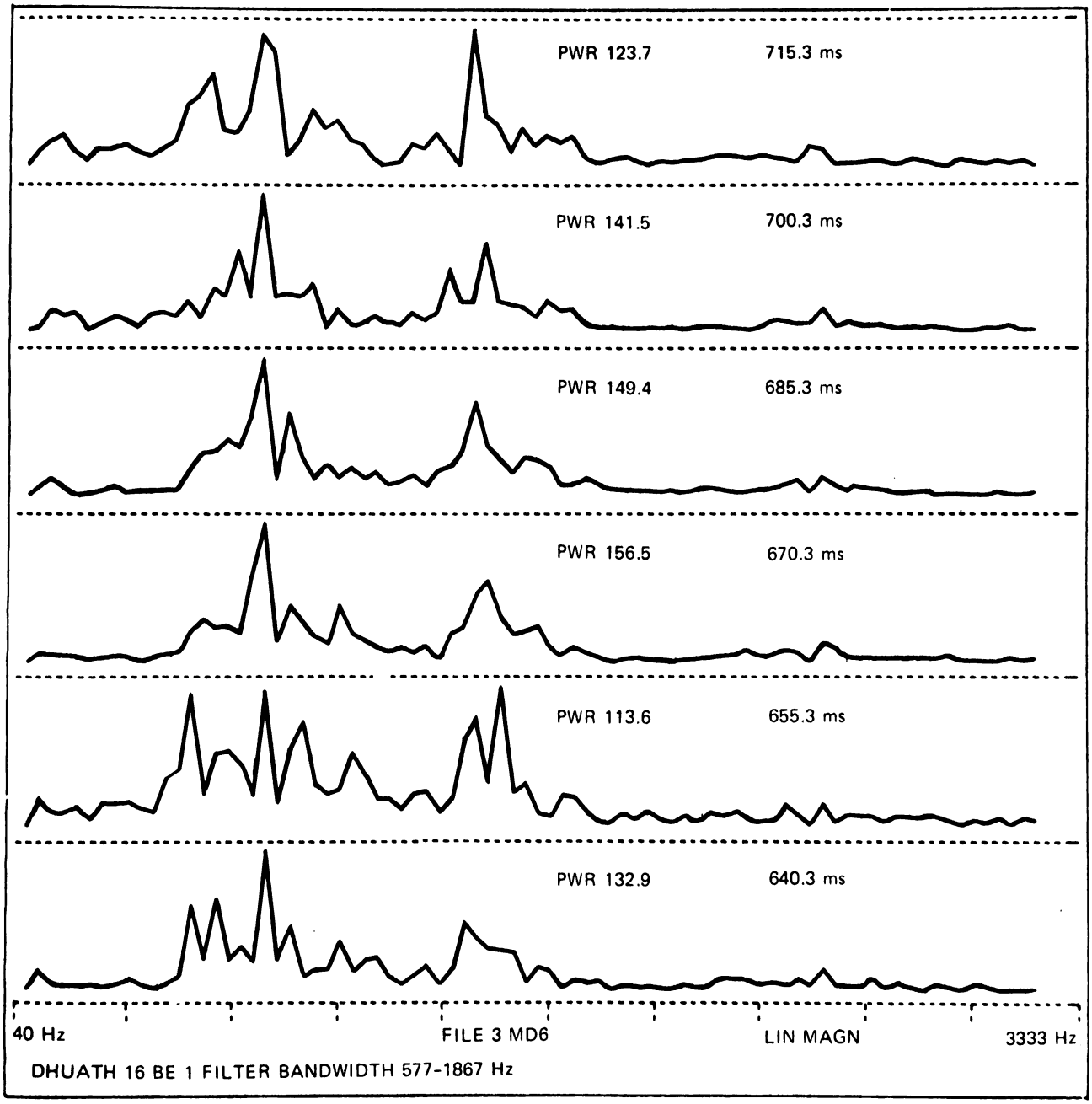


FIGURE D-3 DHUATH 16 BE 1 FILTER BANDWIDTH 577-1867 Hz
(Concluded)

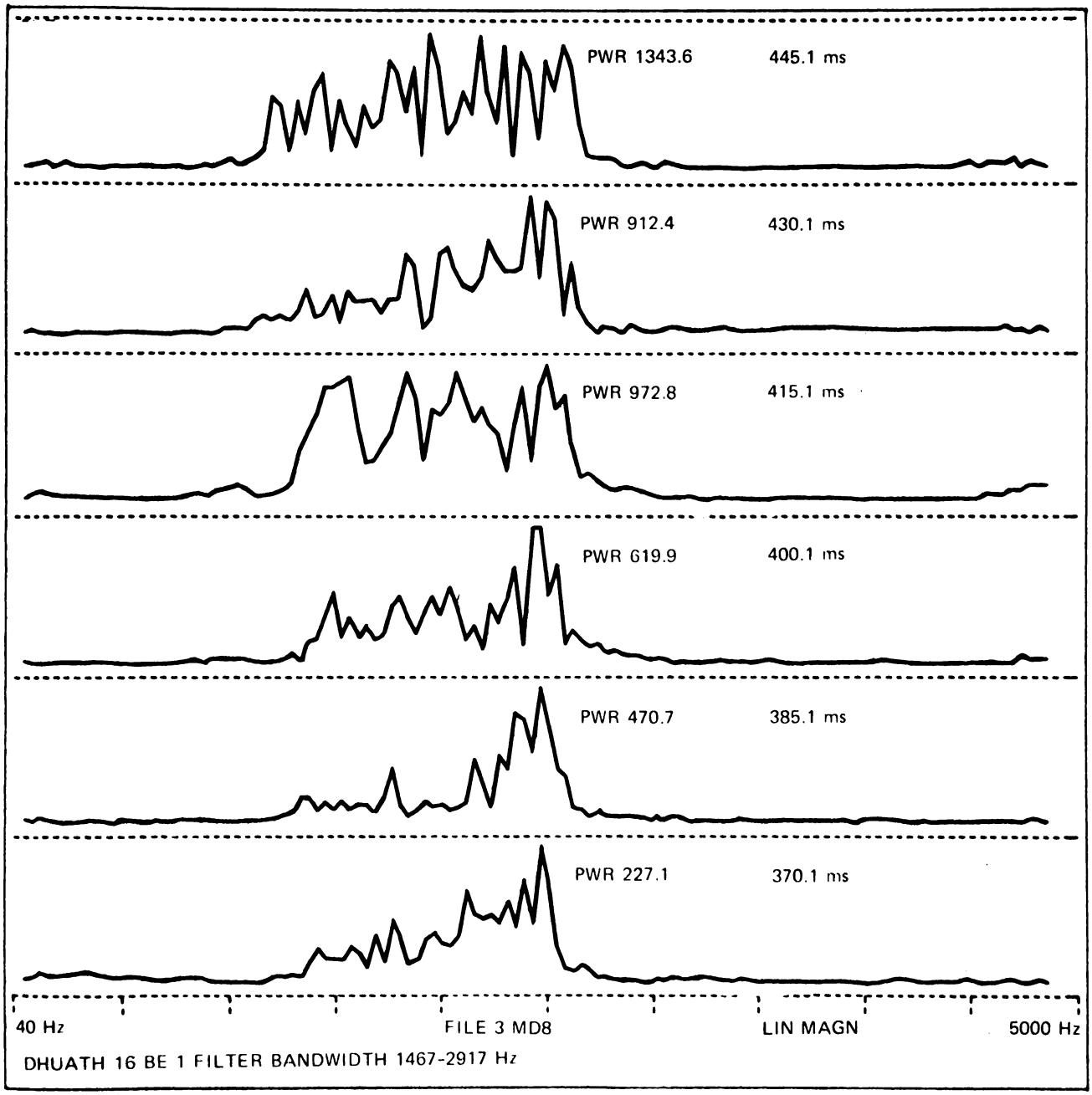


FIGURE D-4 DHUATH 16 BE 1 FILTER BANDWIDTH 1467-2917 Hz

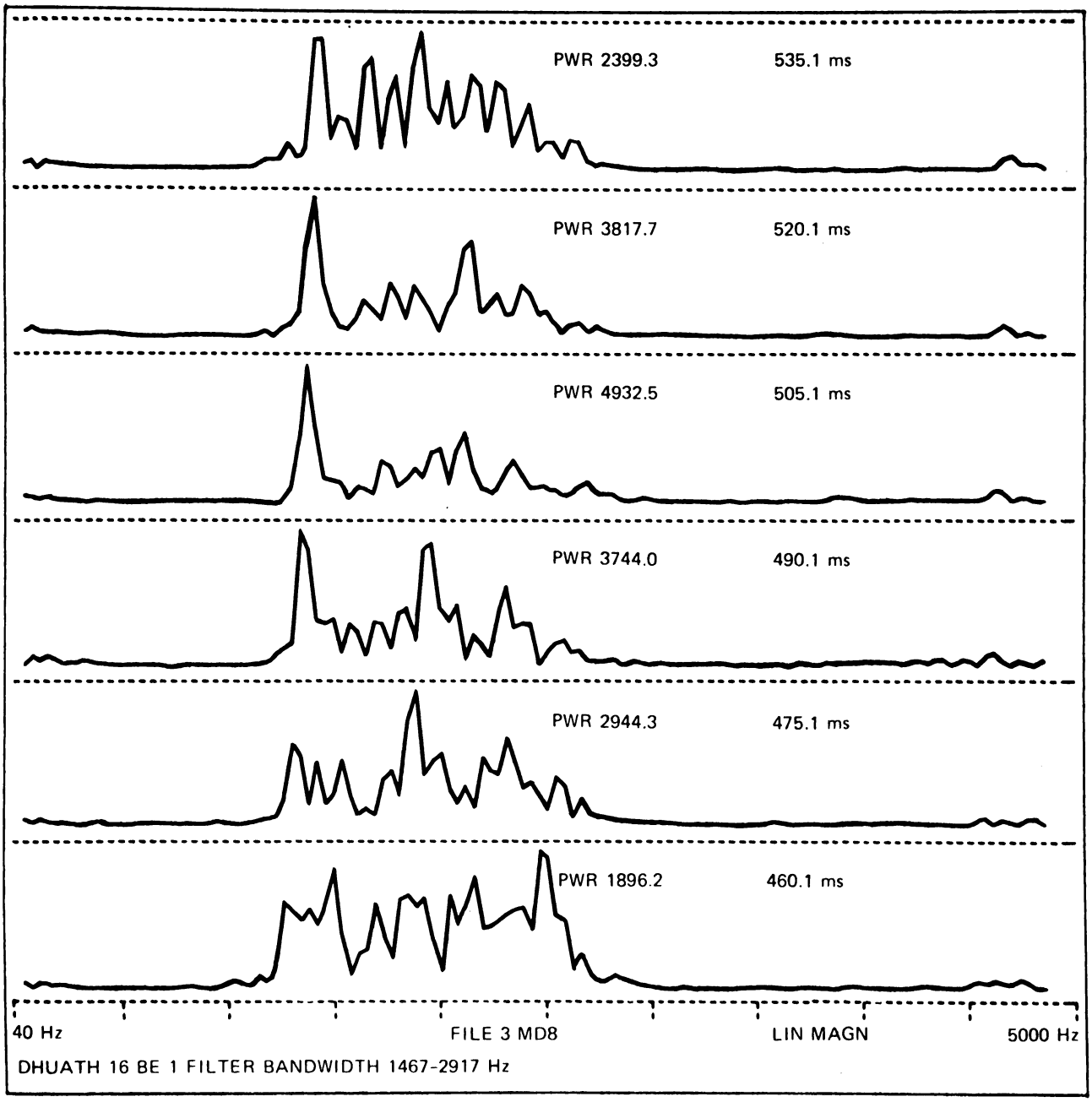


FIGURE D-4 DHUATH 16 BE 1 FILTER BANDWIDTH 1467-2917 Hz
(Continued)

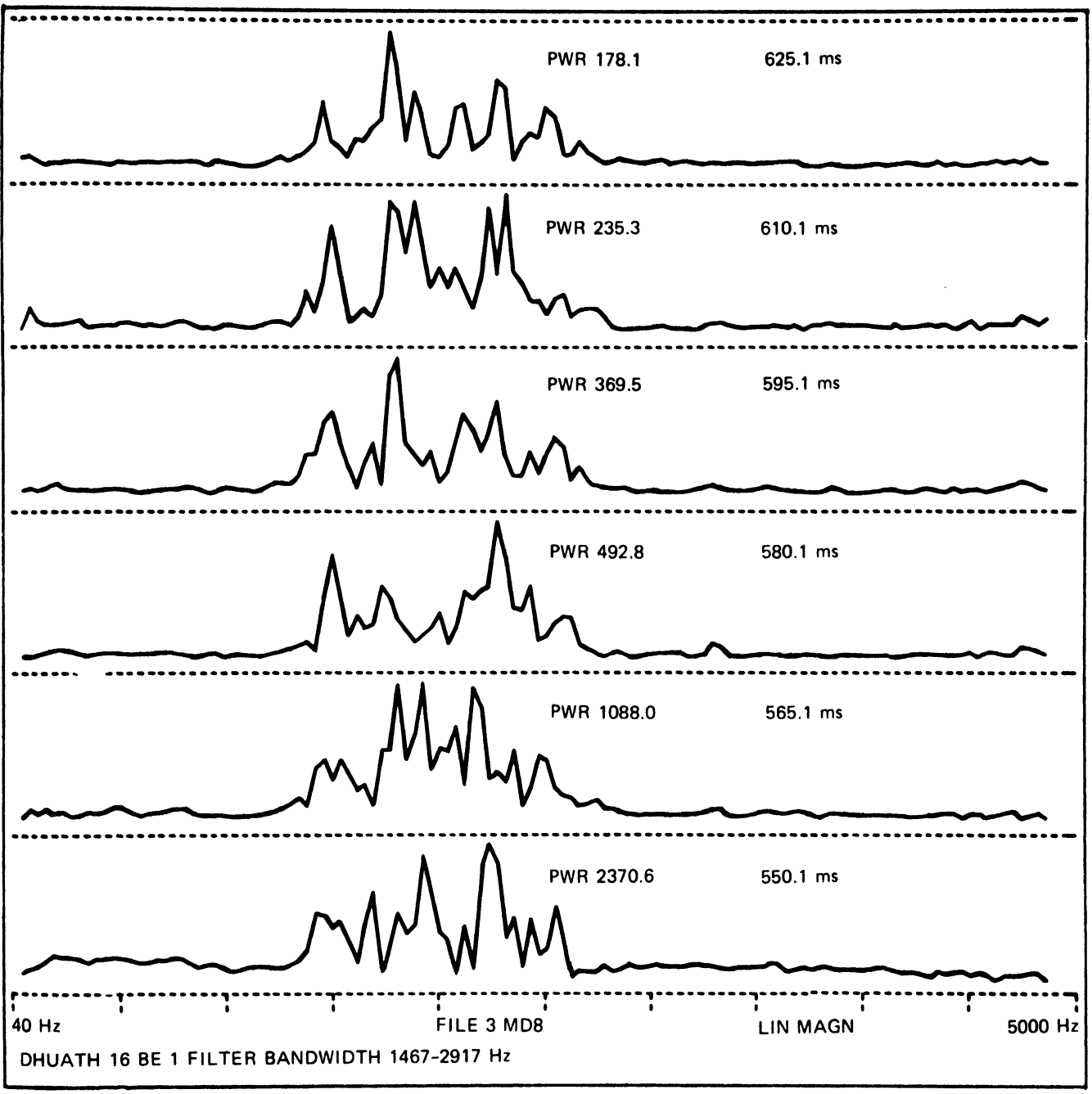


FIGURE D-4 DHUATH 16 BE 1 FILTER BANDWIDTH 1467-2917 Hz
(Continued)

APPENDIX E

INSTANTANEOUS ESTIMATORS OF TIME-VARYING PARAMETERS

It has been shown that representation of a class of acoustical signals can be reduced to estimation of the time-varying frequency and envelope and their derivatives. A common estimator of the instantaneous frequency is a sliding average of the zero crossings of x.

$$\tilde{\omega}(t) = K_1 \int_{t-\tau}^{t} z(\sigma)\, d\sigma \qquad \text{(E-1a)}$$

Or for discrete samples*

$$\tilde{\omega}_n = K_1 \sum_{j}^{n}{}_{k}\, z_j \qquad \text{(E-1b)}$$

Where K_1 is a normalizing constant

$z_j = 1$ if $x_{j-1} \leq 0$ and $x_j > 0$

or $x_{j-1} \geq 0$ and $x_j < 0$

$= 0$ otherwise.

A reasonable estimator of the envelope of $\hat{x}(t)$ is the sliding mean of the absolute value of the real part.

$$\tilde{a}(t) = \frac{1}{\tau} \int_{t-\tau}^{t} |x(\sigma)|\, d\sigma \qquad \text{(E-2a)}$$

---

* We will denote all estimators by adding a tilde (the estimate of ω is $\tilde{\omega}$) and the sliding sum of length k over the index j from n-k+1 to n as $\sum_{j}^{n}{}_{k}$

for discrete samples

$$\tilde{a}_n = \frac{1}{k} \sum_{j}^{n} \Big|_{k} |x_j| \qquad \text{(E-2b)}$$

Another estimator for the envelope of $\hat{x}(t)$ can be derived from equation (II-A-3c), that is, an average of the magnitude of $\hat{x}(t)$. The Hilbert transform of an arbitrary function can be obtained by means of a complex digital filter (Crystal and Ehrman[58]) operating on the real signal.

The computation of these estimators involves a non-linear, no-memory operation followed by a low-pass filter.

The problem of removing the oscillatory terms from the state variable differential equations as discussed in Sec. II-A and the selection of τ (or k) is analogous to the selection of the cutoff frequency for the lowpass filter shown in Figure E-1. An effective measure for stationary signals is the mean square bandwidth. Abramson[60] has shown that the mean squared bandwidth (see II-A-14) of v(t), the result of the nonlinear, no-memory operation, is computed by the following formula:*

$$B_v^2 = \frac{E\left\{(v')^2\right\} E\left\{x^2\right\}}{E\left\{v^2\right\}} B_x^2 \qquad \text{(E-3)}$$

---

*We denote the derivative of a function, v, with respect to its argument as v'.

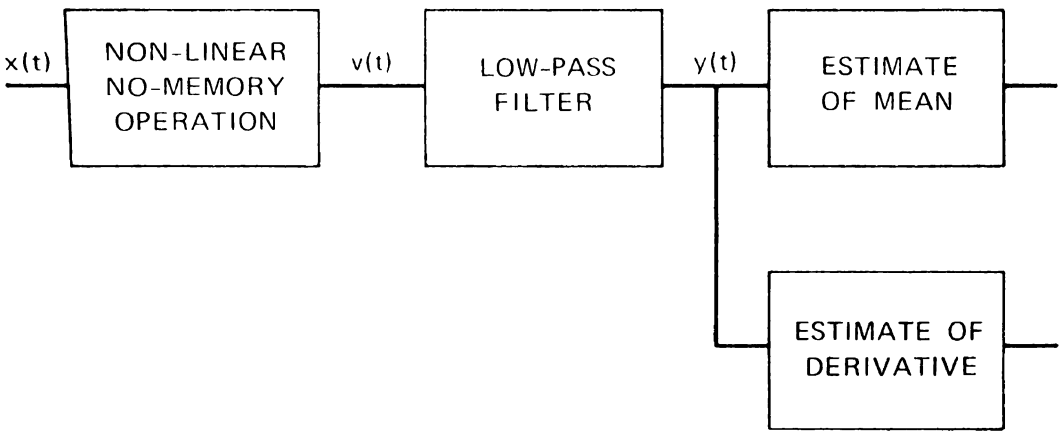


FIGURE E-1 OPERATIONS FOR PARAMETER ESTIMATION

For the situations we are considering, B_v^2 is equal to a constant related to the non-linear operation times the mean squared bandwidth of the input. For example, the bandwidth of the envelope function (using II-A-14) for stationary Gaussian processes

$$\hat{x}(t) = a(t)e^{j\alpha(t)}$$

$$B_a^2 = \frac{E\left\{\dot{a}^2\right\}}{E\left\{a^2\right\}} = B_{x^s}^2 \Big/ 2 \qquad \text{(E-4)}$$

where x^s is the shifted (low pass) version of $\hat{x}$. Use of a full wave detector (absolute value) as an estimator of the envelope gives a bandwidth (Abramson)

$$B_v^2 = \frac{E\left\{(g')^2\right\} E\left\{x^2\right\}}{E\left\{g^2\right\}} B_x^2 = \frac{E\left\{\left(\frac{x}{|x|}\right)^2\right\} E\left\{x^2\right\}}{E\left\{|x|^2\right\}} B_x^2$$

$$= B_x^2 \qquad \text{(E-5)}$$

The mean frequency can be derived by converting v(t) to an analytic signal with discontinuous phase

$$\hat{v}(t) = b(t)e^{j\beta(t)} = |x(t)| + j|x^h(t)| \qquad \text{(E-6)}$$

$$\hat{x}(t) = a(t)e^{j\alpha(t)}$$

where

$$b(t) = a(t)$$

$$\beta(t) = \alpha(t) + \zeta(t)$$

$\zeta(t)$ is a step function which increases by π whenever $x(t) = 0$.

Using (II-D-2), the mean frequency, $\bar{\omega}$, is given by:

$$\bar{\omega} = \frac{\int_0^\infty b^2(t)\dot{\beta}(t)\,dt}{\int_0^\infty b^2(t)\,dt} = \frac{\int_0^\infty a^2(t)\alpha(t)\,dt}{\int_0^\infty a^2(t)\,dt} + \frac{\int_\Gamma a^2(t)\,d\zeta(t)}{\int_0^\infty a^2(t)\,dt}$$

where $\Gamma = \{t \mid x(t) = 0\}$

Since the discontinuities at $t \in \Gamma$ are steps (first order), the last integral is zero. Consequently, $\bar{\omega}$ becomes

$$\bar{\omega} = \frac{\int_0^\infty a^2(t)\alpha(t)\,dt}{\int_0^\infty a^2(t)\,dt} \qquad \text{(E-7)}$$

For signals generated by time-varying differential operators, the mean square bandwidth is not an effective criterion. Rather, the instantaneous fluctuations of the bandwidth must be considered. We can fix an upper bound by using a Chebyshev inequality for stochastic processes (Parzen)[69] for the time interval $[t_1\ t_2]$.

$$P\left[\sup_{t_1\le t\le t_2} \left|b_{xs}(t)\right| > \Delta\omega\right] \le \frac{1}{\Delta\omega^2} E\left\{\sup_{t_1\le t\le t_2} \left|b_{xs}(t)\right|^2\right\} \qquad \text{(E-8a)}$$

$$E\left[\sup_{t_1\le t\le t_2} \left|b_{xs}(t)\right|^2\right] \le \tfrac{1}{2}\left[E\left\{\left|b_{xs}(t_1)\right|^2\right\} + E\left\{\left|b_{xs}(t_2)\right|^2\right\}\right] + \int_{t_1}^{t_2} E\left\{\left|b_{xs}(t)\right|^2\right\}^{\frac{1}{2}} E\left\{\left|b_{xs}(t)\right|^2\right\}^{\frac{1}{2}} dt \qquad \text{(E-8b)}$$

$[\,A\,]$ is the probability of the event A

$[\,\cdot\,]$ is the statistical expectation with respect to P.

$b_{xs}(t)|$ is constant (time-invariant generating equation), we have
amping coefficient b^r)

$$\left[\,_{t\,\le t\le t}\ |b_{xs}(t)| \ > \Delta\omega\right] \ \le \ b^r/(\Delta\omega)^2$$

have seen, the magnitude of b_{xs} is related to both the effective band-
for deterministic system impulse responses and effective bandwidth for
arying differential operators with white noise driving functions. We may
$\Delta\omega$ to the bandwidth of our bandpass pre-filters and then use this rela-
ip to determine a length of average $[t_1,t_2]$ (which is related to the
frequency). The bound, then, depends on not only the instantaneous
of our state variables, but also the instantaneous values of their
tives. We can use these relationships to investigate the properties
cific estimators.

ne estimation of the derivatives of the state variables is, in general,
sy for noisy observations. A more stable estimator is derived by alge-
anipulation of the stochastic derivative (Parzen)[69] of a time series
$-1\}$ with finite second moment. (y_n may be the discrete envelope samples,
or the zero crossing samples, z_n)

$$\overset{d}{=} \underset{\Delta\to 0}{\text{L.I.M.}} \frac{y_n - y_{n-1}}{\Delta}$$

.I.M. is the usual limit in mean definition and Δ is the sample time
each discrete sample. For computer applications, Δ is fixed and the
average of the square of Δy_n is more appropriate (for locally ergodic
es).

$$\Delta\tilde{y}_n^2 = \frac{1}{k\Delta}\sum_{j}^{n} k\,(y_j - y_{j-1})^2 \qquad \text{(E-9)}$$

$$= \frac{1}{k\Delta}\left[2\tilde{s}_n^2 + \frac{1}{k}\,y_{n-k}^2 - y_n^2 - 2\sum_{j}^{n} k\,y_j y_{j-1} - \tilde{m}_n^2\right]$$

where

$$\tilde{m}_n \overset{d}{=} \frac{1}{k}\sum_{j}^{n} k\,y_j$$

$$\tilde{s}_n^2 \overset{d}{=} \frac{1}{k}\sum_{j}^{n} k\left(y_j - \tilde{m}_n\right)^2$$

Thus, the sliding variance is a factor in the mean square sliding stochastic derivative (and has a shorter name). Reliable estimation of a significant derivative requires a small value of k (the number of points averaged) while reduction of stochastic variation requires a large value. By using the sliding variance, these requirements are partially reconciled by eliminating terms primarily due to stochastic noise. Also, the sliding variance is more stable than a simple difference of the sliding mean which reduces to $\frac{1}{k}\left(y_n - y_{n-k}\right)$.

Let us summarize the alternatives for selection of a total estimation process.

i.) Sub-interval length - This is the number of points to be summed corresponding to the first low-pass filter in Figure E-1. In order to compute the proper sliding averages, these intervals are non-overlapping rather than sliding.

ii.) Sliding average length - The value, k, in the formulas for the various estimators relates to both standard deviation and mean value.

) Envelope estimator - Either absolute value of the input signal or a Hilbert envelope (the square root of the sum of the squares of the input signal and its Hilbert transform).

Derivative estimators - For each of the envelope estimators we may define three derivative estimators:

One-Point Difference for Sliding Mean

$$(1) \qquad \tilde{\dot{a}}_1 = \frac{1}{k}\sum_{j}^{n}{}_k \; y_j^a - \frac{1}{k}\sum_{j}^{n-1}{}_k \; j_j^a \qquad \text{(E-10a)}$$

Sliding Standard Deviation

$$(2) \qquad \tilde{\dot{a}}_2(n) = \left[\frac{1}{k}\sum_{j}^{n}{}_k \left(y_j^a\right)^2 - \left(\frac{1}{k}\sum_{j}^{n}{}_k \; y_j^a\right)^2\right]^{1/2} \qquad \text{(E-10b)}$$

Mean Square Derivative

$$(3) \qquad \tilde{\dot{a}}_3(n) = \left[\frac{1}{k}\sum_{i}^{n}{}_k \left(y_j^a - y_{j-1}^a\right)^2\right]^{1/2}$$

where y_j^a is the j-th estimate of the envelope.

ote that the last two estimators only give the magnitude of the e but not the sign.

rder to make choices between these different alternatives, we resentation criterion to compare waveforms (which will be the timates of the underlying signal properties). We note that mean or is inappropriate for the types of comparisons we wish to The reason is its insensitivity to very sharp derivative An alternative criterion can be derived by use of the Chebyshev discussed in Equation (E-8). By algebraic manipulation of ality using the weighted difference between the two waveforms, ive at a criterion that gives a better comparison. For two

waveforms, $y_1(n)$ and $y_2(n)$, $n = 1, 2, \ldots N$ define the Chebyshev weighted error by:

$$\xi_c = \left[\frac{1}{2}\left(\frac{e^2(1)}{y_2^2(1)} + \frac{e^2(N)}{y_2^2(N)}\right) + \frac{1}{N}\sum_{n=2}^{N}\left[\left(\frac{e(n) - e(n-1)}{y_2(n) - y_2(n-1)}\right)^2\left(\frac{e(n)}{y_2(n)}\right)^2\right]^{1/2}\right]^{1/2} \quad \text{(E-11)}$$

where

$$e(n) = y_1(n) - y_2(n)$$

The assumption of local ergodicity must be invoked to relate this measure to the probability of exceeding a bound as in Eqn. (E-8) (much the same as the justification for mean square criteria). However, Eqn. (E-11) can be used to compare estimators. In Figure -2, two estimates and a smooth envelope are shown. The rapid variations are averaged by the mean square error computation so that the value for the two estimators is approximately equal (0.098 and 0.101). However, using the Chebyshev weighted measure, the difference in the two estimators is apparent, indicated by a calculated value of 0.2974 for the rapidly varying one and 0.1689 for the smooth one.

Envelope and frequency estimators must work in different situations ranging from slow but large magnitude variation and possibly a smooth frequency transition (such as during vowel formant portions) to rapid, small amplitude and frequency changes (which occur during fricatives). We will first consider the typical vowel onset which occurs in the order of 50 to 100 milliseconds (see Figure 3). In order to compare our different estimators, we will use the following idealized vowel onset waveform:

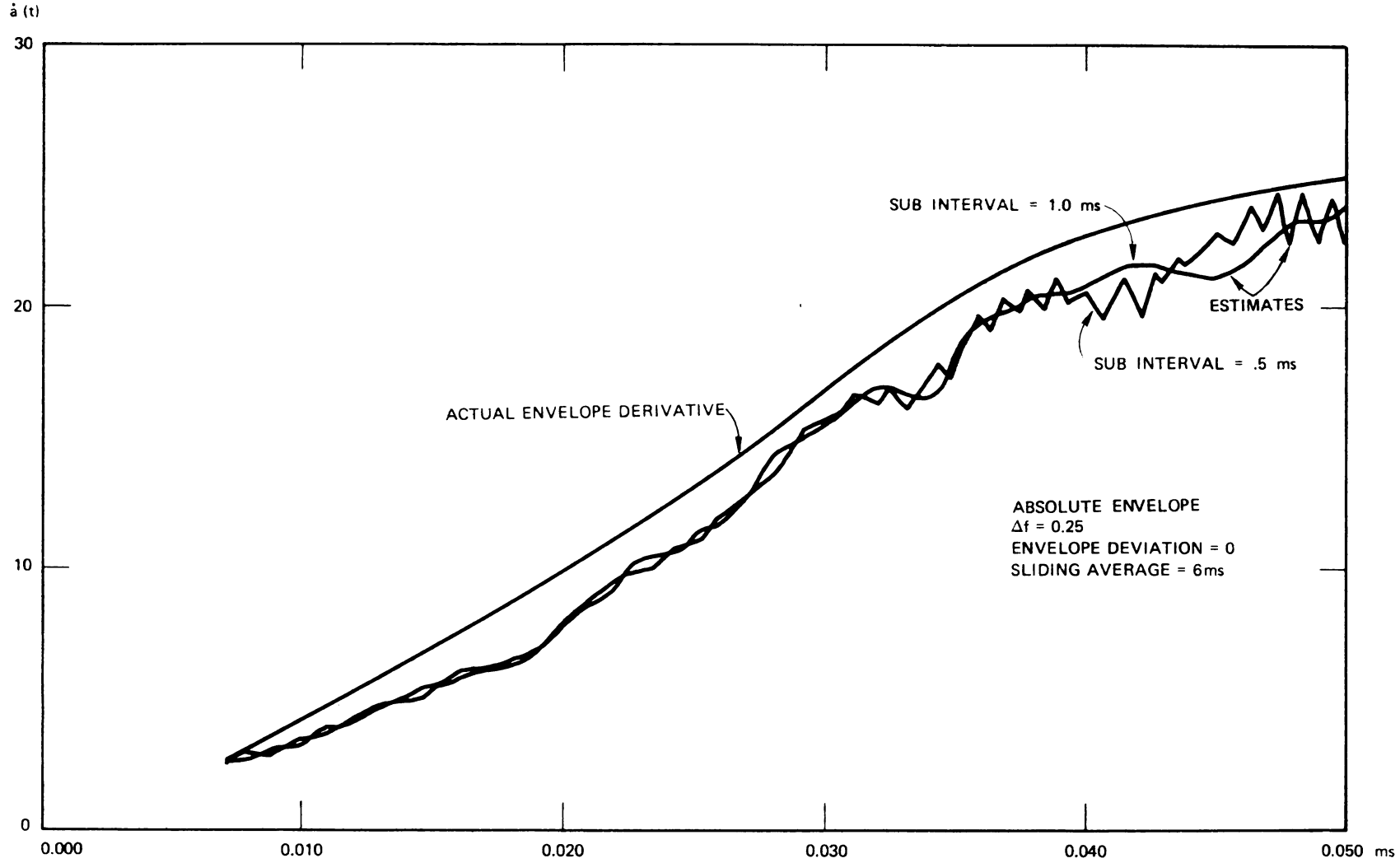


FIGURE E-2 TWO ENVELOPE DERIVATIVE ESTIMATORS

Let $\hat{x}(t) = a(t)e^{j\alpha(t)}$ $\quad 0 \le t \le .050$ (E-12)

Where

$$a(t) = a_o + \int_o^t \arctan\{a(s)/100\}\eta(s)\,ds$$

$$d(t) = \omega_o t + B_o \int_o^t \left\{3(s/.05)^2 - 2(s/.05)^3\right\} ds$$

a_o is initial amplitude (= 10.)

ω_o is initial frequency (= 2π 2000)

B_o is frequency deviation

$\eta(\cdot)$ is a I.I.D. r.v. with Gaussian distribution with mean independent of $a(\cdot)$ and $\alpha(\cdot)$ at time t with mean $E(\eta) = 1$ and standard deviation $\sigma^2(\eta) = a_2$. By selecting values for the parameters; frequency deviation; B_o, and amplitude noise standard deviation; a_2, we can generate time functions with complex nonlinear behavior in order to investigate the stability properties of the estimators we have chosen. Figure E-2 shows the ideal envelope derivative ($B_o = a_2 = 0$) and two envelope derivative estimators for the Hilbert envelope of x(t).

Even for this idealized model, we have five parameters to change in order to investigate the properties of the three envelope derivative estimators: (1) absolute or Hilbert envelope, (2) variation of subinterval length, (3) variation of sliding average length, (4) amount of frequency deviation, and (5) envelope standard deviation. Typical values for variations of number of these parameters are shown in Tables E-1, E-2, and E-3 and Figures E-2, E-3, and E-4. Table E-1 and E-2 show variation of the sliding average length for B_o=0 and various values of envelope standard deviation for both absolute and Hilbert envelope derivative estimators. Figure E-3 shows a typical plot from Table E-2. Table E-3 indicates variation

induced in envelope derivative estimators by frequency changes. Figure E-4 shows the Chebyshev weighted error for the three envelope derivative estimators as the subinterval length is increased (evaluation of the stability of the derivative estimators is sufficient to tell us about the estimation of the envelope itself). Rather than discuss all these data in detail, we will state the choice of estimator subinterval length and sliding average length and give reasons for that choice.

We note from Table E-1 and E-2 that the Hilbert and absolute envelope estimators have almost exactly the same values of Chebyshev weighted error when there is no frequency deviation. Because of the additional complexity in computing the Hilbert envelope, this would recommend the absolute value envelope as an estimator. However, Table E-3 shows that for frequency deviations the absolute value estimator gives an order of magnitude higher Chebyshev weighted error than the Hilbert envelope estimator.

Note that the behavior of the sliding standard deviation as an estimator of envelope derivative behaves much more stably and gives, in most cases, a lower Chebyshev weighted error. Table E-3 for absolute value envelope estimator shows this very dramatically. For this estimator, sub-interval lengths on the order of 0.5 to 2 milliseconds give approximately the same Chebyshev weighted error. This result can be anticipated from the form of the mathematical relationships between the three derivative estimators since both the 1-point sliding difference and mean square derivative have unaveraged terms that vary as the random samples. For this reason, as shown in Figure E-4 they are very dependent on the variation of the sub-interval average values.

TABLE E-1 ENVELOPE DERIVATE CHEBYSHEV WEIGHTED ERRORS USING HILBERT ENVELOPE ESTIMATOR

nvelope td. Dev.	Estimator 1	2	3
.2	.0540	.0916	.0544
.4	.1076	.1830	.1093
.6	.1613	.2742	.1649
.8	.2150	.3650	.2208
1.0	.2687	.4553	.2769

4 ms sliding average

1 ms sub interval

Envelope Std. Dev.	Estimator 1	2	3
.2	.0615	.0470	.0616
.4	.1230	.0939	.1227
.6	.1845	.1409	.1832
.8	.2461	.1879	.2432
1.0	.3076	.2348	.3037

6 ms sliding average

1 ms sub interval

avelope td. Dev.	Estimator 1	2	3
.2	.0256	.0396	.0280
.4	.0515	.0793	.0552
.6	.0776	.1196	.0816
.8	.1041	.1593	.1070
1.0	.1308	.1995	.1316

8 ms sliding average

1 ms sub interval

Envelope Std. Dev.	Estimator 1	2	3
.2	.0413	.0226	.0419
.4	.0827	.0456	.0830
.6	.1241	.0688	.1232
.8	.1658	.0922	.1622
1.0	.2074	.1157	.2002

10 ms sliding average

1 ms sub interval

TABLE E-2 ENVELOPE DERIVATIVE CHEBYSHEV WEIGHTED ERRORS USING ABSOLUTE VALUE ESTIMATOR

Envelope Std. Dev.	Estimator 1	2	3
.2	.0543	.0916	.0548
.4	.1085	.1830	.1102
.6	.1628	.2741	.1662
.8	.2170	.3649	.2226
1.0	.2712	.4552	.2791

4 ms sliding average

1 ms sub interval

Envelope Std. Dev.	Estimator 1	2	3
.2	.0614	.0470	.0616
.4	.1228	.0941	.1226
.6	.1843	.1412	.1831
.8	.2458	.1883	.2433
1.0	.3073	.2353	.3037

6 ms sliding average

1 ms sub interval

Envelope Std. Dev.	Estimator 1	2	3
.2	.0256	.0396	.0280
.4	.0514	.0793	.0553
.6	.0776	.1192	.0817
.8	.1040	.1593	.1072
1.0	.1307	.1995	.1318

8 ms sliding average

1 ms sub interval

Envelope Std. Dev.	Estimator 1	2	3
.2	.0413	.0227	.0419
.4	.0827	.0456	.0830
.6	.1241	.0687	.1231
.8	.1657	.0921	.1622
1.0	.2072	.1157	.2002

10 ms sliding average

1 ms sub interval

TABLE E-3 EFFECTS OF FREQUENCY CHANGE ON ENVELOPE DERIVATIVE ESTIMATORS-1 MS SUBINTERVAL

Sliding Average	Estimator 1	2	3
4	1.805	1.831	2.031
6	1.405	1.172	1.644
8	1.155	.8593	1.432
10	.8641	.6001	1.199

Absolute envelope derivative vs. sliding average
Freq. dev. = 0.25

Sliding Average	Estimator 1	2	3
4	.2066	.2599	.2057
6	.1077	.1369	.1122
8	.0990	.0693	.0993
10	.0836	.0383	.0826

Hilbert envelope derivative vs. sliding average
Freq. dev. = 0.25

Freq. Dev.	Estimator 1	2	3
.05	.7520	.2745	1.286
.10	.6717	.3400	.9978
.15	.6482	.2533	1.295
.20	1.035	.3862	2.532
.25	.8641	.6001	1.199

Absolute envelope derivative vs. freq. dev.
sliding average = 10 ms

Freq. Dev.	Estimator 1	2	3
.05	.0129	.0103	.0136
.10	.0750	.0255	.0852
.15	.1685	.0345	.2001
.20	.1354	.0388	.1472
.25	.0836	.0383	.0826

Hilbert envelope derivative vs. freq. dev.
sliding average = 10 ms

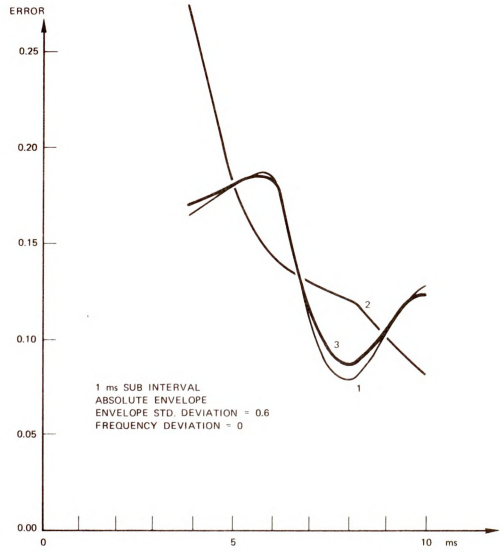


FIGURE E-3 CHEBYSHEV WEIGHTED ERROR FOR ENVELOPE DERIVATIVE ESTIMATORS AS A FUNCTION OF SLIDING AVERAGE LENGTH

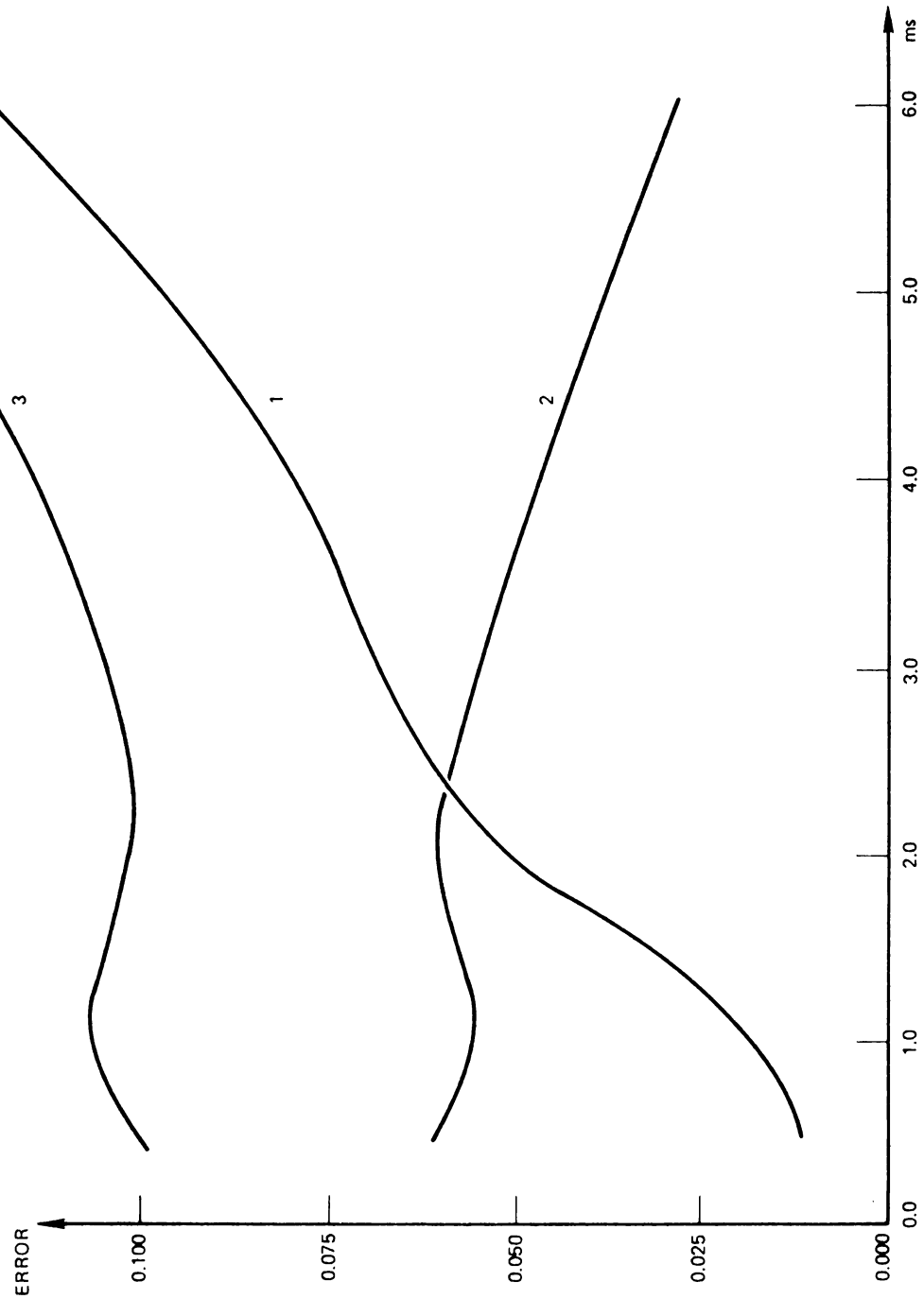


FIGURE E-4 CHEBYSHEV WEIGHTED ERROR FOR ENVELOPE DERIVATIVE ESTIMATORS AS A FUNCTION OF SUB-INTERVAL LENGTH

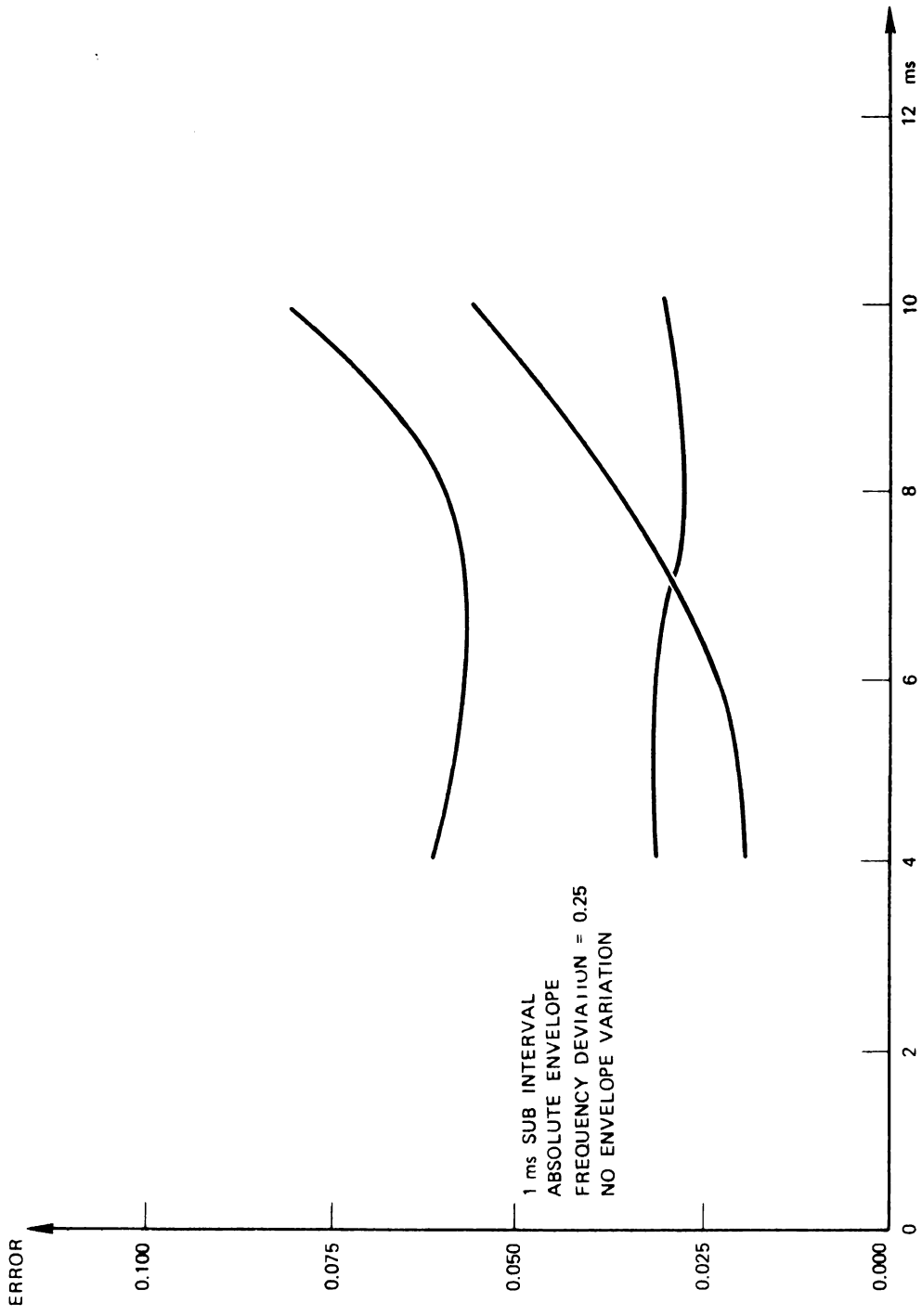


FIGURE E-5 CHEBYSHEV WEIGHTED ERROR FOR ENVELOPE DERIVATIVE ESTIMATORS AS A FUNCTION OF SLIDING AVERAGE

Figures E-4 and E-5 show an increase in the estimation error for larger values of subinterval length and sliding average. Reference to Figure 13 and the corresponding discussion of distortion in linear filters induced by large frequency changes explain this increase. The intuitive notion of longer averaging time can be misleading in this complex situation. The error rates in Figure E-4 and E-5 were derived for a(t) = 100.

There appear to be two sources of variance in envelope estimation: the first is induced by the small number of samples, which would require longer averaging times, and the second is the distortion caused by the time-varying parameters, which would require shorter averaging times. We must select a compromise value, which appears to be approximately 1 ms subinterval length and 6-10 ms sliding average length.

We may conclude that, for this idealized speech acoustical signal, the most stable estimator is the Hilbert envelope with sliding standard deviation as a derivative estimator. The sliding mean of the absolute value of x(t) gives some envelope estimation distortion, primarily during epochs with changing frequency, but requires much less computation.