22

RADC-TDR-64-312
Final Report

VOICE IDENTIFICATION TECHNIQUES

TECHNICAL DOCUMENTARY REPORT NO. RADC-TDR-64-312

September 1964

Intelligence Applications Branch
Rome Air Development Center
Research and Technology Division
Air Force Systems Command
Griffiss Air Force Base, New York

Project No. 4027 , Task No. 402701

(Prepared under Contract No. AF 30(602)-3043 by Information Sciences Laboratory, Data Systems Division, Litton Systems, Inc., 335 Bear Hill Road, Waltham, Massachusetts, 02154. Author: W. Floyd.)

FOREWORD

This report has been prepared in the Information Sciences Laboratory within the Data Systems Division of Litton Systems, Inc., a division of Litton Industries. The work reported here has been performed over a period of 12 months, under Contract Number AF30(602)-3043, entitled "Voice Identification Techniques". This project has been completed under the direction of the Intelligence and Information Processing Division within the Rome Air Development Center.

Several individuals within the Information Sciences Laboratory have made contributions to the development of speech processing techniques reported here. Experimental speech processing equipment has been constructed under the guidance of Mr. Arthur Crooke; computer programs have been prepared by Mr. Paul Connolly, Mr. Vito Maglioni, Miss Sarah Foster, and Miss Helen O'Shea. Mr. William Floyd has served as Project Director and Dr. George Sebestyen has provided technical consultation.

In addition to the other individuals in the Information Sciences Laboratory who contributed to this work, thanks are due to Mr. Mark Weiss and the Federal Scientific Corporation for their courtesy in providing an IBM Card Source Deck for the formant tracking routine utilized in the extraction of spectral clues.

AIR FORCE, Canterbury Press, Rome, N. Y., 9/16/64, 144

Key Words: Acoustics: sound pitch; sound signals.
Mathematics: statistical analysis; statistical data; statistical distributions.
Computers and Data Systems.

ABSTRACT

The problem of designing a machine which can identify automatically a speaker by processing samples of his speech has been investigated. Through construction of special purpose equipment and digital computer simulation, the fundamental operations of speech clue extraction and voice recognition have been implemented in a manner designed to render voice identification by a machine insensitive to the effects of noise and speech bandwidth truncation. Results of experiments indicate that the extraction of certain spectral, voice fundamental and speaking cadence chara teristics of speech can lead to reliable identification of a speaker's voice, even in undesirable speaking, and speech transmission environments.

PUBLICATION REVIEW

This report has been reviewed and is approved. For further technical information on this project, contact Mr. Bruno Beek, RADC (EMIAP), Ext. 25122.

Approved: Bruno Beek
BRUNO BEEK
Electronic Engineer
Electromagnetic Processing Section

Approved:
ROBERT J. QUINN, JR.
Colonel, USAF
Chief, Intel and Info Proces Div

FOR THE COMMANDER: Irving J. Gabelman
IRVING J. GABELMAN
Chief, Advanced Studies Group

EVALUATION

The "Voice Identification Techniques" program was initiated to provide the Air Force with a capability of automatic talker recognition. The present development employs a pattern recognition technique for the learning and decision portions in the overall recognition system. In general, two types of recognition have been examined: (1) detection of a single speaker from a given group of speakers, and (2) identification of each of a pre-specified group of speakers.

The program has undergone two phases of study; in the first phase automatic speaker recognition was demonstrated on a computer with a manual input of the various speaker characteristics. The second phase was concerned with the automatic extraction of the speakers characteristics, where the speech signal has a truncated bandwidth and is degraded by noise.

At the present time, speaker identification can be performed with low probabilities of error and low probabilities of false alarm. The system has the potentiality of being completely automatic in the learning and recognition phases; however, this has not been completely demonstrated.

A future effort will stress (1) automatic processing in a noisy and limited speech bandwidth environment; and (2) processing of foreign languages and dialects.

The results of this detailed investigation and continued effort will eventually merge with the Word Recognition program to provide an automatic system for speaker identification and the transcription of spoken English into printed text. The process subsequently will lead to the same capability in other languages.

ROBERT J. QUINN, JR.
Colonel, USAF
Intel and Info Proces Div

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF ILLUSTRATIONS (cont.)

LIST OF ILLUSTRATIONS (cont.)

LIST OF TABLES

1. INTRODUCTION

This report presents a summary of work performed on Contract AF-30(602)-3043 entitled "Voice Identification Techniques". The purpose of this project has been to execute certain steps toward the development of techniques for automatically processing speech signals to render an identification of the speaker. The basic ideas underlying the approach taken on this project were published in the final report on Contract AF-30(602)-2499.[1*] This earlier work demonstrated the potential of certain operations for rendering reliable speaker recognition decisions, but involved manual simulation of some of these operations. Thus, one of the goals of the current project has been to automatize these operations. A second goal has been to ascertain the effects of noise degradation and bandwidth truncation of speech signals on the recognition performance of a speaker identifier.

Over the past few years several approaches to the problem of voice identification have been investigated, and varying degrees of success attained.** Most of the methods which are aimed at a minimum reliance on human judgment have been demonstrated or tested on a relatively small quantity of speech and primarily through the use of general purpose computers for simulating the recognition processes. In spite of varying approaches to the problem, all of these methods may be regarded as particular methods of performing the operations indicated by the block diagram in Figure 1. The incoming speech is processed first by a Speech Clue Extractor, which performs measurements leading to the generation of clues*** which are designed to allow for convenient

---

* References are designated by numerical superscripts corresponding to the reference numbers in the Reference List at the end of this report.

** Including investigations at Bell Laboratories,[2,3] and Lincoln Laboratory,[4].

***Various other terms have been used to describe such quantities, including "parameters", "measurements", "attributes", "characteristics", "features", and "properties".

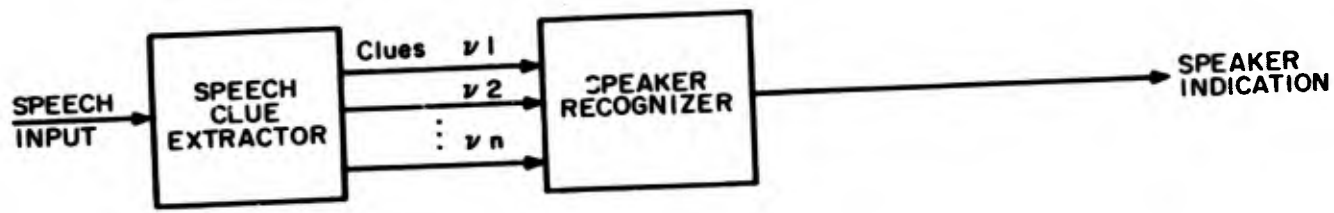


Figure 1. Generic Block Diagram of a Speaker Recognizer

identification of the speaker through their measurement. It is usually possible to regard each of the speech clues generated by the Clue Extractor as a voltage amplitude versus time waveform. Examples of such clues are the pitch of the speaker's voice, formant frequency locations, and speech envelope amplitude.

The next and final major processing step involved in identifying a speaker is performed in the Speaker Recognizer. In this unit the clue waveforms generated by the Clue Extractor are generally sampled, quantized and compared in some way with stored information obtained through earlier examinations of the speakers whom it is desired to recognize. On the basis of this comparison between incoming clue values (arising from an unknown speaker) and stored information on previously specified speakers, a decision is rendered and emitted by the speaker recognizer.

It is possible to utilize humans to perform one, both, or neither of the clue extraction and speaker recognition operations. For instance, today most of the working systems for speaker authentication simply employ a human to perform both tasks, without any attempt at conscious formulation of speech clues as an intermediate step. However, the use of humans for either task opens up possibilities for malfunctioning of a voice recognition system which an automatic device would preclude. For instance, in an application involving a fixed set of speakers over a long period of time, a human's a priori anticipation of one of the known speakers may prevail and cause him to miss a deceiver. Or, on the other hand, for applications requiring frequent changes in the speaker sets, a human's ability to distinguish between speakers may not be very high, particularly in the presence of noise.

In considering automatic ways to solve the voice recognition problem, it is necessary that an explicit investigation of candidate speech clues be undertaken. The primary quality sought in a set of clues is that the clue waveform patterns take on different values corresponding to intervals of

time during which different speakers are talking. Once a set of speech clues has been selected, the problem of recognizing the speaker becomes a standard problem of pattern recognition.

Several anatomical explanations accounting for the uniqueness of a speaker' s voice have been given in the past. It has been anticipated that speaker identification capability is enhanced by the uniqueness of our vocal cavities and articulators. The vocal cavities (like resonant circuits) cause energy to be concentrated in specific regions of the spectrum dependent on the cavity sizes and their method of coupling. The major cavities affecting speech are the throat, the nasal and the two oral cavities formed in the mouth by the positioning of the tongue. The size and the manner in which the vocal cavities are coupled evidently can account to a great extent for the identifiability of a specific person' s voice. This is particularly true of the nasal cavity which is not controllable or manipulatable by the talker. These cavities correspond to the third and fourth formant frequencies.

While these stationary properties of the vocal cavities are important in determining the speaker' s characteristics, evidently the manner in which the articulators are manipulated during the utterance of connected speech is of even greater significance. The articulators are the jaw muscles, the teeth, and the tongue; and their systematic interplay results in speech. Thus, even though most persons can produce sounds whose formant frequencies ocassionally agree identically (within measurement errors) with formant frequency combinations occurring in another person' s voice, the relative frequency with which one person produces a specific formant combination differs from the relative frequency with which another person produces the same sound. The manner of transition from one sound to the next also differs for different speakers.

During an earlier study performed in the Information Sciences Laboratory[1], a set of thirteen speech clues exhibiting some of these dynamic and static attributes were examined and found useful in representing speech

at any given instant. These were the four formant frequency locations, their derivatives, speech signal amplitude and its derivative, pitch frequency and its derivative, and the binary clue designating whether or not a sound is voiced or unvoiced. The results obtained using these clues indicated that in a controlled environment speaker recognition can be performed automatically with very low error rates. In the current investigation, emphasis was therefore placed at the outset on the use of these same clues in a noisy, truncated bandwidth speech signal environment. This emphasis has entailed the replacement of the manual formant extraction simulation techniques with a completely automatic combination of special purpose equipment and a digital simulation program, design and construction of a noise-resistant pitch extractor, and the development of modified data collection and intermediate storage techniques. The resulting speech processing system has allowed for the investigation of additional speech clues which are expected to exhibit improved resistance to the effects of corrupting environmental conditions.

The experimentation performed with this speech processing system has produced indications of the effects of variations in signal-to-noise ratio, the duration of speech signals available for rendering decisions, the number of speakers to be recognized, the types of clues utilized, and bandwidth truncation of speech signals. Detailed descriptions of the special purpose equipment, computer programs, and data processing experiments and their results are presented in Section 2 of this report. Also discussed are salient aspects of techniques for selecting and extracting the speech clues utilized, and the pattern recognition techniques utilized to perform the voice recognition unit functions.

The major conclusions and recommendations concerning further development and use of the voice identification techniques investigated on this project are presented in Section 3. Four Appendices and a Reference List complete the report.

2. TECHNICAL DISCUSSION

The accuracy with which the voice identification problem can be solved with an automatic device is strongly dependent on several factors, ten of which are listed in Table 1. During the course of this project, equipment has been constructed and experiments performed to ascertain the effects of variations in several of these factors on the performance attainable with a certain class of speaker identification devices.

In particular, two types of recognition have been examined: (1) detection of a single pre-specified voice in the presence of other voices, and (2) identification of each of a pre-specified group of speakers. These have been designated the speaker detection and classification problems, respectively. No knowledge of uttered text has been presumed for these experiments, and no restrictions on the spoken text have been imposed. Further, it has not been presumed that any of the spoken text available for learning the peculiar characteristics of a speaker's voice is contained in the spoken text on which the recognition is to be based.

With the exception of these and the ninth factor (speaker emotional and physical states) listed in Table 1, variations in all of the remaining factors have been examined to some extent during the course of this project. The major steps involved in this investigation have been the following:

- Selection of speaker identification speech clues
- Implementation of clue extractors through either equipment construction or digital computer simulation
- Selection of Speaker Learning and Recognition Methods

- Preparation of speech recordings for different environmental conditions
- Processing speech signals in experiments to ascertain recognition performance attainable with the selected learning and recognition methods under the different environmental and decision-making conditions.

Each of these phases of the project is discussed in the following four subsections.

1. TYPE OF RECOGNITION DESIRED
   (Simple Detection or Specific Identification)
2. KNOWLEDGE OF TEXT
   (Pre-specified or Unknown)
3. BANDWIDTH TRUNCATION OF SPEECH
4. NOISE CORRUPTION OF SPEECH
5. SIZE OF SPEAKER GROUP
6. MAKE-UP OF SPEAKER GROUP
   (All male or mixed, common language or not, etc.)
7. AMOUNT OF SPEECH PROCESSED FOR LEARNING
8. AMOUNT OF SPEECH PROCESSED FOR RENDERING A DECISION
9. VARIABILITY OF SPEAKER EMOTIONAL AND PHYSICAL STATES
10. COMPLEXITY OF EQUIPMENT
    (Number of Clues Utilized, Clue Resolution Capability, Storage Capacity of Recognition Unit, etc.)

Table 1. Ten Factors Affecting Accuracy of a Voice Identification System

2.1 SELECTION OF SPEECH CLUES FOR VOICE IDENTIFICATION

As mentioned in the Introduction to this report, it has been anticipated that a voice identification capability may be based on the uniqueness of a speaker's vocal cavities and articulators, and the manner in which the articulators are manipulated during the utterance of speech. During an earlier study[1] performed in the Information Sciences Laboratory a set of thirteen speech clues exhibiting some of these static and dynamic characteristics were examined and found useful for representing speech at any given instant. These were the four formant frequency locations, the speaker's pitch, the voice amplitude, derivatives of these six quantities, and a binary clue designating whether or not a speech sound is voiced or unvoiced. The results of experiments performed with these clues indicated that fairly high recognition performance could be obtained through their measurement, under ideal conditions (e.g., in the absence of any noise).

Thus at the outset of the current investigation, primary emphasis was placed on completely automatizing the extraction of these same clues, in a manner which would render their measurement insensitive to the effects of noise, bandwidth truncation, and a limited duration of speech signals available for the learning and recognition tasks. However, it was anticipated that the formant-related clues would be more sensitive to the effects of noise than the pitch- and speaking cadence-related clues, and that therefore the original set of thirteen clues should be augmented. To this end the question of clue selection for voice identification in a non-ideal environment has been re-examined.

It is not fundamentally necessary that a candidate speech clue be aimed at the measurement of any particular, pre-identified speech characteristic, such as first formant position, in order for the clue to be useful for voice identification. Rather, it is primarily only necessary that clue values take on different values during periods of speech uttered by different speakers.

Thus, for instance, a set of speech clues might be (and has been on occasion) generated by periodically sampling the envelope detected outputs of a bank of band-pass filters, beginning at the onset of each voiced interval, and continuing over a prescribed period of time. If n filters are employed, and m samples are taken then each of the mn different filter output samples generated by this process could be regarded as a clue coordinate in an (nm)-dimensional measurement space. This type of speech clue set is generally selected on the basis of empirical evidence that all of the clues taken together represent speech sufficiently well to preserve the salient differences (whatever they may be) between voices.

Although the notion of selecting such a "sufficient" set of speech clues is appealing in that a detailed investigation of the "salient characteristics" of speech signals generated by a single speaker is obviated by its adoption, there are definite drawbacks to this approach. One reason for this is that it is not sufficient that different clue values result from the speech of different talkers.

The other desiderata for a set of speech clues, particularly if automatic voice identification is to be accomplished within reasonable economic constraints, are that

- The speech clue generated by a speaker should be insensitive to: (a) the environment in which the speaker is talking, (b) any noise or bandwidth truncation imposed on the speech signal during transmission, and (c) normal variabilities in the speaker's manner of talking.
- The speech clue values generated by a speaker should occupy positions in the associated measurement space which allow for the use of economical recognition techniques to perform the speaker identification task.

If a speaker is talking in a noisy environment, or the speech signals are limited to a 3 kc band, or no information is available on the spoken text, then the difficulty of selecting speech clues which will still provide a high degree of discriminability between speakers is much more severe than it would be under ideal conditions. Or if a set of speech clues generates multiply-connected regions* in the measurement space for each of the speakers involved, then fairly complex operations may be required to perform the recognition function with either an automatic device or a human.

These additional considerations have dictated a systematic approach to the selection of speech clues for use in the voice identification task, rather than attempting a grand coup through the utilization of all potential clues within sight.

The number of potential candidates for voice identification clues which have been suggested over the years is quite large. Foremost among those which have not been listed previously are formant amplitudes and bandwidths, instantaneous spectral mean and variance, average pitch value and variation within each voicing interval, average (over a few seconds) voiced interval length, and relative spacing between successive speech element boundaries.

As outlined above, the utility of these and other speech clues for voice identification must be measured first in terms of the degree to which different speakers generate different clue values in the presence of whatever disturbance the speech signals may be subjected to. Secondarily, the way in which clue values are distributed in their associated measurement space must be evaluated to determine the complexity required of recognition routines to accomplish the discrimination between speakers. The major steps involved in determining the utility of speech clues in these terms are:

---

*To be discussed in Section 2.2

- Preparation of speech recordings under prescribed environmental conditions, for each of a group of speakers.
- Processing these recordings to extract a quantity of samples of the speech clues for each of the conditions and speakers involved.
- Compilation of statistics descriptive of the behavior of these speech clues for each of the conditions and speakers.
- Recording and processing these statistics in a form which sheds light on the degree to which clues take on different values under different conditions and for different speakers, and which indicates the type of recognition operations required to distinguish between speakers.

Considerable effort is required to perform the operations indicated in the last three of these four steps. Specifically, in order to extract a statistically significant quantity of speech clue values within reasonable bounds on cost and time, automatic clue extractors must be available either in the form of special purpose equipment, or through the use of a general purpose computer. If a computer is utilized to perform the clue extraction operations, then a capability for essentially real-time input of speech data is highly desirable.

The statistics referred to in the third step above may also present severe processing requirements. On the one hand, it may be regarded as sufficient only to compile first order probability distributions of each of the clues being studied, with a view toward selecting those clues which, when each is examined alone, take on different values during intervals of speech generated by different speakers. The data processing involved in this approach to speech clue selection is quite modest. On the other hand, it is quite clear from our experience that good speech clues will sometimes be discarded by this approach. Instead of examining each speech clue separately, it is necessary to examine clues jointly.

The illustration in Figure 2 serves to indicate how this necessity arises. When the two speech clues, v_1, and v_2, are measured simultaneously, the clue values resulting from two speakers occupy two completely non-overlapping regions in the two-dimensional clue measurement space.

If, however, either of the two clues is examined separately, then the clue values which result from the two speakers overlap considerably. Thus, on the basis of univariate examinations of each of these clues, neither would be regarded as useful for discriminating between speakers; yet, taken together these clues separate the two speakers perfectly.

The need is therefore for estimates of the multivariate distributions of speech clue values for different speakers. Unfortunately, as the number of clues increases, the amount of data required to achieve statistical significance in such estimates also increases. This effect can generate a severe data processing requirement if a large number of speech clues are involved.

Once a quantity of speech clue data has been processed to ascertain the distribution of clue values in measurement space for different speakers, not only the question of how well the clues separate speakers, but also how complex the recognition technique must be, can be answered. If, for, instance, in the problem of distinguishing between two speakers, clue values for each speaker lie within non-overlapping, singly-connected regions as illustrated in Figure 2, then a simple linear discriminant function (see Section 2. 2) may be utilized; i. e., a sample pair of clue values may be associated correctly with one of the speakers according to which side of a straight line* the sample falls. This type of operation is quite simple to instrument. However, a more likely distribution of clue values for two speakers is illustrated in Figure 3, in which the clue values from one speaker lie in multiply-connected regions. A more complicated (nonlinear) method of partitioning the measurement space into regions corresponding to speakers

*In measurement spaces having a larger number of dimensions, a hyperplane.

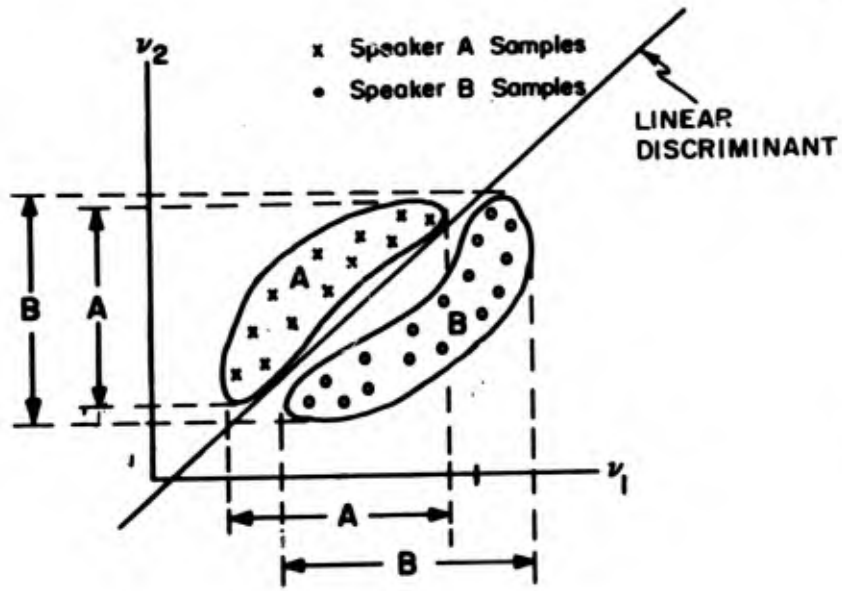


Fig. 2. Separation of Two Speakers with Two Speech Clues

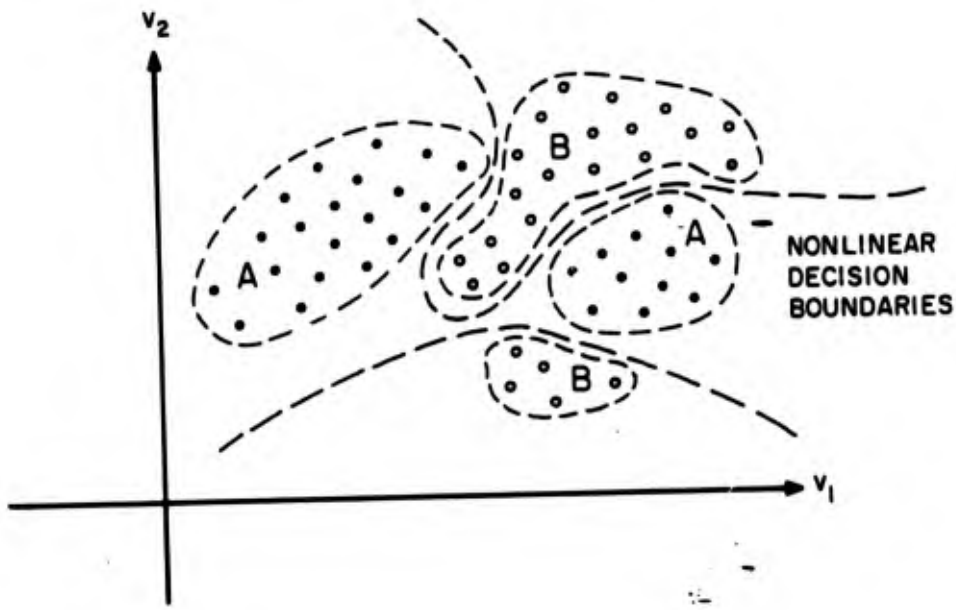


Fig. 3. Multi-Modal Distributions of Clue Values for Two Speakers

is required in this situation. This situation is discussed further in Section 2. 2.

Within the guidelines of this systematic approach to the selection of speech clues for voice identification, two groups of candidate clues have been examined. The first group, hereinafter referred to as the "spectral clues", are essentially those utilized during the earlier study. Specifically, this group consists of the following sixteen clues: three formant frequency locations and their derivatives and amplitudes, pitch and its derivative, the speech envelope and its derivative, two estimates of normalized speech signal energy, and a voicing indication.

The formants have been extracted in the following way. The speech signal is passed through a high-frequency pre-emphasis network (6 db per octave above 1 kc) and a short time-constant AGC circuit, into a 40-channel real-time spectrum analyzer. During voiced intervals of speech the spectrum analyzer output is sampled and A/D converted. As is the case for all of the spectral clues, the spectrum analyzer output is sampled periodically at a rate of 50 samples per second. The resulting sequence of digitized "instantaneous spectra" is processed in a digital computer to produce estimates of formant frequency locations. The essential operations are (a) location of peaks in each instantaneous spectrum, (b) comparison of peak location patterns for adjacent samples, (c) establishment of tentative tracks of associated peak locations, and (d) selection of formants from the set of tentative tracks generated within a voiced interval.

Having determined estimates of formant locations in the manner just indicated, their derivatives and amplitudes have been estimated by means of simple differences between adjacent sample values and noting the relative amplitude of the corresponding spectral channel, respectively.

All of the remaining seven spectral clues have been extracted directly with special purpose equipment, details of which are presented in Section 2. 3.

The second group, which has been called the "rudimentary" speech clues, is characterized by the absence of spectrum envelope information, and the generation of a single sample for each interval of time within which a switch in the speech processing equipment has indicated that a speaker' s vocal chords were vibrating: the average pitch, the average value of the derivative of pitch, the average value of the magnitude of the pitch derivative, the maximum value of pitch, the minimum value of pitch, the pitch peak-minimum order, the average value of the derivative of the speech envelope, and the duration of the voiced interval. These clues have been extracted through a combination of equipment and digital simulation.

Experiments with both spectral and rudimentary clues are reported in Section 2. 4. The equipment utilized in their extraction is described in Section 2. 3, and the class of pattern recognition techniques utilized to process these clues to render automatic voice identifications are discussed in the following Section 2. 2.

2.2 Automatic Voice Recognition Techniques

In the past four years research in the pattern recognition field has yielded a variety of techniques for automatically processing clues for rendering classification or detection decisions. Many of these recognition techniques have made dramatic entrances to the information processing arena, bearing a variety of mnemonic labels. The titles "Perceptron, Cybertron, CHILD, SCEPTRON, Conflex, Cynthia, Cyclops and Adaline," all refer to techniques or devices associated with the problem of associating a pattern (of clue values) with one of a set of classes of events.

In spite of the wide variety of names and sources* of such techniques, essentially all of the pattern recognition methods proposed or developed to date can be described and analyzed in a single, fairly simple way. Specifically, the diagram in Figure 4 shows the basic elements of the classification problem of pattern recognition. Underlying the pattern classification problem is the pre-specification of a set of classes of events which it is desired that an automatic means be developed to distinguish between. In the voice identification problem, each speaker involved in the ensemble may be considered a class. In the speaker detection problem there are essentially two classes of interest: (1) the desired speaker and (2) all other speakers. When individual identification of a prescribed group of M speakers is desired, there are M classes. To perform the discrimination between classes as they occur, a pattern recognition system is composed of two basic units: an observation system and a recognition system. The observation system accepts whatever events as occur in the real world as its inputs, and emits outputs which may be called "observations", or as we have been calling them, "clues", which represent a particular view of the real world. The function of the recognition system is to process the clues submitted to it, and emit at various times indications of estimated class membership of the real world events. In the current application, the "real world events"

*Including psychologists, mathematicians, and engineers.

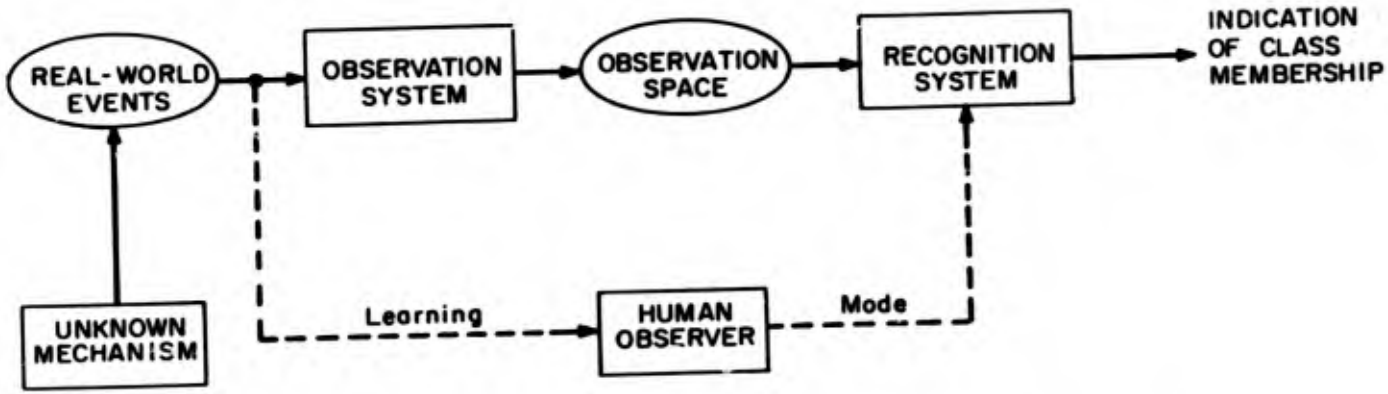


Fig. 4. The Classification Problem Model

available for processing by a pattern recognition system are speech signals. The observation system consists of the spectrum analyzers, pitch estimators, voicing detectors and other clue extractors which may be utilized as discussed in the preceding section of this report. The clue values emitted by the observation system may be viewed as points or patterns in an "observation", or "measurement" space, in which each clue is associated with a coordinate direction.

As outlined briefly in the beginning of this section, the function of the recognition system is to partition the measurement space into non-overlapping regions, associating each region with one of the pre-specified classes. To attain this capability, there are generally two phases of operation of a recognition system. During the first, learning phase, sample events are submitted to the recognition system for the purpose of allowing the system to locate or "learn" the proper positions of the class boundaries in measurement space. In most systems, a human participates in this phase of operation at least to the extent of informing the recognition machine of the correct class membership of the "learning" samples. This has been called "learning with a teacher". The current stage of development of other systems which "learn without a teacher" is such that they should not be considered for the voice identification application.

Following the learning phase of operation, most pattern recognition systems maintain fixed class boundaries in measurement space, and operate automatically without human intervention.

All pattern recognition systems differ from one another in two fundamental aspects: (1) the degree and nature of the human' s participation in the selection of class boundaries during the learning mode of operation, and (2) the geometrical constraints on the types of class boundaries which can be selected. For most applications, the differences between the ways in which the human participates in class boundary selection are largely academic. In some systems a "reward

and punishment" routine is used to adapt the class boundaries to the learning samples, and in other systems the adaptation is performed automatically. But these differences may be regarded as a matter of personal taste.

With respect to the second characteristic in which recognition systems differ – the constraints on class boundaries which can be selection – most applications require careful consideration. In particular, many pattern recognition systems proposed to date may be classed as linear discriminant techniques; that is, the class boundaries which can be adjusted in measurement space are hyperplanes, or linear combinations of the clue values. In two dimensions this means that the machines perform the recognition task by observing which side of a straight line a sample falls on. In Figure 2, for instance, it is easily seen that all the patterns generated by speaker A lie on one side of the dotted straight line, and the patterns generated by Speaker B line on the other side of the same line. In this example, the task of a linear discriminant recognition machine would be to locate the proper position of this straight line so that correct identification of subsequently observed clue patterns could be accomplished.

While a linear discriminant suffices to partition the measurement space perfectly well in the example in Figure 2, it is easily seen in Figure 3 that there exists no straight line which places all the clue patterns for speaker A on one side and those for speaker B on the other side. It is clear that if the clues generate such a multiply-connected region as illustrated for speaker A in Figure 3, then a more sophisticated method of partitioning the measurement space into regions corresponding to classes must be utilized.

During the past three years several approaches to the selection of non-linear class boundaries have been investigated by personnel in the Information Sciences Laboratory. In the usual case in which very little is known in advance about the distribution of clue patterns, we have found that a good

approach to the selection of class boundaries is to utilize a machine learning method which involves estimation of the joint probability distribution of clue values over the measurement space. Once estimates of these statistics are available, the decision-theoretic optimum method of deciding class membership of new samples may be implemented. The decision procedure may be stated simply*:

> Given a sample pattern of clues, this pattern is classified as belonging to that class for which the conditional probability of occurrence of the pattern is highest.

The most important details of this approach to automatic pattern recognition are reviewed in Appendix II of this report. The important point to note here is that the boundaries between class regions in measurement space generated by this technique are basically unlimited. In the example of Figure 3, for instance, the probability density function of points in the regions labeled A would be high for speaker A and low for speaker B, and the other way around for points in the region labeled B. Thus, by the above decision rule all of the points in the A regions would be associated with speaker A and similarly for speaker B.

As stressed in the earlier discussion of desirable speech clue qualities, the distribution of clue values in the associated measurement space has a direct effect on the cost and complexity of a recognition device. Linear discriminants, for instance, are relatively easily instrumented. On the other hand, the decision-theoretic, or maximum likelihood method outlined above may be (but not necessarily) much more expensive to implement. The course which we recommend to balance the potential complexities of non-linear recognition techniques against the potential simplicity afforded by

---

*This is the simplest form, which follows from the assumptions of equal a priori occurrence of classes, and equal cost associated with all misclassification errors.

linear methods is to first ascertain the distribution of clue values in the measurement space for the different classes involved, and then choose the simplest method of fixing the class boundaries which will perform the recognition task adequately. The latter step may involve only linear operations or may demand other types of boundaries. In either case, the collection of statistics dictated by the first step appears to be indispensable.

During the course of this project, two different methods of estimating probability density functions of clues, corresponding to the two types of clue sets investigated ("rudimentary" and "spectral"), have been utilized. With the rudimentary clues, histograms were constructed over a fixed cell structure, as the means of estimation. This method was dictated by the limited amount of data available for these clues.

A much more sophisticated estimation procedure has been utilized with the spectral clues. Specifically, the "local reference representation" method described in Appendix II has been employed with gaussian functional forms with adaptive mean and variance parameters. A complete discussion of this method appears in Reference 10. Results obtained for selected subsets of clues are presented in Subsection 2.4 below.

2.3 DESCRIPTION OF SPEECH PROCESSING EQUIPMENT

The special purpose speech processing equipment utilized to generate data and clue samples for voice identification experiments on this project is outlined in the block diagram of Figure 5, and illustrated in Figure 6. The primary input sources are (a) recorded speech signals and (b) noise generators. Most of the experiments reported in the next section have been conducted for broadband noise conditions, as provided by the GR-1390B noise generator. To create the noisy, bandlimited speech signals utilized in these experiments, originally high-quality speech recordings were added to the noise in the combiner, and the resulting waveform passed through a tunable bandpass filter (allison 2ABR). By monitoring a VU meter, the signal-to-noise ratio* could be controlled through a step attenuator at the noise generator output.

Primarily to allow for greater resolution in high-frequency spectral peaks, a 6 db per octave pre-emphasis circuit was incorporated at the input to the clue extraction equipment. The speech envelope (E), its derivative (E') and the speaker's pitch (F_o) were obtained from the resulting waveform. To obtain a reliable, relatively smooth estimate of pitch from noisy, bandlimited speech a special extractor was constructed for this project. This unit is described in Appendix IV. For bandlimiting to 3 kc (300-3300 cps) this device produces a reliable pitch indication for signal-to-noise ratios above 4 db. As discussed in Appendix IV, a more sophisticated pre-filtering arrangement would allow for reliable operation at lower signal-to-noise ratios.

As a preliminary to voicing detection and spectral analysis, the speech waveform is passed through short-term (20 msec) AGC circuit. Voicing has been detected with a conventional comparison of envelope-detected outputs of high- and low-pass filters.

*Signal-to-noise ratio (in db) has been defined as the difference between the average peak VU reading during an utterance in the absence of noise and the VU reading in the presence of noise alone, both measured after bandlimiting.

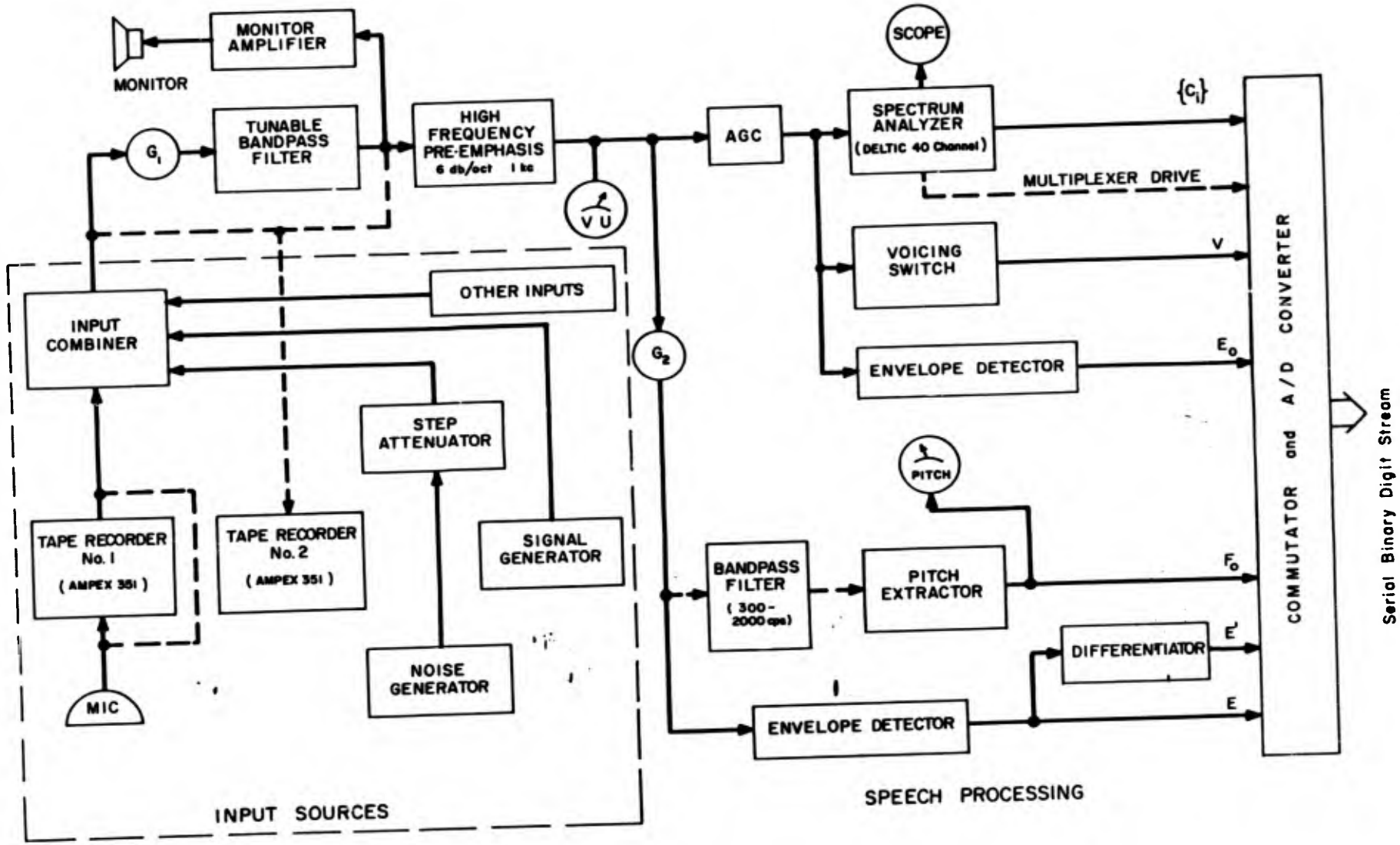
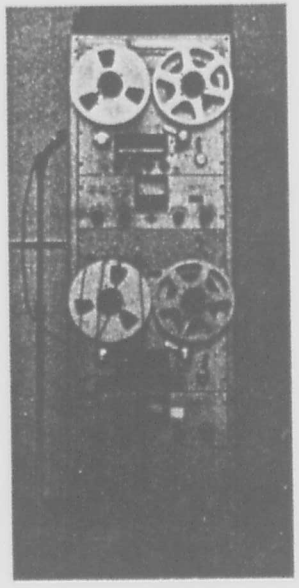
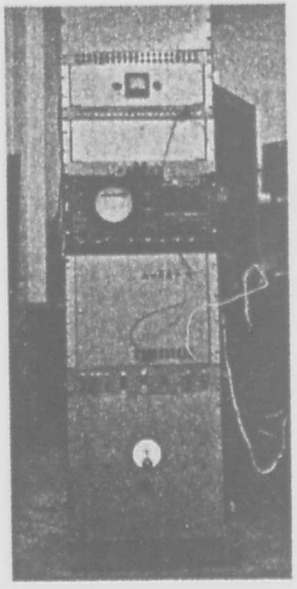


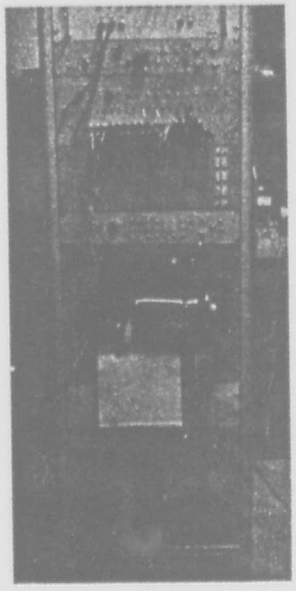
Fig. 5. Block Diagram of Speech Processing Operations



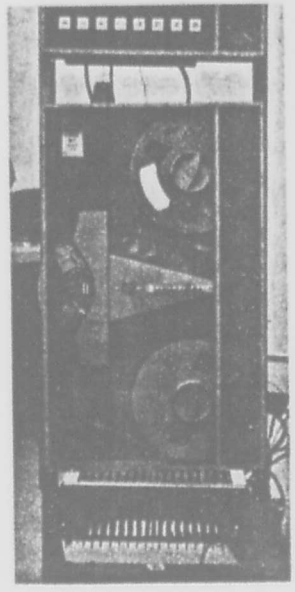
a) Recording



b) Speech Processing



c) Commutation and A/D Conversion



d) Magnetic Tape Transport

3043-102

e) Recomp II

Fig. 6. Speech Processing Equipments

Spectrum analysis is performed through the use of a delay line time compressor (DELTIC), followed by a single scanning filter. Although this unit is capable of 50 cps resolution, the analyzing filter bandwidth was set to 230 cps for this project, in order to allow for some smoothing of the spectral energy density envelope. Details of the major design considerations for this device are presented in Appendix III.

To provide a smooth spectral input to a formant extraction routine, the analyzing filter in this device scanned the 3 kc range from 300 cps to 3300 cps in 40 steps, or channels, covering 75 cps per channel. The resulting 40-character samples $\{c_i\}^{40}$, were combined with the other clues in a multiplexer and A/D converted. Sample outputs of the spectrum analyzer are shown in Figure 7. The clue samples and spectrum analyzer outputs were thus transformed into a serial digital data stream, suitable for recording on magnetic tape, or for direct input to a digital computer.

Because of a delay in delivery of the digital computer originally planned for this project, the resulting digital data were recorded on magnetic tape for later use with other computers, including the Recomp II shown in Figure 6. Details of the operations performed on this data appear in the following subsection.

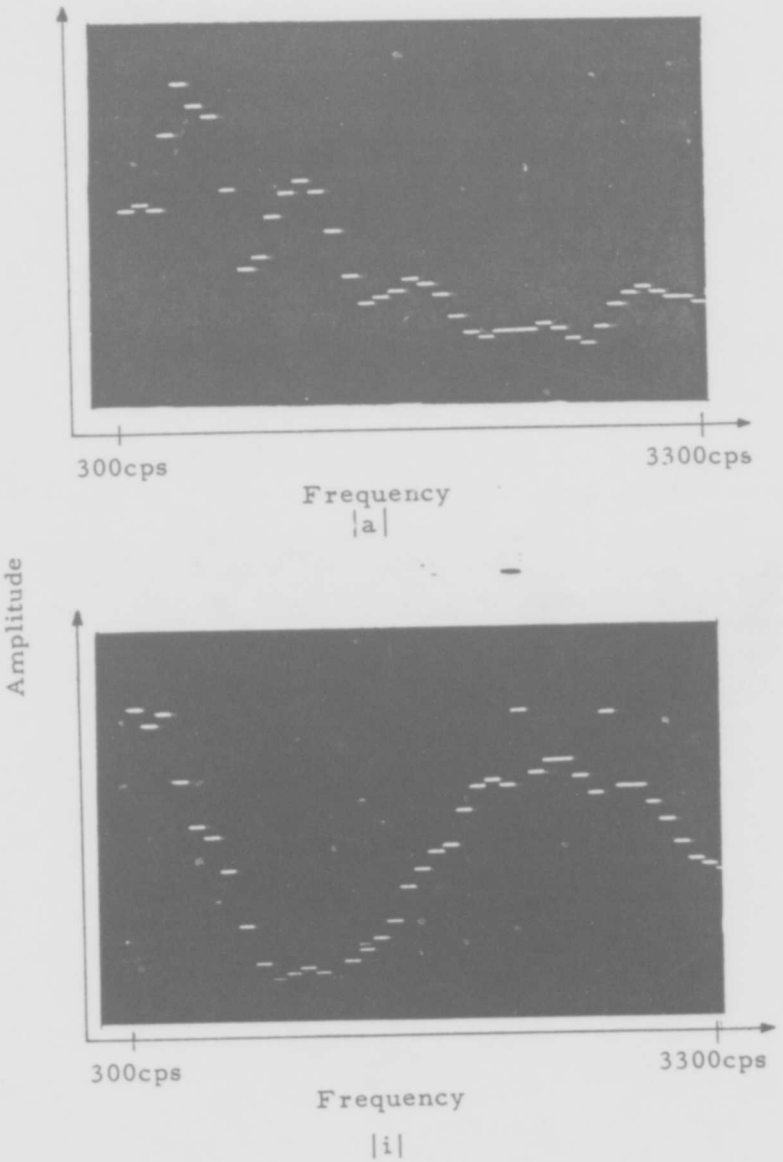


Fig. 7. Short Term Digitized Energy Density Spectra for the Vowel Sounds |a| and |i| .

2.4 VOICE IDENTIFICATION EXPERIMENTS

The equipment described above has been utilized along with general purpose digital computers to process a significant quantity of speech to ascertain the utility of the techniques described in Sections 2.1 and 2.2 for voice identification. In the following subsections the data processing goals and methods are discussed and illustrated, and the major results of experiments performed with noisy, bandlimited speech signals are reported.

2.4.1 Data Processing Program

The speech processing originally planned for this project involved operating on the speech signals with special purpose equipment (primarily spectrum analysis and pitch extraction) to produce digitized data samples for real-time input to a Computer Control Corporation DDP-24 computer. However, a major change in the manner of execution of this project resulted from a delay in delivery of this computer. The change involved the addition of a series of data recording and format conversion steps, reprogramming, and the use of different computers. The candidate speech clues, and the learning and recognition processing steps have been modified to some extent to accommodate these changes. However, the results obtained serve to indicate the feasibility of completely automatically extracting clues and processing clue samples to render speaker identification decisions in the presence of noise, and using bandwidth truncated speech signals.

The processing steps actually executed on the project are summarized in Figure 8 in terms of data recording and format conversion operations, and the use of three different digital computers: The Autonetics Recomp II, Digital Equipment Corporation PDP-1, and the IBM 7090. The basic data samples were derived from 30 seconds of speech from each of ten speakers*,

*The original, noise-free speech recordings were supplied by the agency.

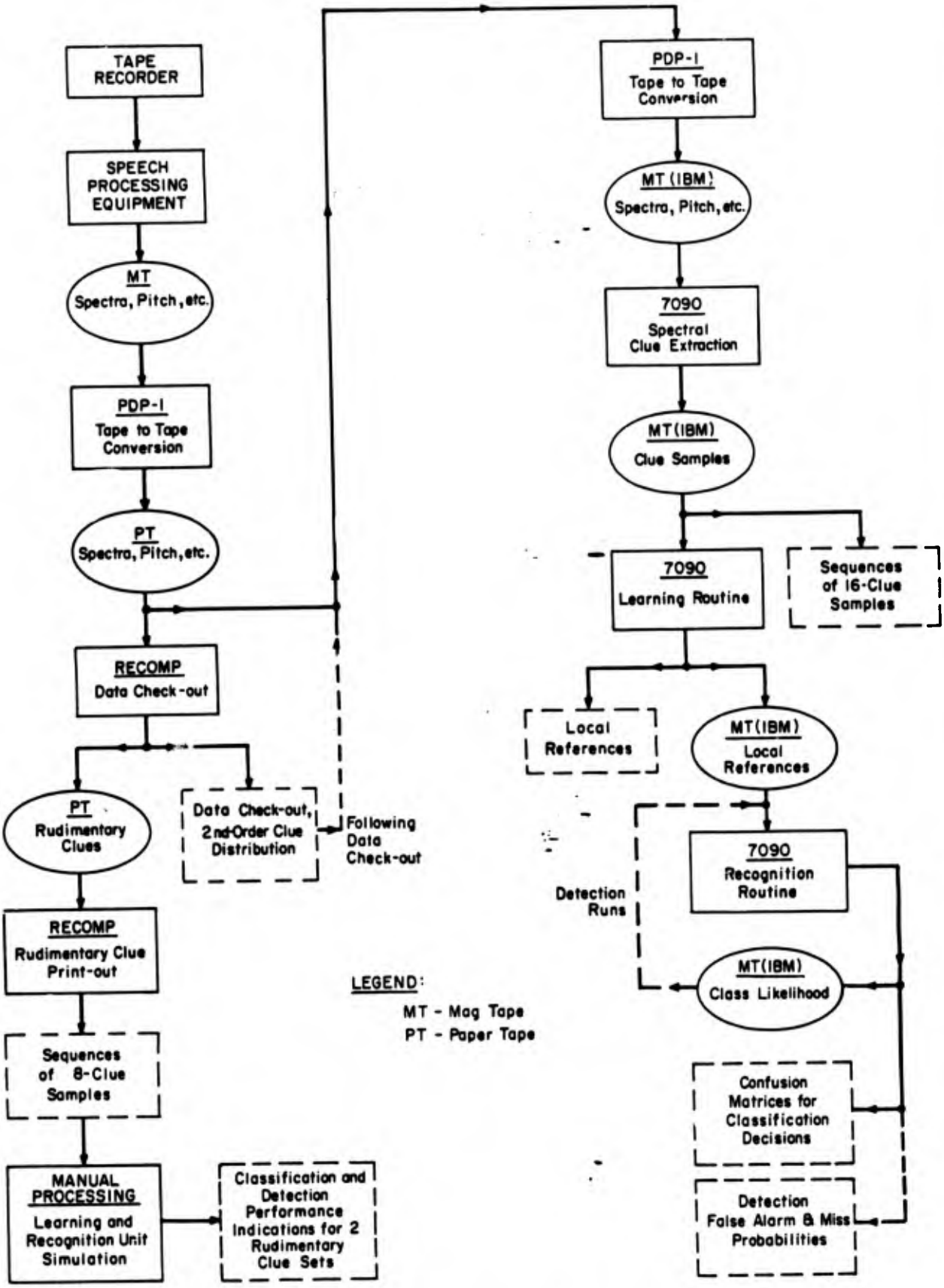


Fig. 8. Data Processing Steps for Demonstrating Automatic Clue Extraction and Speaker Identification Capability

and for each of four (4) signal-to-noise ratios (a total of 20 minutes of speech). Approximately half of the data samples were assigned to the "learn series" to be used in the automatic learning operations, and the remainder were assigned to the "recognition series". Salient characteristics of the ten speakers are listed in Table 2.

Each data sample generated by the Speech Processing Equipment consisted of 48 six-bit characters, as indicated in the diagram. Samples of this type are generated at a rate of 50 samples per second, for a data rate of 14, 400 bits per second. The first major step in processing this data is

Character Position	1 - 40	41	42	43	44	46	47	48
Quantity	Spectrum Analyzer Outputs	Synchro-nization	Voicing	Pitch	Enve-lope	Normal-ized Enve-lope	Enve-lope Deriva-tive	Synchro-nization

Data Sample Format

conversion to a format suitable for input to the IBM 7090 computer. To do this, however, it was required that the data samples to be recorded on magnetic tape in their original format, and that these recordings be processed through the PDP-1 computer to produce an equivalent paper tape recording, which then served as input to an IBM-format conversion program using the same computer. As a result of being constrained to use two different magnetic tape decks, operating at different speeds, it became necessary to check the paper tapes prior to this conversion step. The checking was accomplished with the Recomp II computer.

SPEAKER NUMBER	SEX	AVERAGE PITCH (cps)	AVERAGE DURATION OF VOICED INTERVAL(msec)
1	M	137	201
2	F	212	123
3	M	112	169
4	F	210	123
5	M	119	171
6	M	121	155
7	M	145	130
8	M	139	131
9	F	159	170
10	F	183	170

TABLE 2. Pitch, Sex, and Talking Rate Characteristics of Ten Speakers

Upon completion of this check-out, the data samples were converted to IBM format, and stored on magnetic tape. A total of approximately 30,000 data samples were obtained for four signal-to-noise ratios (40 db, 30 db, 20 db, and 16 db), and ten speakers. All of these samples were generated within voiced intervals of speech.

Speech Clue Extraction

Simultaneous with the paper tape check-out with the Recomp, clue samples were generated by the Recomp II for a set of 8 rudimentary speech clues, to be discussed below.

Following this check-out step as mentioned above, the data samples were recorded in IBM format on magnetic tape which served as input to a speech clue extraction program written for the IBM 7090. We refer to these latter clues as Spectral Clues in order to distinguish them from the Rudimentary Clues produced by the Recomp Computer. The 16 spectral clues generated by this step are as follows:

F_1 = First Formant Position
F_2 = Second Formant Position
F_3 = Third Formant Position
F_1' = First Formant Position Derivative
F_2' = Second Formant Position Derivative
F_3' = Third Formant Position Derivative
A_1 = First Formant Amplitude
A_2 = Second Formant Amplitude
A_3 = Third Formant Amplitude
V = Voicing Indication
F_o = Pitch
F_o' = Pitch Derivative
E = Speech Envelope
E_o = Normalized Speech Envelope
E' = Speech Envelope Derivative
P = Data Sample Power

A clue sample consisting of these 16 component clues is generated in this step for each data sample generated within voiced intervals of speech. As indicated in Figure 8, these clue samples were stored on magnetic tape for use in the 7090 learning routine.

The eight rudimentary clues generated by the Recomp II are:

$\overline{F_o}$ = average pitch

$\overline{F_o'}$ = average pitch derivative

$|\overline{F_o'}|$ = average pitch derivative magnitude

F(max) = maximum pitch

F(min) = minimum pitch

ω = Pitch Max-Min order

$\overline{E'}$ = average speech envelope derivative

S = voiced interval duration

Each of these rudimentary clues is referenced to a single voiced interval. For example, F_o(max) is the maximum value of the speaker' s pitch observed during a single voiced interval. Thus, these clue values are generated at the end of voiced intervals only. As an indication of their rudimentary nature, note that for 6-bit clue quantization approximately 18 bits per second are generated for each rudimentary clue, in comparison with approximately 300 bits per second for each spectral clue.

Learning and Testing Data Processing

The processing of clue values for learning speaker decision boundaries in a clue space, was performed in different ways for the two types of clues. As indicated in Figure 8, the spectral clues were processed in the 7090 using an automatic learning routine known as SPEAR (Statistical Property Estimation Regeneration). This routine is a sophisticated version of ASSC (Adaptive Sample Set Constructor) program utilized earlier on Contract No.

AF-30(602)-2499[1]. This program produces a set of "local references" from a sequence of clue samples. Each local reference consists of a reference clue sample, a "variance" sample $\underset{\sim}{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_c)$ whose components correspond to the clues, and a probability weighting for a single speaker. A set of local references is generated in this program. *

Following this learning step, spectral clue values were processed through a testing routine* in the 7090 to ascertain the accuracy with which the local references for a given clue set can achieve separation of speakers. Since the 7090 routines being utilized were not as efficient as those originally planned for this project, only a few learning and recognition experiments have been conducted with the spectral clues. Specifically, a spectral clue subset consisting of 6 clues has been investigated for four signal-to-noise ratios. Both classification tests and detection tests have been performed using four male speakers with similar pitch (Speakers number 1, 3, 5, and 6 in Table 2). The six clues utilized are F_o, F_o', F_1', F_2, F_3, and E'. These six clues were selected because they represent the three (a priori) most useful formant-related clues, the two available pitch-related clues, and an elementary "talking cadence" clue. Although a larger number of clues could have been processed (up to 16), the primary concern was to determine the utility of formant- and pitch-related clues.

With the change from the DDP-24 computer, the original program for extracting formant information could not be utilized. A replacement program was obtained through the courtesy of Federal Scientific Corp.[5] Unfortunately, this program proved to be less than ideal in two important respects: (a) For the frequency resolution being utilized (75 cps channel spacing, using a 230 cps filter), the criteria for selecting formants from a set of candidate tracks did not work very well, and (b) in processing

*See Section 2.2 and Appendix II for a discussion of this type of recognition routine.

continuous spoken text (as opposed to isolated words) no formant selection is made for very often a majority of the samples occurring within a voiced interval. The net effect of these limitations was that values of the formant-related clues were usually missing in the clue samples. Although the learning and testing routines have been devised to make optimum use of clue values which are available, the processing performed with the spectral clues thus consisted of tests of performance attainable with the pitch-related clues. As a consequence, extensive testing with these clues was not warranted.

To test the notion that relatively easily extracted non-formant-related clues offer significant potential for speaker identification, some of the eight rudimentary clues were processed through the learning and classification steps in a different manner, as indicated at the lower left in Figure 8. Specifically, two different subsets of clues were processed for each of ten speakers, and for the four signal-to-noise ratios. One clue set consisted of the clues $(S, \overline{F_o})$ and the other consisted of $[F_o(\max), F_o(\min), \omega]$. Both of these clue sets were limited by our being constrained to simulate the learning and recognition steps by manual processing. The learning routine was implemented by construction of a histogram estimate of the distribution of clue sample values for each speaker over the corresponding clue space. For the clue set $(S, \overline{F_o})$ a uniform two-dimensional histogram cell structure was utilized with a resolution of 80 msec in S, and approximately 25 cps in $\overline{F_o}$. With this coarse resolution, it was not anticipated that extremely good results could be obtained. However, an increased resolution was not warranted by the amount of available data (approximately 90 clue samples for each speaker and each signal-to-noise ratio). The same pitch resolution (25 cps) was employed for the clues $F_o(\max)$ and $F_o(\min)$, and two 2-dimensional histograms were utilized for the second clue set, since ω is a binary clue.

Since the learning and recognition tests for the rudimentary clues were confined to two-dimensional measurement spaces, the speaker decision boundaries in this space can be illustrated readily. Figure 9 shows the histogram probability density estimates generated for two of the ten speakers tested (a male and a female), using the rudimentary clues (S, $\overline{F_o}$). These estimates were obtained by processing the "learn series" data indicated by the solid points in Figure 10. By comparing the two probability density values in each histogram cell, the decision region boundary shown in Figure 9 is generated.

Both detection and classification tests have been conducted with both of the rudimentary clue sets for all of the ten speakers, three signal-to-noise ratios, and three decision interval durations.

2.4.2 Results of Experiments

The curves in Figure 11 indicate the typical speaker classification performance attainable with a 2-dimensional rudimentary clue set ($\overline{F}_o$, S), as a function of the number of speakers in the group to be classified. The shading in these curves spans the variation in probability of correct identification corresponding to signal-to-noise ratios ranging from 40 db to 16 db. This slight variation indicates that the pitch extractor works quite well over this range, as was expected. In fact, from observations of the pitch extractor output (Appendix IV), it appears likely that with this same rudimentary clue set, essentially the same performance would be attained down to signal-to-noise ratios on the order of 3 db. As indicated in Figure 11, significant improvement in performance results from processing a longer portion of a speech waveform before rendering an identification decision.

While the classification performance indicated in Figure 11 is not high enough for most practical applications, it is nevertheless quite encouraging for the following reasons:

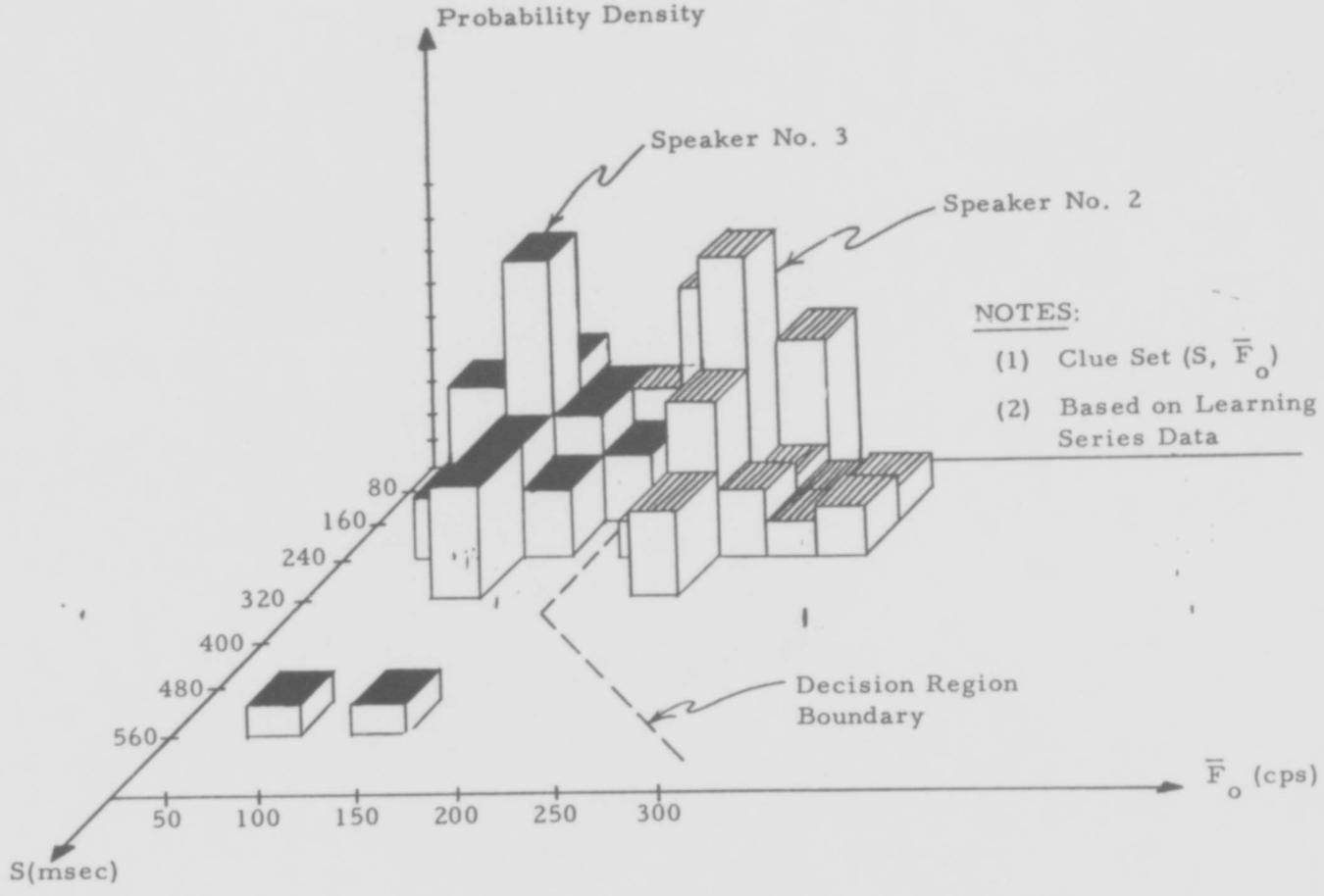


Fig. 9. Histogram Estimates of Two-Dimensional Probability Density Functions for Two Speakers

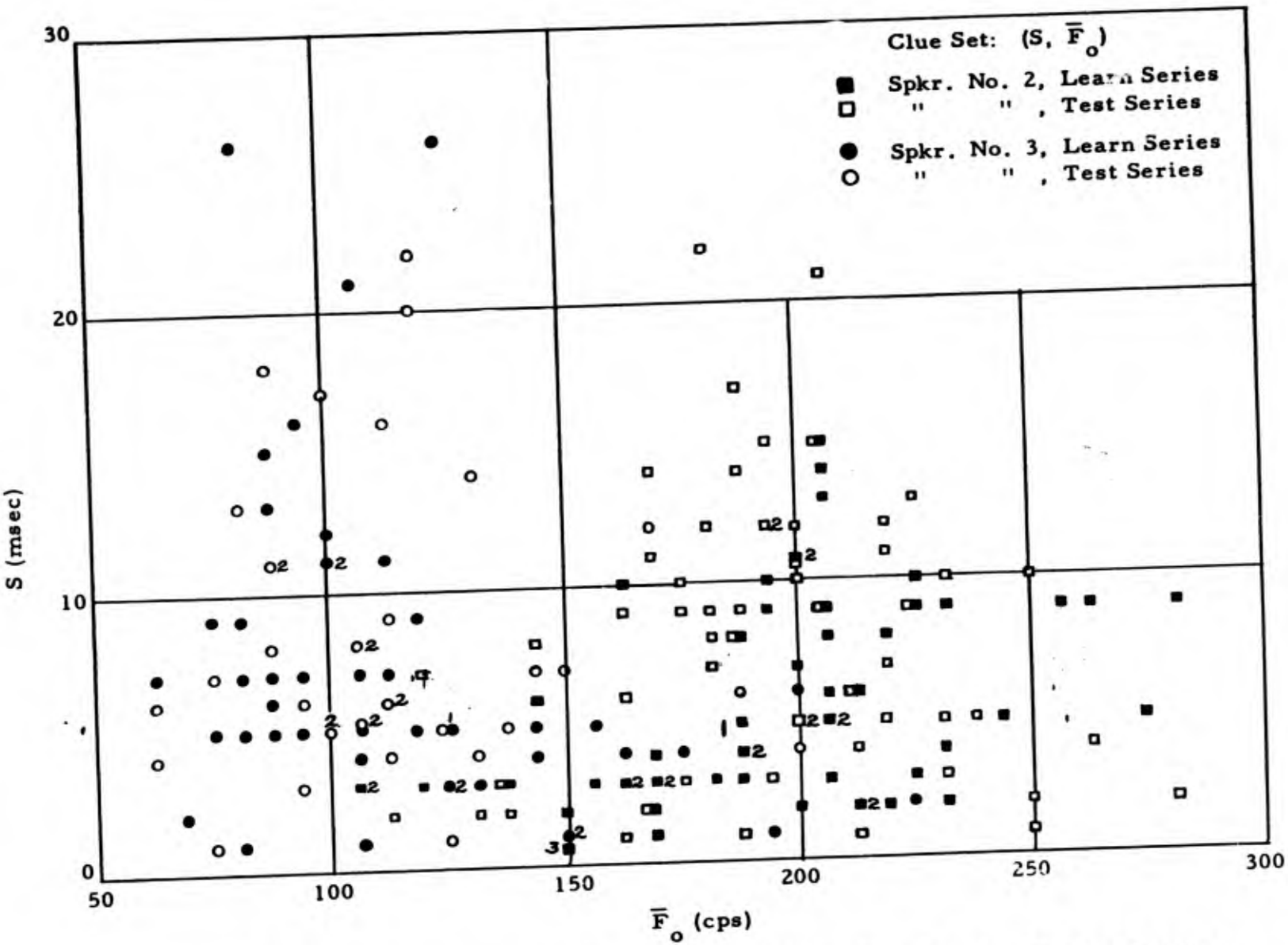


Fig. 10. Two-Dimensional Learning & Testing Clue Samples for Two Speakers

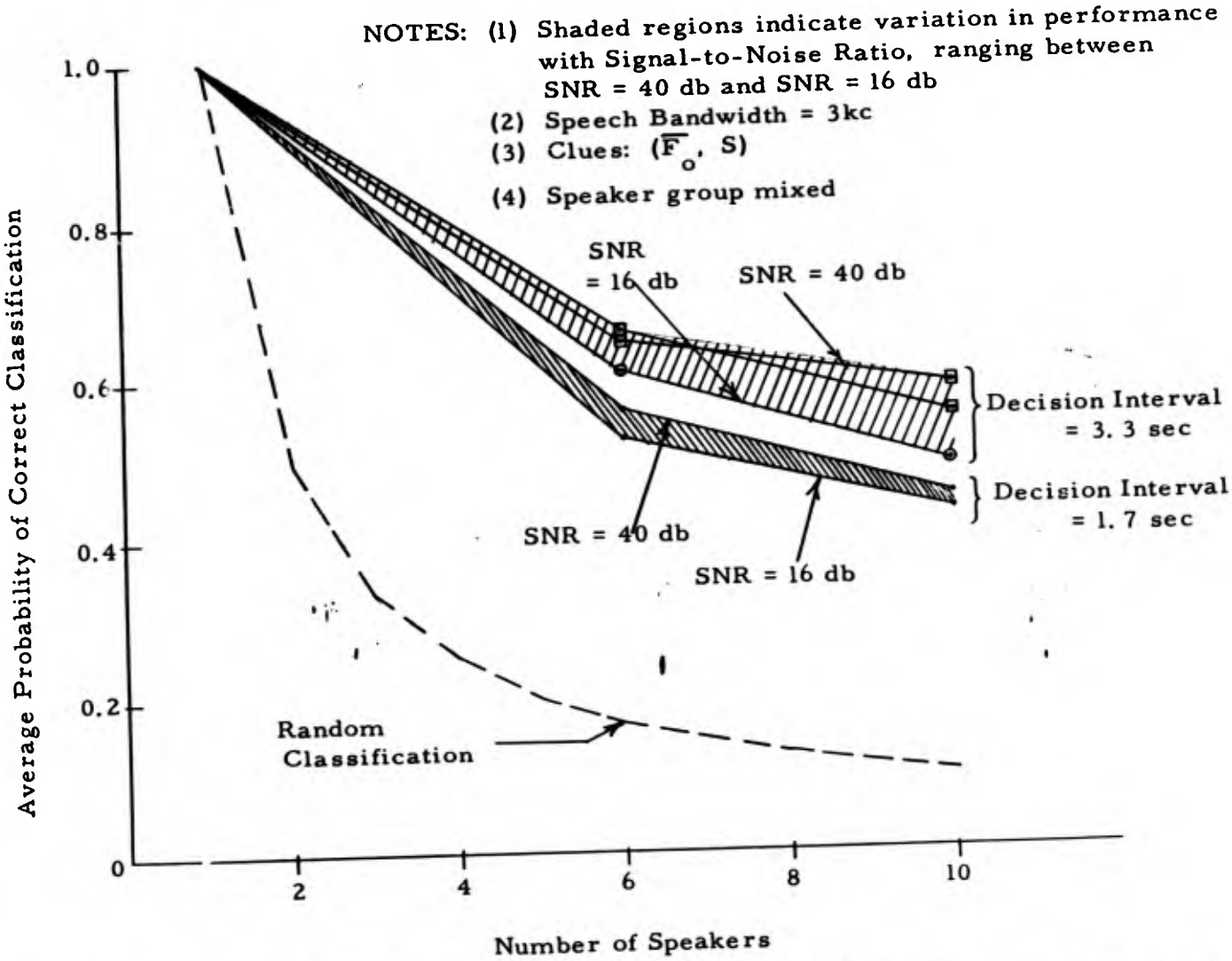


Fig. 11. Classification Performance versus Number of Speakers for Two Decision Intervals and a Two-Dimensional Rudimentary Clue Set

- Only two clues were utilized.
- Each of these clues is a rudimentary measurement.
- Each clue has been coarsely quantized (25 cps resolution for $\overline{F}_o$, 80 msec for S).
- Straightforward histogram estimates of the clue distributions have been utilized.
- The speech bandwidth is truncated and does not include the pitch fundamental.

With the utilization of more clues, the identification performance will improve significantly, as discussed in Section 2.1 (and as illustrated for the detection classification problem in Figure 13). The restriction to essentially 2-dimensional rudimentary clue sets by the data processing program on this project has prevented the direct examination of sets with higher dimensionalities.

The detection performance attainable with the same rudimentary clue set is indicated in Figure 12. Again the variation in performance with signal-to-noise ratio is slight. The average probability of correct decision is only about 75 percent, but for the reasons cited above, this portends excellent performance with the utilization of several rudimentary clues in a more sophisticated recognition routine.

As an indication of the improvement afforded by the use of multiple clues, the 3-dimensional rudimentary clue set [F_o (max), F_o (min), ω] was evaluated also with the coarse histogram estimation recognition method. As shown in Figure 13, just the addition of a single binary clue (ω), the average detection performance is improved from 75% to 80% correct recognition. As mentioned above, a longer decision interval (than the 1.7 seconds for which the curves in Figures 12 and 13 were generated) also produces an improvement.

The effect of a restricted learning interval on classification performance for 10 speakers is illustrated in Figure 14. Under the stated conditions (signal-

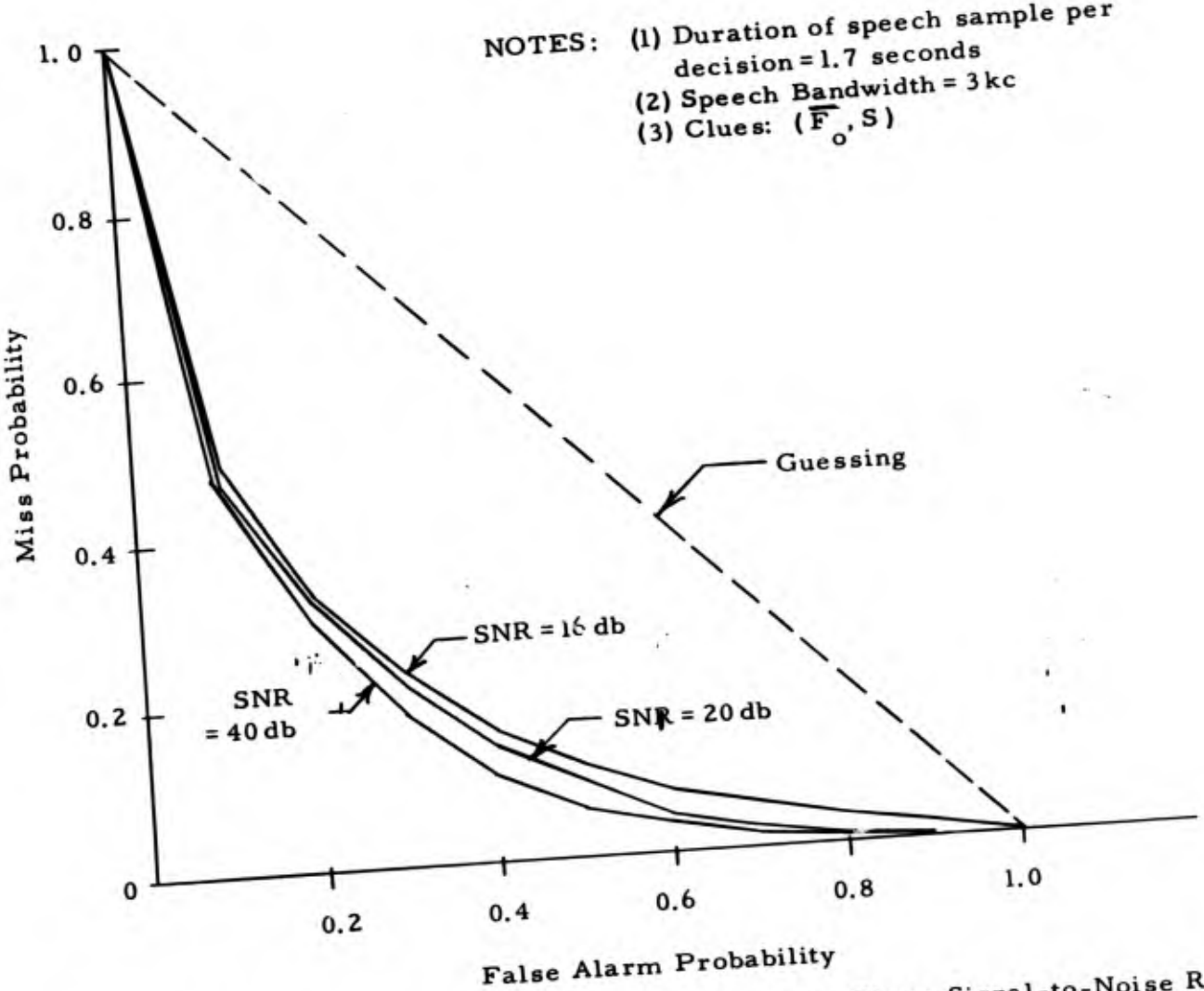


Fig. 12. Average Speaker Detection Performance for Three Signal-to-Noise Ratios, with a Mixed Group of Speakers, using a Two-Dimensional Rudimentary Clue Set

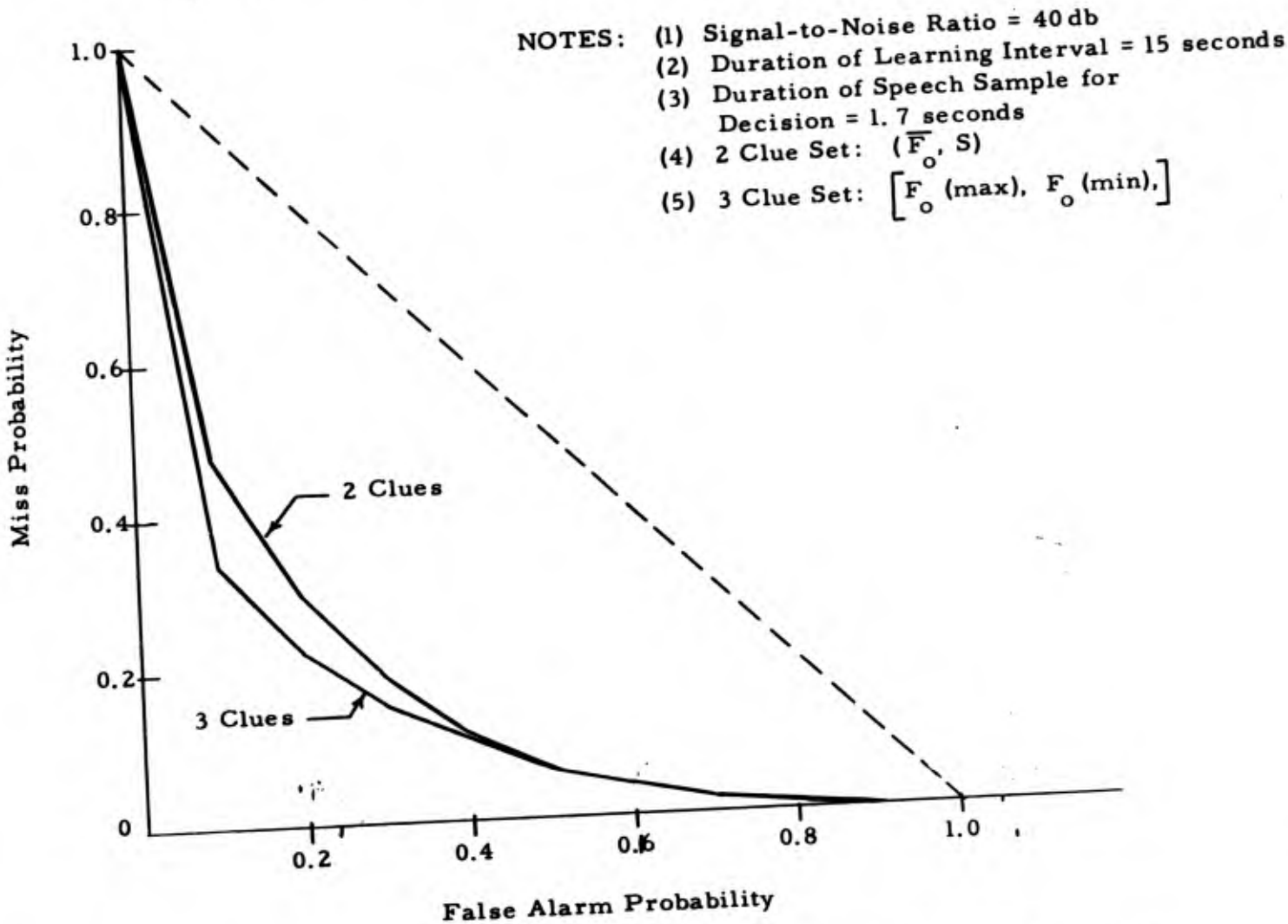


Fig. 13. Average Speaker Detection Performance for a Mixed Group of Speakers, and for Two Rudimentary Clue Sets

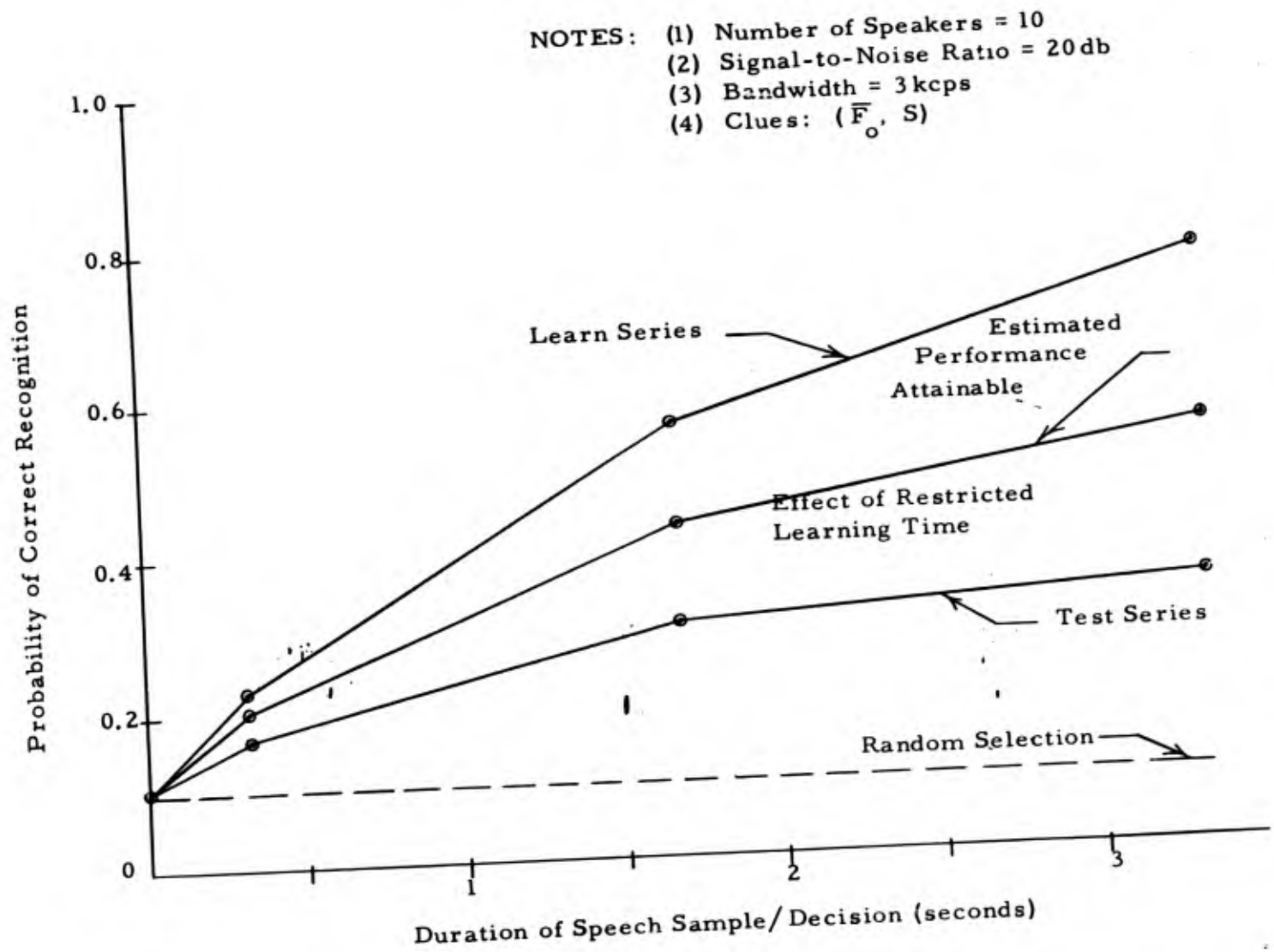


Fig. 14. Estimated Effect of Restricted Learning on Classification Performance for 10 Speakers

to-noise ratio = 20 db, speech bandwidth = 3 Kc) the same 2-dimensional rudimentary clue set was used to classify both the learn series data and the test series data. Since only approximately 45 rudimentary clue samples are obtained from 15 seconds of speech, the histogram estimates on which the decisions are based cannot be regarded as very accurate, even with the coarse cell structure utilized. Thus, for these experiments to the test series data (which are not involved in the clue distribution estimates) can be expected to produce significant variations from the learning estimates. The estimated performance attainable with a longer learning interval (1 minute) has been based on an average of the learn series and test series classification results, as indicated in Figure 14.

The data processing associated with the spectral clues has involved the use of three separate general purpose digital computers, to perform the clue extraction, learning, and recognition operations originally planned for a single computer (see Figure 8). The changes in data format and clue extraction routines required by the late delivery of the latter computer have dictated a less intensive examination of the spectral clue distributions. In particular a subgroup of 4 speakers* have been tested for detection recognition. The typical average correct identification performance obtained with a group of six spectral clues [F_o, F_o', F_1', F_2, F_3, E'] is reflected by the curves in Figure 15 for a signal-to-noise ratio of 30 db, and three decision intervals. For a decision interval of 3 seconds of speech, the average correct recognition probability is 78%. The value of processing long periods of speech prior to rendering a decision is indicated by the significantly poorer performance attainable with only 60 milliseconds of speech. The variation in performance with signal-to-noise ratio is indicated for this clue set in Figure 16, using 1.5 seconds of speech per decision.

The relatively greater susceptibility of the spectral clues to noise indicated

---

*Speaker Numbers 1, 3, 5, and 6 in Table 2.

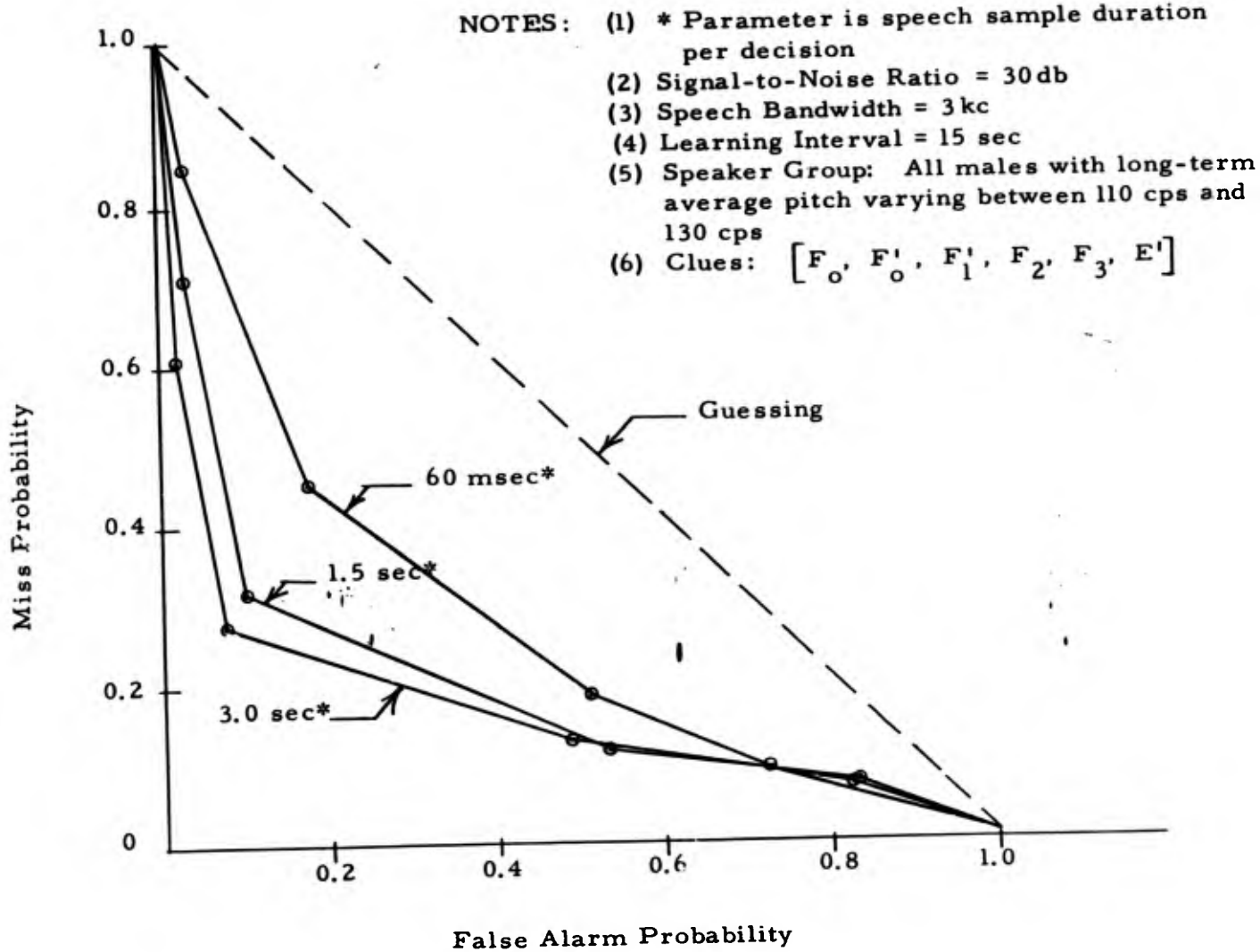


Fig. 15. Detection Performance for All-Male Speakers and Three Decision Intervals, using Six Spectral Clues

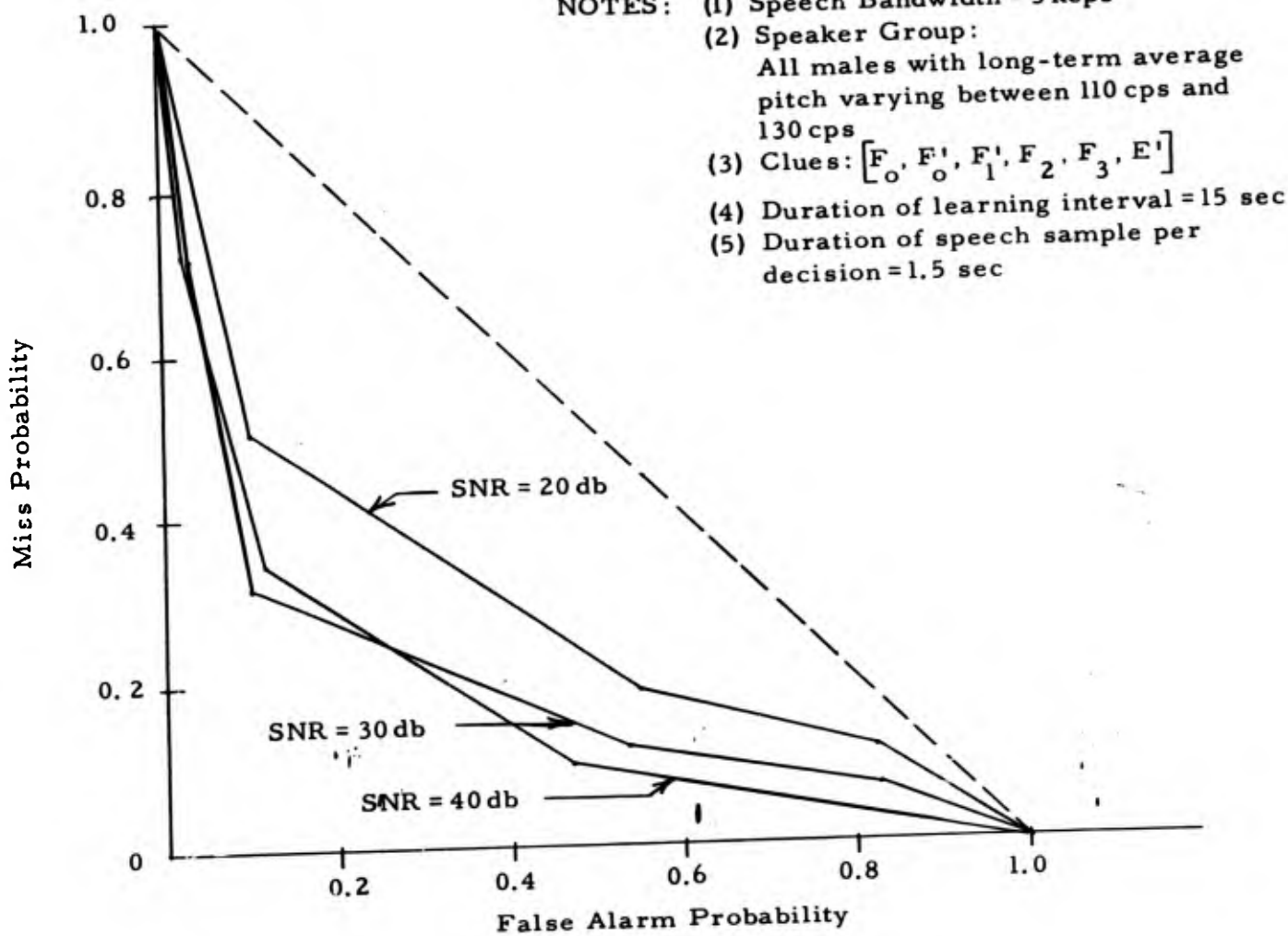


Fig. 16. Detection Performance for All-Male Speakers, and for Four Signal-to-Noise Ratios, using Six Spectral Clues

3. CONCLUSIONS AND RECOMMENDATIONS

The primary goal of demonstrating the automatic extraction of useful speech clues for voice identification in the presence of noise and speech bandwidth truncation has been attained through a combination of special purpose equipment and general purpose digital computers. Two types of speech clues have been investigated: (a) spectral clues, derived from characteristics of the envelope of short-term voiced speech energy density functions, and (b) so-called rudimentary clues, derived from the voicing fundamental and speaking cadence characteristics.

The spectral clues alone had been expected to provide high detection performance at least in the absence of noise, but were found to exhibit two undesirable characteristics: formant frequency location errors are occasionally very large, and formant estimates were generated for only a small portion of each voiced interval of speech. The latter characteristic was an inherent property of the formant extraction routine used on this project, but could be changed.*

However, the former characteristic is a limitation on any automatic formant extraction routine, particularly in the presence of noise and speech bandwidth truncation.

With these limitations on utilization of spectral clues, automatic speaker identification routines exhibited lower performance than had been expected. Specifically, typical test results using a set of six clues indicate that a single speaker can be detected with false alarm and false dismissal error probabilities equal to approximately 25 percent, in the absence of noise, for 3 Kc bandlimited speech, and using a decision interval of approximately 1.5 seconds. At a signal-to-noise ratio of 16 db this figure is degraded to about 33 percent error.

*Use of this particular formant extraction routine was dictated by a change in digital computers during the project.

With a 3 second decision interval the error probabilities for 40 db and 16 db signal-to-noise ratios become 22 percent and 29 percent respectively.

Experiments performed with a few rudimentary speech clues have resulted in encouraging speaker classification performance figures for noisy and bandwidth truncated speech. While these figures are lower than desired for a practical application, the conditions under which the experiments were conducted indicate that much higher performance is readily attainable. Specifically, experiments with rudimentary clues:

(1) Were generally restricted to the use of only 2 clues per experiment
(2) Utilized coarse quantization of clues (25 cps for pitch, 80 msec for talking rate clues)
(3) Involved a restricted amount of learning data
(4) Utilized a crude approximation to the desired recognition procedure.

All of these restrictions were imposed by a change in the data processing facilities utilized on this project.

Typical results of experiments with rudimentary clues indicate that for a decision interval of approximately 3.3 seconds (10 voiced intervals), the probability of correctly identifying one of a group of the speakers (6 males, 4 females) varied between 0.5 and 0.6 within the signal-to-noise ratio range (16 db - 40 db). For a group of six speakers (4 male, 2 female), these figures are 0.60 to 0.66. These probabilities compare with 0.10 and 0.17 for random selection in the 10-speaker group and the 6-speaker group, respectively. The single speaker detection performed was approximately 25 percent error of both types. All of these figures are for a 2-dimensional rudimentary clue set, (average pitch and voiced interval duration).

The net results of the data processing performed with both types of clue sets may be summarized as follows:

(1) Bandwidth truncation to (300 cps – 3300 cps) does not prohibit the extraction of useful speaker identification clues.

(2) Completely automatic extraction of formant locations, particularly in the presence of noise, still poses a problem, particularly for implementation in a relatively simple device.

(3) In lieu of formant locations, spectral peak location and amplitude patterns may be used to represent spectral analysis information.

(4) Other clues not derived from spectrum envelope characteristics exhibit a high potential for performing the discrimination between speakers; these include pitch-derived clues and talking rate or cadence characteristics.

(5) These latter clues are relatively insensitive to short-term fluctuations in the speech signal due to the presence of noise. In particular, it appears that significant degradation in a short-term average pitch extractor (70 msec in the device used on this project) does not occur until the signal-to-noise ratio of a bandlimited speech signal falls well below 10 db.

(6) Use of several of these clues simultaneously can be expected to provide significant improvement in performance over the figures obtained through the use of coarsely quantized pairs of clues.

(7) Significant improvement in speaker identification performance can also be obtained by allowing for longer decision intervals, i.e. for intervals of several seconds.

(8) For these potentially useful clues, apparently a longer learning interval, say a full minute, may be required to realize their full potential.

With the special purpose speech processing equipment utilized on this porject much more extensive speaker identification experiments may be conducted, upon completion of installation of a new digital computer. All of the cumbersome data processing compromises and intermediate storage problems encountered during the current project will be removed. With a capability for processing speech data in real time, a large variety of candidate speech clue sets may be evaluated with a large quantity of speakers and speech materials. The selection of appropriate clue sets for incorporation in devices to be used in special voice identification applications can therefore be accomplished on a sound basis.

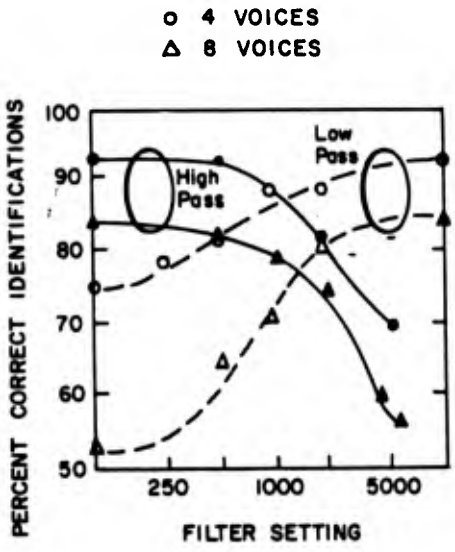
APPENDIX I
VOICE IDENTIFICATION LISTENING TESTS

To ascertain a range of signal-to-noise ratios and frequency bands within which it would be reasonable to expect a machine to perform the task of identifying a voice whose utterances are immersed in noise, a series of small scale listening tests have been conducted to ascertain the human's ability to do the job.

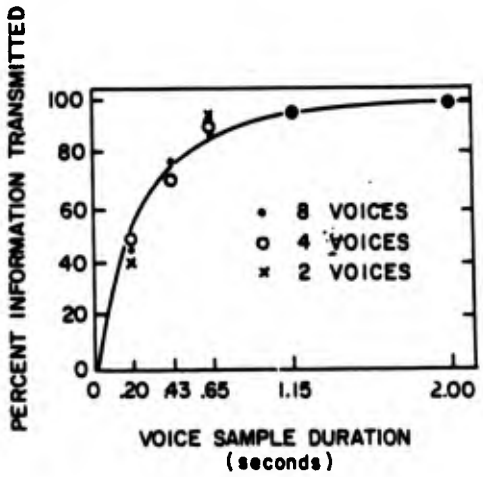
A cursory review of the literature prior to conducting these tests suggested that considerable work remains to be done before the human's ability to identify voices, particularly in a noisy environment, will be known on a quantitative basis. In 1954, some experiments were conducted* which resulted in the performance curves reproduced in our Figure I-1. Part (a) of this figure shows the capability of a panel of seven listeners to identify a speaker, as a function of the degree to which the speech signals have been limited in frequency. Results are shown for both four and eight speakers. Performance is measured by an estimate of the probability of correct classification, i.e. the percent of correct identifications. Since our primary interest is in 3Kc communication channels (300 cps to 3300 cps), we could anticipate from these curves that in the absence of noise, less than 85 percent correct recognition of eight speakers would be achieved by a human, and no more than 90 percent correct recognition of four speakers.

A very important parameter associated with these curves (Figure I-1(a)), is the duration of the utterance on which a decision is based. For these curves, each utterance consisted of a monosyllabic word. Presumably the average duration of these utterances was a fraction of a second. The variation in performance

*Reference 7.



a) Probability of correct identification as a function of speech bandwidth truncation, for 4 and 8 voices



b) Variation in speaker identification performance with speech signal duration

Fig. I-1. Performance of a Human in a Voice Identification Task Utilizing Monosyllabic Words (from Reference 7)

as a function of speech sample duration is shown* in part (b) of Figure I-1. The percent information transmitted is the performance indicator in this case, and it is clear from this curve that approximately one second of speech must be provided to a human in order to achieve high speaker identification performance.

Also in 1954, another project** was undertaken to ascertain the human's ability to identify a previously heard voice, when the speech signals are subjected to selective frequency filtering, altered sound pressure level, and two types of noise: broadband and "propeller-type aircraft noise". The tests conducted under this project involved a large number of listeners (at least twenty for each test), and were conducted for a wide variety of high and low pass filtering frequencies, as well as several signal-to-noise ratios. The duration of speech signals used to render a decision was approximately 2 seconds (5 syllables).

For our purposes, the major results of this project, for the effects of band-limiting, were the indication of relatively little degradation in identification performance for high-pass filtering with cut-off below 150 cps and for low-pass filtering with cut-off above 2000 cps. The reported effects of broadband noise on speaker identification performance are shown in Figure I-2. The 2, 3, and 4- speaker results were obtained during an early phase of the project (as reported in Figure 5 of Reference 8), and the 5- speaker results were obtained during a later phase (as indicated on Page 10 of Reference 9). The salient aspects of these curves are (1) the poor performance (approximately 70 percent correct classification even in the absence of noise) for 3 and 4 speakers, and (2) the indication of almost pure guesswork on the part of a human attempting to recognize which of five speakers has spoken.

The wide variation between the 5-speaker result and that for 3 and 4 speakers, may possibly be explained by the fact that the 5-speaker tests were conducted very carefully, with the sound level adjusted for each speaker using matched earphones

* From Reference 7.
** From Reference 8.

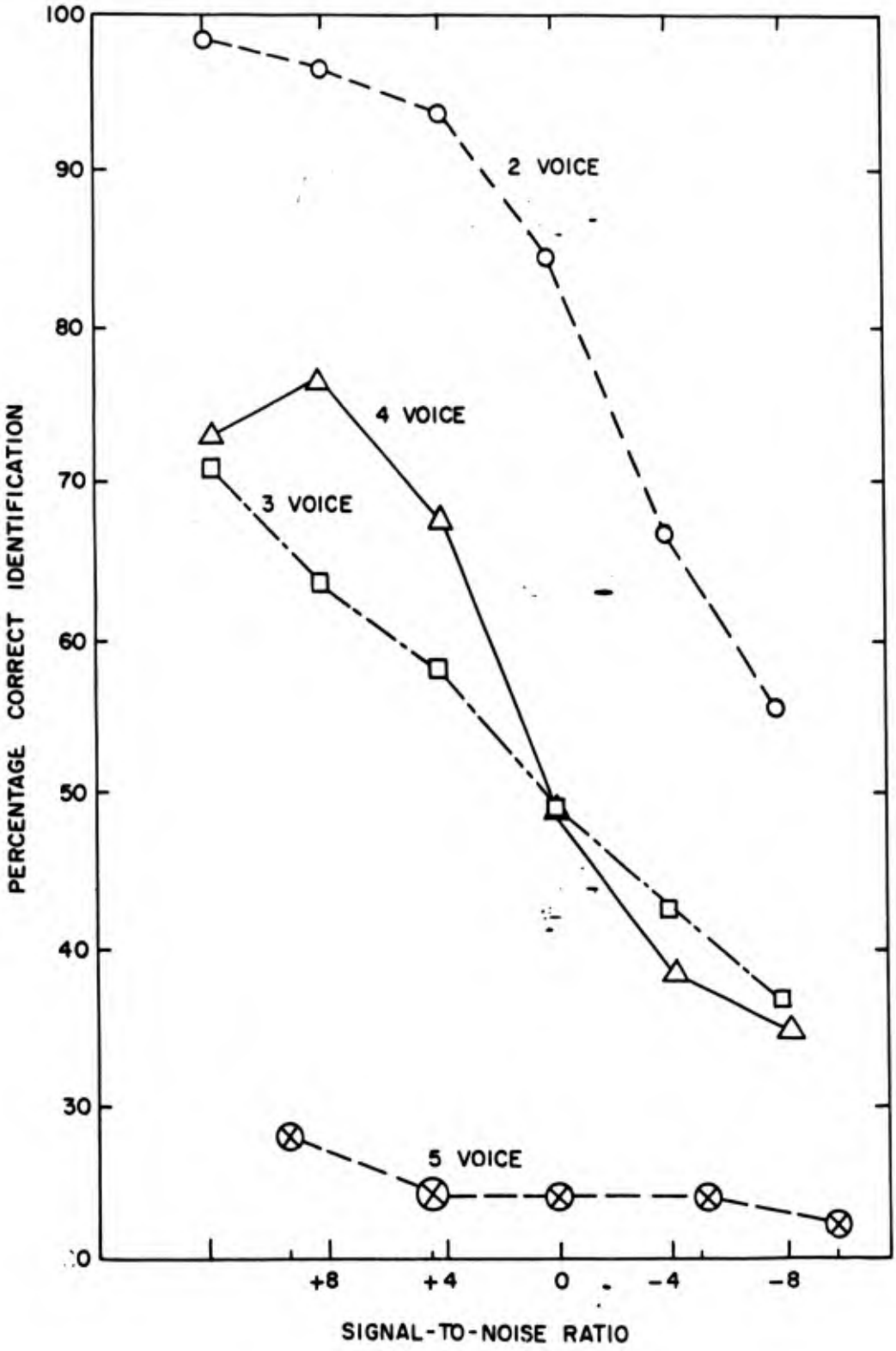


Fig. I-2. Speaker Identification Performance of a Panel of Human Listeners Using Five-Syllable Speech Samples

for presentation of speech to the listeners, whereas the 3 and 4-speaker tests utilized a single loudspeaker presentation to the listeners, with sound level measured only at selected points in a room.

The capability of a human to recognize previously heard voices is not indicated by the results of these two projects to be extremely high. In fact, it would appear that a human can improve on guessing only slightly for signal-to-noise ratios less than 0 db, and for speaker compliments greater than 5, particularly if the speech sample on which a decision is based has a duration on the order of a second.

Although the previous tests were conducted carefully and reported in detail, it appears that, particularly for the tests conducted in the presence of noise*, some of the errors may have been introduced by mislabeling a speaker during the tests due to lack of familiarity with the group of speakers involved. To check this possibility, a series of six tests has been conducted to determine a human's ability to recognize which of 5 speakers (each of whose voices was very familiar to each listener) has uttered a monosyllabic word. Each speech sample in 5 of these tests was bandlimited to the range 300-3300 cps, and was corrupted by combining the original, noise-free utterances with noise. Five of the six tests were conducted for the five signal-to-noise ratios** +40 db, + 20 db, +12 db, +8 db, and +4 db, and the 6-th test was for infinitely clipped (unfiltered) speech. The listening panel was composed of five listeners.

The six tests were conducted as follows. Each of the 5 speakers (numbered one through five) identified himself by uttering the following two sentences:

> "This is speaker number (correct number) speaking.
> You are very familiar with my speech".

Following these familiarization utterances, one hundred words were

---

* References 8 and 9.

** Signal-to-noise ratio is defined as $10 \log \frac{S}{N}$ where S is the peak power of the bandlimited speech signal indicated by a VU meter, and N is the noise power within the band.

uttered (20 words by each speaker), with approximately three seconds of silence between words. Words from the different speakers were presented in random order. During each three second interval of silence following utterance of a word, the 5 listeners (each of whom was very familiar with each speaker's voice), wrote on a Listening Test Sheet the number of the speaker who he believed uttered the word. A total of five hundred different words were utilized in the six tests.

Results of these tests are summarized in Figure I-3. The percentage correct decisions rendered by the five listeners considered individually are shown in the light curves. The heavy curve shows the average percentage correct decisions for the listening panel. Except for Listener Number 5, all curves tend to reflect a monotonic degradation in speaker identification capability as the signal-to-noise ratio (for broadband noise) is reduced. Also, the effects of clipping on a human's ability to recognize a speaker appears to be equivalent roughly to a broadband noise signal-to-noise ratio of 0 db. The actual recognition performance exhibited by the listening panel indicates that even good listeners cannot achieve better than 90 percent correct decisions at signal-to-noise ratios below about + 10 db.

Of major interest to this project is the change in recognition performance which would occur for larger groups of speakers. While a variety of (all indefensible) models can be postulated to provide an extrapolation to larger speaker groups, it is perhaps more realistic to resist the temptation to use these models and simply be content with the knowledge that recognition performance decreases as the number of speakers increases.

Comparison of the results obtained from these six tests with those obtained on the previously described projects suggests that familiarity with the speakers' voices does help a human listener. However, it is doubtful that better than 90 percent correct recognition of one set of thirty familiar speakers' voices could

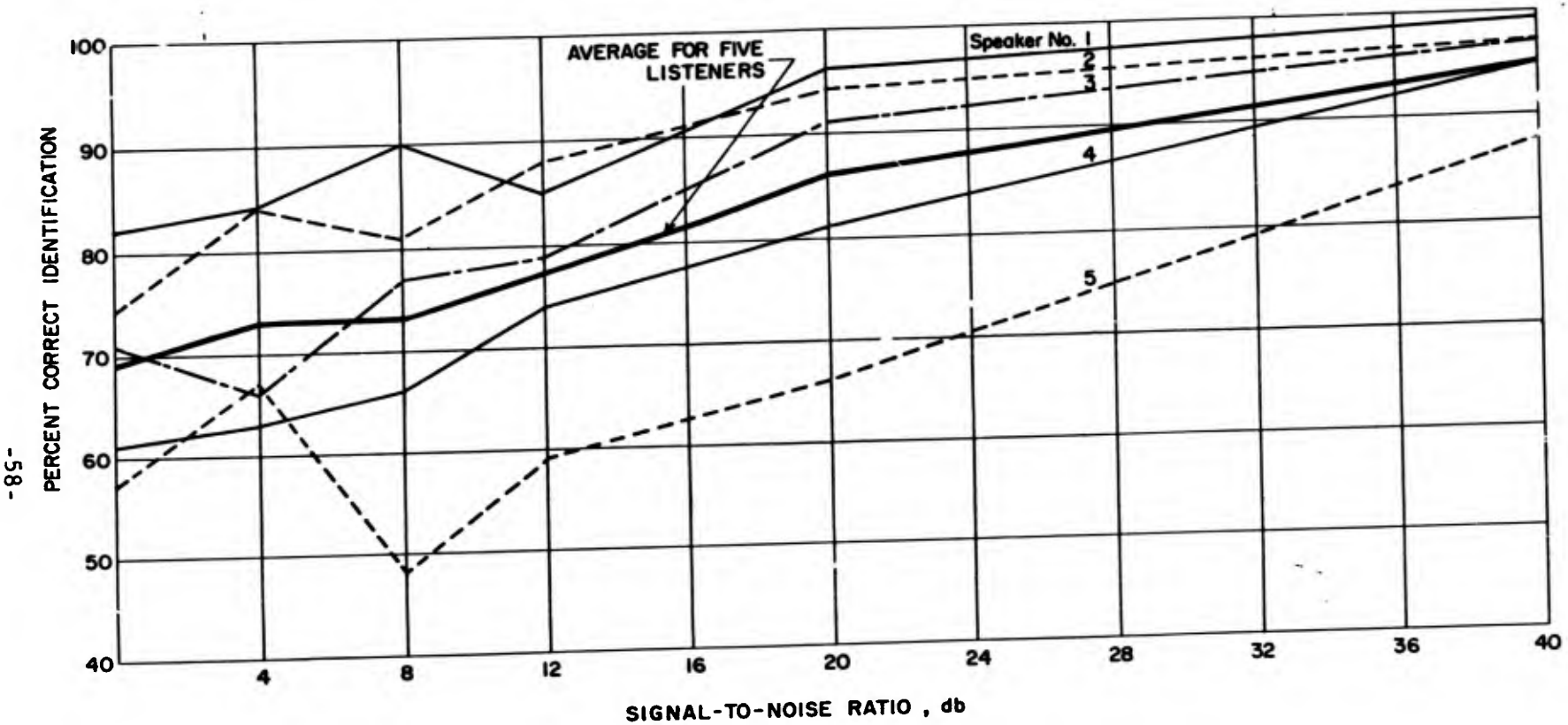


Fig. I-3. Speaker Identification Performance of a Panel of Five Listeners and a Group of Five Speakers, Using Monosyllabic Words as Speech Samples

be achieved by even a "good" listener when the bandlimited speech sample duration is limited to approximately one second and is corrupted to the extent of a 20 db signal-to-noise ratio. For the automatic speaker recognition tests, therefore, it was deemed reasonable to conduct speaker recognition tests for signal-to-noise ratios greater than +10 db.

APPENDIX II

CLASSIFICATION BY LIKELIHOOD FUNCTION ESTIMATION

Consider the problem of deciding which of M classes has given rise to an observed event, $\underset{\sim}{x} = (x_1, x_2, \ldots)$, and suppose that the statistics of events and classes are known, i.e., the joint probability density function of $\underset{\sim}{x}$ and m is known, where m denotes the class label: m = 1, 2, ..., M. The decision-theoretical optimum method of processing a measured event $\underset{\sim}{x}$ to render the classification is well known. Specifically, $\underset{\sim}{x}$ should be regarded as a member of the k-th class if the cost of deciding in favor of the k-th class is less than that of deciding in favor of any of the other classes. This is stated in Eq. (2.1).

$$\sum_{m=1}^{M} P_m p_m(\underset{\sim}{x}) \left[C_k^{(m)} - C_j^{(m)} \right] \leq 0 \text{ for all } j \neq k, \; j=1, 2, \ldots, M, \qquad (2.1)$$

where

$C_j^{(m)}$ ≡ the cost (i.e. loss) associated with deciding that $\underset{\sim}{x}$ belongs to the j-th class when in fact $\underset{\sim}{x}$ belongs to the m-th class,

P_m ≡ the a priori probability that an event from class m will occur,

and $p_m(\underset{\sim}{x})$ ≡ the conditional probability density function of $\underset{\sim}{x}$, given that $\underset{\sim}{x}$ belongs to the m-th class.

This method of decision-making minimizes the average risk associated with the classifications.* If, as is appropriate with many practical classification

*Evidently the basic form of the procedure is the same for other optimization criteria.

problems, the cost or loss is the same for all misclassifications, then Eq. (2.1) reduces to the following decision rule: decide $\underset{\sim}{x}$ is a member of the k-th class if

$$P_k p_k(\underset{\sim}{x}) \geq P_j p_j(\underset{\sim}{x}) \text{ for all } j \neq k,\ j = 1, 2, \ldots, M \tag{2.2}$$

Further, if the a priori probabilities are the same for all classes ($P_m = 1/M$ for all m), then Eq. (2.2) becomes: decide $\underset{\sim}{x}$ is a member of the k-th class if

$$L_{\underset{\sim}{x}}(k) \geq L_{\underset{\sim}{x}}(j) \text{ for all } j \neq k,\ j = 1, 2, \ldots, M, \tag{2.3}$$

where $L_{\underset{\sim}{x}}(m) = p_m(\underset{\sim}{x})$ is commonly called the likelihood function of m given the event $\underset{\sim}{x}$. When class a priori probabilities are the same, the likelihood function is equal to the a posteriori probability of class occurrence; i.e., $L_{\underset{\sim}{x}}(m) = p_m(\underset{\sim}{x}) = p_{\underset{\sim}{x}}(m)$.

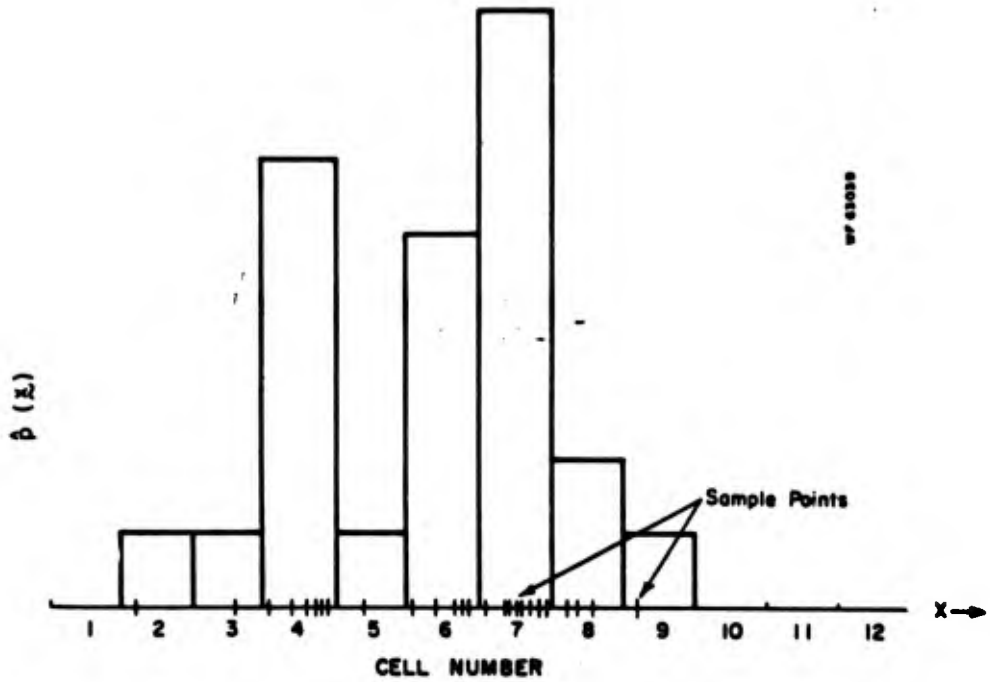
Thus, we see that if the statistics of events and classes are known, than an optimum (from the standpoint of minimizing risk) method of establishing classification decision boundaries in observation space is known, and the only hurdle which remains is implementation of this procedure. Unfortunately, however, this result can only be used as a guide to solving any practical classification problem, because the statistics of events and classes are usually not known precisely.

In most practical problems, all the information available on the statistics of events is contained in the values of a finite number, N, of the sample events processed in the learning mode of operation of a recognition system. A reasonable way to proceed in this situation is to generate an estimate of the likelihood function (or equivalently, the probability density function) of the different classes over the observation space, and render classification decisions in the manner

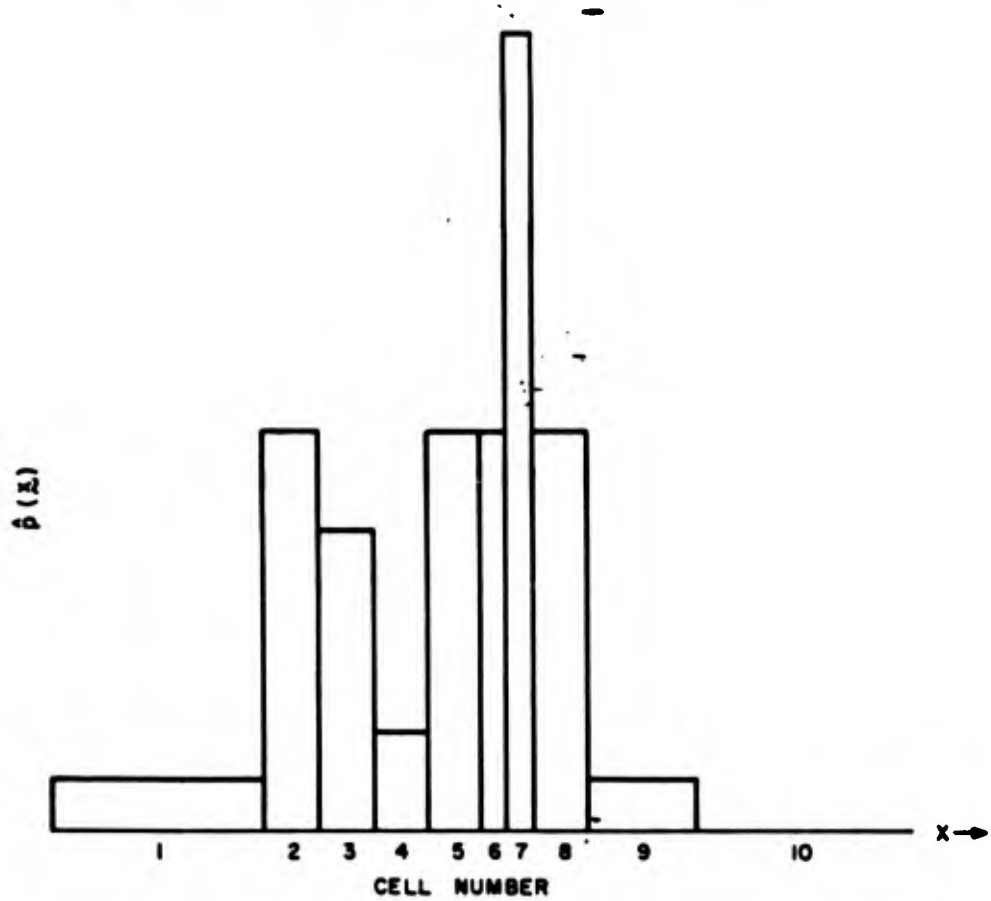
dictated by decision theory using the estimated quantity in lieu of the "true" function. This is the basis for most of the classification methods which have been investigated in the Information Sciences Laboratory.

With this approach to establishing classification decision boundaries in observation space, the method of estimating probability density functions plays the key role. The degree to which the estimate corresponds to the true function determines the similarity between the decision boundaries actually utilized and those which would minimize the misclassification probability. In addition, and perhaps equally important for advancing the development of automatic recognition systems, the form of the estimate should be selected to minimize the equipment complexity (primarily the storage requirements and operating speeds) associated with its implementation.

Although there are many methods of estimating probability density functions, two approaches to the problem stand out as most suitable for consideration in a recognition system. The first consists of estimation through histogram construction by counting the number of occurrences of events in pre-specified regions (cells) in the observation space. Such an estimate is illustrated in Figure II-1(a) for a one-dimensional observation space, and N = 20 samples in the learning set of data. The area of each vertical bar is an estimate of the probability that $\underset{\sim}{x}$ will occur within the range of values defined by the boundaries of the (in this case, one dimensional) cell. This probability estimate for any cell is provided by the ratio of the number of learning samples which fall in the cell, to the total number of learning samples. In general, the probability density function $p(\underset{\sim}{x})$, of a multidimensional random variable, $\underset{\sim}{x}$, is assumed to be constant over the cell, and equal to the ratio of the estimated probability of obtaining a sample within the cell to the hypervolume of the cell.



a) Uniform Cell Structure



b) Adapted Cell Structure

Fig. II-1. Histogram Estimates of the Probability Density Function of a One-Dimensional Random Variable

In symbols,

$$\hat{p}(\underset{\sim}{x}) = \frac{N_j}{N} \cdot \frac{1}{V_j} \tag{2.4}$$

where

$N_j \equiv$ the number of learning samples which occur within the j-th cell,

and $V_j \equiv$ the volume of the j-th cell.

The caret symbol indicates that an estimate, rather than a true probability density function is obtained.

Straightforward application of this method of estimation requires a priori specification of the cell structure (size, shape and number in observation space) over which the histogram is to be constructed. To reduce storage requirements it would be desirable to keep the number of cells small. However, to represent the probability density function accurately in regions where this function is sensitive to small changes in $\underset{\sim}{x}$, the cells should be small, which would make the number of cells large. A third factor which must be considered in selecting a cell structure is that the accuracy of estimation of the probability that $\underset{\sim}{x}$ will occur in a given cell, is proportional to the number of learning samples which occur within the cell. Thus, the minimum resolution which should be attempted with a cell structure is limited not only by the (unknown) character of the true probability density function, but also by size of the learning set.

Since the character of the (unknown) probability density function plays such an important role in determining the appropriateness of a given cell structure, it is reasonable to utilize the only information available on this function (the learning set) to select the cell structure. This could be accomplished during the learning mode of operation of a recognition system by adjusting the cell

structure (according to a pre-established criterion) with each exposure of the system to a new learning sample. There are many ways of implementing this procedure. One would be to start with a coarse cell structure consisting of a few rectangular polytopes (e.g. hypercubes), and then increase the cell structure resolution by subdividing existing cells to avoid a violation of the constraint that no more than 4 (say) samples should be allowed in a single cell. This method of adjusting cell structure is illustrated in Figure II-1(b) for the same 25 one-dimensional samples used to construct the (uniform cell structure) histogram in Figure II-1(a). Even though the modified cell structure involves two less cells than the uniform structure, considerable greater resolution is attained with the modified structure. If the range of possible values of the observation variable (x) is partitioned into segments corresponding to unchanging values of the probability density function, then both the uniform cell structure and the adapted structure would require 9 quantities to represent their corresponding histograms, although the adapted structure attains a higher resolution.

Of course, the accuracy of an estimation based on rules for adapting the histogram cell structure to the learning samples must be evaluated before the utility of such a procedure can be ascertained. The purpose of this histogram illustration is to point out the possibility of using an adaptive procedure for estimating probability density functions (and therefore, decision boundaries). The significant difference between this approach to adaptation of decision boundaries in observation space and most of the other methods which have been proposed in the past few years is that this approach makes a conscious attempt to approximate the class probability densities without any prior assumptions about the distribution of events in the observation space. Having estimated the densities, the procedure known to be "optimum" is used to construct the decision boundaries. While constraints on the number and type of boundaries

which can thus be generated do exist with this approach, these constraints impose no serious limitations on the distribution of events in observation space for a successful separation of classes.

The second important approach to estimation of probability density functions, called local reference representation, takes the adaptation procedure oulined above for histogram estimates as a point of departure, but implements this approach in a somewhat different manner. As before, the observation space is partitioned adaptively into regions called cells; however, the role of the estimation process and the geometrical disposition of these cells are not necessarily the same as for histogram construction. First of all, cells are created only in those portions of the observation space where learning samples have been observed. Since it is expected that in most practical problems a very high percentage of the volume of the observation space is empty, this serves to reduce significantly the storage requirements. Secondly, the size, shape and height of a cell is determined from an examination of the local behavior of the learning samples in the neighborhood of the cell in question. From the local behavior of the learning samples a component function is generated which represents the learning samples in the immediate neighborhood of the cell.

The entire process of local reference representation can be regarded as an adaptive method of approximating the probability density by expanding it in a set of non- a priori specified component functions. The component functions represent and typify each of the different significant manifestations of members of the class by creating a cell corresponding to each of the different manifestations. The component functions also describe the local characteristics of each "typical" concentration of learning samples and they shape the cells.

Figure II-2 illustrates the behavior of these component functions for a one-dimensional random variable. The process by which such an estimate of the probability density function is constructed encompasses three basic steps:

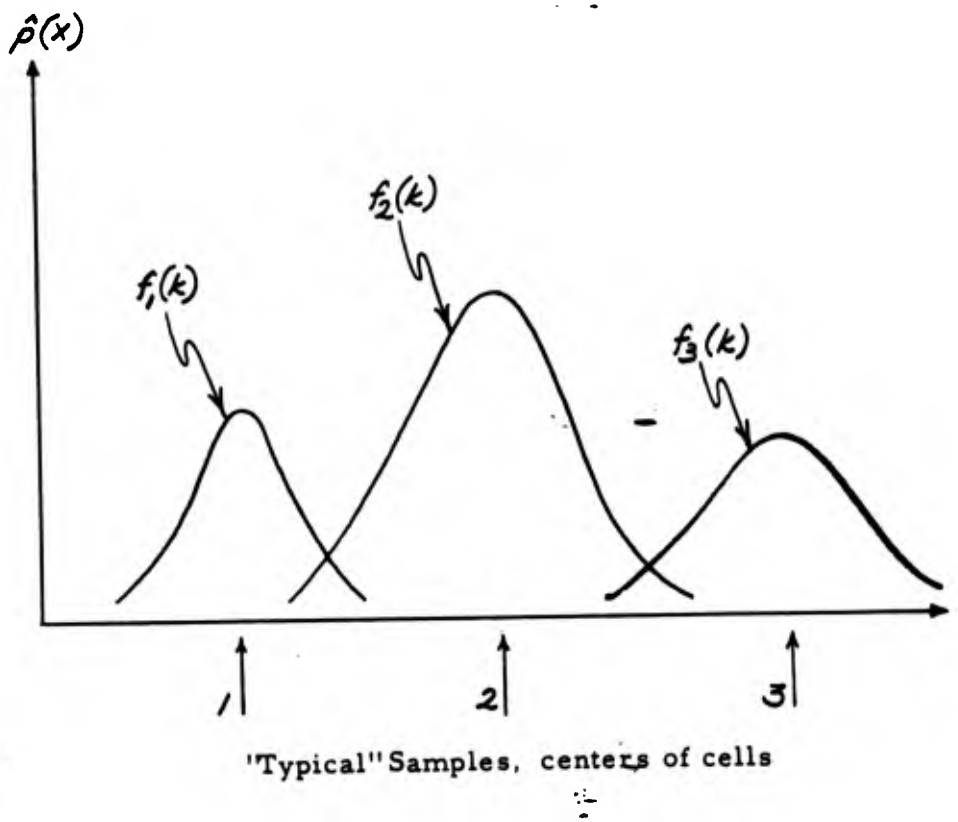


Fig. II-2. Local References of a Probability Density

- A cell structure consisting of c cells is generated by the learning data.
- Corresponding to each cell, one of a class of functions, $\{f(\underset{\sim}{x})\}$, is selected according to values of learning data samples occuring within that cell.
- The probability density function is estimated by some sort of combination of the selected functions $\{f_j(\underset{\sim}{x})\}$, $j = 1, 2, \ldots, c$, corresponding to the c cells.

Thus, the probability density function is represented by a set of "local references", where each local reference consists of a reference point, a component function, and a cell boundary.

The cell structure is established by adaptive adjustments controlled by the sequence of samples contained in the learning set. Of the many ways in which rules for the adaptation can be established, the following has been studied most extensively. The first learning data sample is established as the "reference point" for the first cell. The second sample is then compared with this reference point according to a criterion which indicates whether this sample should be used to modify the first cell (by adjusting its reference point), or be established as a new reference point for a second cell. If used to modify the first cell, then the criterion by which future learning samples will be compared with the reference point for that cell may also be modified. If the second sample is established as the initial reference point for the second cell, then a second criterion is also assigned to that cell. The third and succeeding learning data samples are compared with each of the established cell reference points (according to their respective criteria) and used to either modify one of these cells, or establish a new one.

The criterion by which new learning data samples are compared with an established cell is constrained to reflect a notion of similarity between the cell reference point and a new sample. The provision for adjustment of the criterion according to the value of a new sample, allows for development of different measures of similarity between events in observation space, according to location of the events in that space, as well as class membership. The criterion associated with a given cell may be regarded as a maximum allowable distance between its reference point and any other point in observation space to be associated with that cell, where "distance" is measured in an adjustable way. Thus, modification of the criterion for a cell changes the cell boundary which consists of all points in observation space equi-distant (at a specified value) from the cell reference point.

Either during or at the end of the process of cell formulation, the samples occurring in, say, the ν-th cell are used to select the component function $f_\nu(\underset{\sim}{x})$, from a pre-established class of functions, $\{f(\underset{\sim}{x})\}$. This set of functions may or may not allow for non-zero values of $f_\nu(\underset{\sim}{x})$ outside of the ν-th cell.

In practice, it is convenient to relate the class of component functions to the way in which distances are measured between points in the observation space and cell reference points. In particular, if the component function for a cell decreases monotonically as the distance between $\underset{\sim}{x}$ and the cell reference point increases, then the process of computing probability density function values at $\underset{\sim}{x}$ may be reduced to the calculation of distances between $\underset{\sim}{x}$ and the cell reference points. An illustration of this relationship is the used of quadratic forms for measuring distances, and Gaussian forms for the component functions.

The last step in the process of estimating probability density functions with typical samples consists of combining the component functions over the entire observation space. One way is to consider the probability density function to be the sum of the component functions:

$$p(\underset{\sim}{x}) = \sum_{\nu=1}^{c} f_\nu(\underset{\sim}{x}) \qquad (2.5)$$

For uncorrelated Gaussian component functions, this method allows for convenient processing of recognition data samples in which some parameter values are missing. Another method of combining component functions is to use the function whose cell reference point is closest to the point at which the probability density function is to be estimated.

APPENDIX III

DESIGN OF A REAL TIME, DELAY LINE TIME COMPRESSOR (DELTIC) SPECTRUM ANALYZER

Some of the basic design considerations for a Delay Line Time Compressor (DELTIC) spectrum analyzer are presented in this appendix. As with most spectrum analyzers, the idea behind a DELTIC device is the storage and repetitive input of a finite segment of the signal to be analyzed into a frequency-scanning filter. The resulting output of the scanning filter is a succession of short-term energy density spectra of the input signal. With a DELTIC device, each segment of input signal is time-compressed, and the analyzing filter commensurately broadened, to allow for the completion of many repetitions in a short period of time. Thus, the total time required to generate each energy density spectrum may be made extremely short, and essentially "real-time" analysis is possible.

The major operations involved in the analyzer are depicted in Figure III-1. The input waveform is sampled periodically at a rate, f_s, which is high in comparison with the rate at which energy density spectra are to be generated,

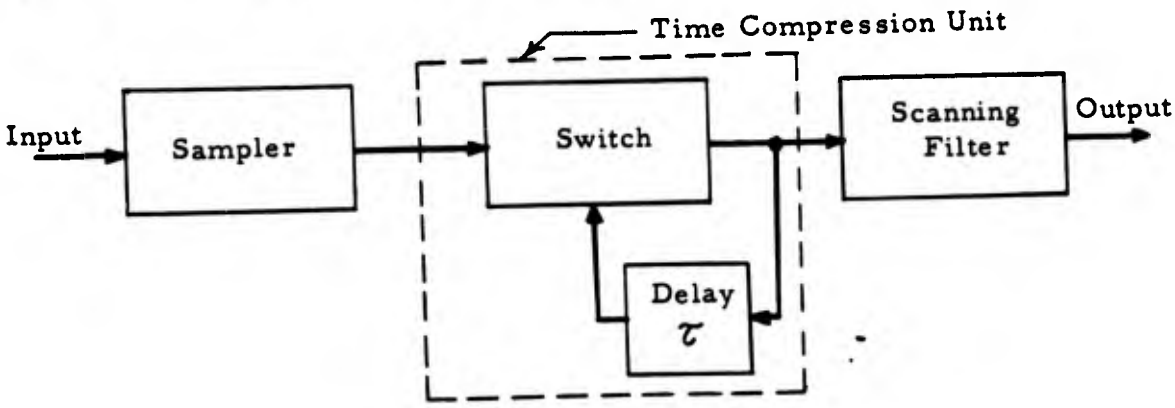


Fig. III-1. A DELTIC Spectrum Analyzer

and the samples are fed to the time-compression unit. The latter unit may be regarded as a delay line of length, $\tau = \frac{1}{f_s} - \epsilon$, which is slightly smaller than the spacing between samples, and a switch which serves to control the number of samples to be employed in the analysis. The time-compressed sequence of samples (or staircase approximation) emitted by the time compression unit are passed to the scanning filter, whose output represents an energy density spectrum.

Although we have examined and established the feasibility of a version of this system in which each repetition of the time compressed waveform is the same, it is possible for the analyzer to be designed to replace the oldest samples with new samples during the time that the scanning filter is sweeping across the frequency band to be analyzed, W. The resulting "skewing" introduced in the energy density spectra will be insignificant, provided spectra are generated at a reasonably high rate.

1. DESCRIPTION OF A DIGITAL DELTIC ANALYZER

For reasons of economy, flexibility and reliability, a digital time compression unit is required. A block diagram of a spectrum analyzer employing such a unit is shown in Figure III-2, and its operation described as follows. The time compression unit is included within the dotted line. The input signal is sampled by the A/D converter at a rate, f_s, samples per second, with N_q bit quantization. The A/D output is stored temporarily in N_q, (I-1)-bit registers, where I is the number of samples taken for each cycle of the delay lines. After I conversions, the samples are shifted into N_q, 1 mc. delay lines. Timing is controlled by a 1 mc. clock and a counter which counts up to N_s, the number of of input signal samples used in the analysis. The delay lines are $N_s - I\mu$ s long so that the data in the delay line precesses by I bits for each recirculation of the delay line. Thus data entering the line during the 1st I counts will pass

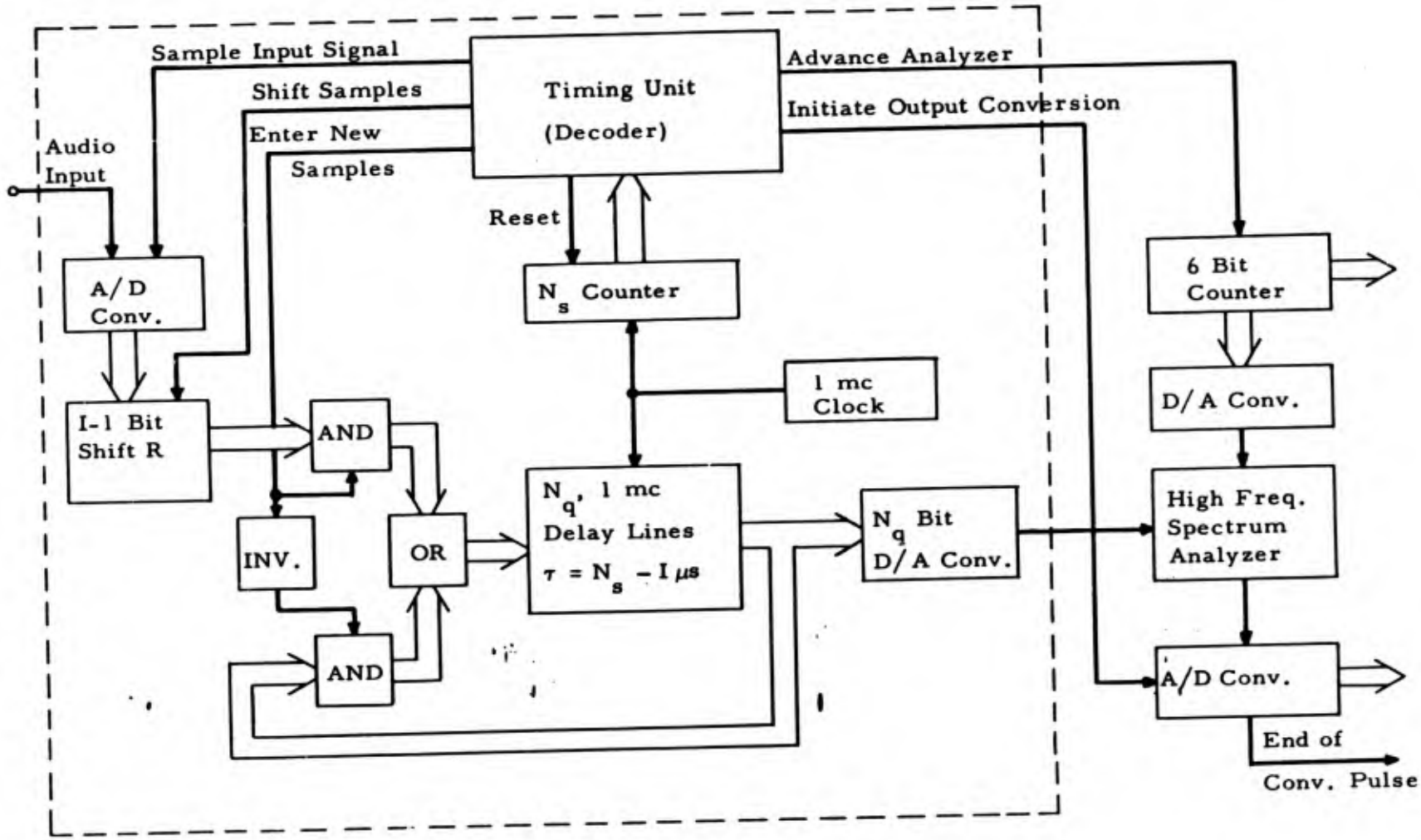


Fig. III-2. Block Diagram of Time Compression Spectrum Analyzer

through the delay lines and have re-entered the lines as the counter resets to zero, and the next I samples will be entered immediately following the first I samples. The output of the delay lines is connected to a D/A converter to produce a speeded up repetitive version of a segment of the input signal. The speeded up signal is then analyzed by a high frequency spectrum analyzer whose output can then be A/D converted for entry to a computer. The 6 bit counter is used to generate the control voltage required to sweep the spectrum analyzer. The counter output therefore is fed to the computer as an indication of the analysis frequency. The timing unit is a set of combinatorial circuits used to decode the output of the counter to generate the necessary timing signals to control the delay lines, A/D converters and spectrum analyzer control counter.

The detailed design of a DELTIC unit involves compromises between the minimum resolution attainable, analysis time required, maximum input signal bandwidth, flexibility in resolution and analyzing filter characteristics, and cost of instrumentation. For the unit developed for use in speech projects in the Information Sciences Laboratory, the trade-off between these factors has been viewed from the standpoint of producing an analyzer which can provide a resolution of 60 cps over at least a 3 kc band. Some of the equations which reflect the different trade-offs involved are derived in the remainder of this appendix.

2. DELTIC DESIGN EQUATIONS

The minimum, real time length, T_d, of the segment of data which must be stored in the delay lines is:

$$T_d = k/F \tag{1}$$

where F is the analyzer resolution (effective filter bandwidth) and k is a factor which depends on the type of analyzing filter used and the accuracy required.

Curves of the step response of N pole Butterworth filters are given in Figure III-3*. Equations for the response of 1- and 3-pole filters to step inputs at frequencies different from the filter center frequencies are derived in Section 3 of this appendix. Curves showing these responses are plotted in Figures III-4 and III-5. As the data is recirculated in the delay line, the first sample of the segment comes out immediately following the last sample. Since the tail end of the data segment may be either in phase (e.g., a sine wave for which an integral number of wave lengths are contained in the delay line) or out of phase (e.g., a sine wave for which an odd number of half wavelengths are contained in the delay line) with the beginning of the segment, the worst case error due to the transient of the "seam" is equal to twice the difference between the filter steady state response and its transient response, as shown in Figures III-3, III-4, and III-5. The value of k in Equation (1) can be determined from these curves. Reasonable values would be about 1 for a single pole filter and 10% accuracy, to about 2 for a three-pole filter and about 4 or 5% accuracy.

The number of samples, N_s, stored in the delay lines is:

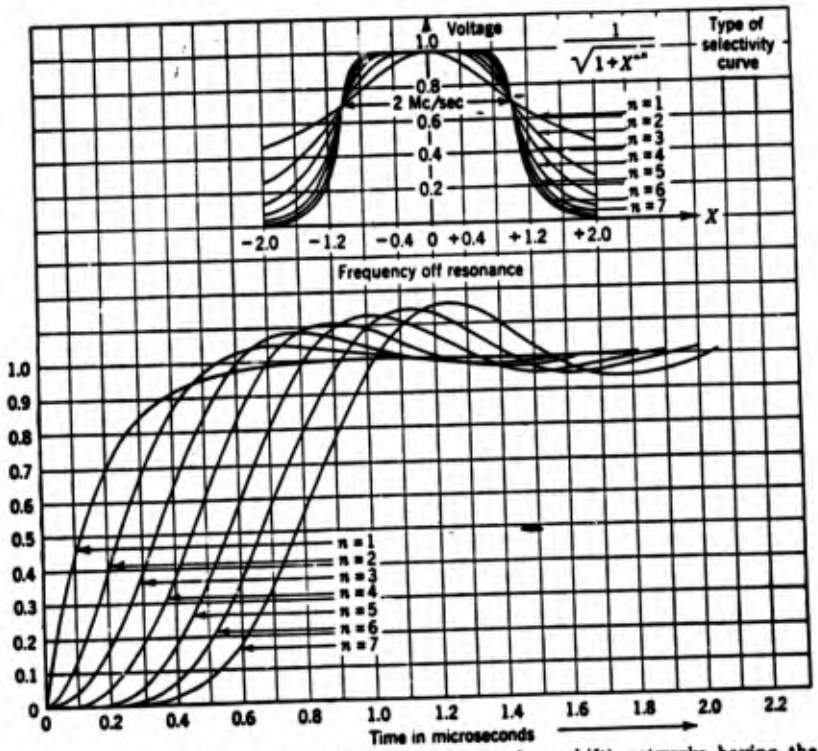
$$N_s = T_d f_s \tag{2}$$

where f_s is the sampling frequency.

The delay line length, τ, is determined by the sampling frequency and the number of samples taken, I, for each recirculation of the delay line:

$$\tau = \frac{I}{f_s} \tag{3}$$

*From Reference 6, p. 283.



—Step-function response of the (minimum phase shift) networks having the absolute value curves shown above.

Figure III-3*

*Reference 6, p. 283.

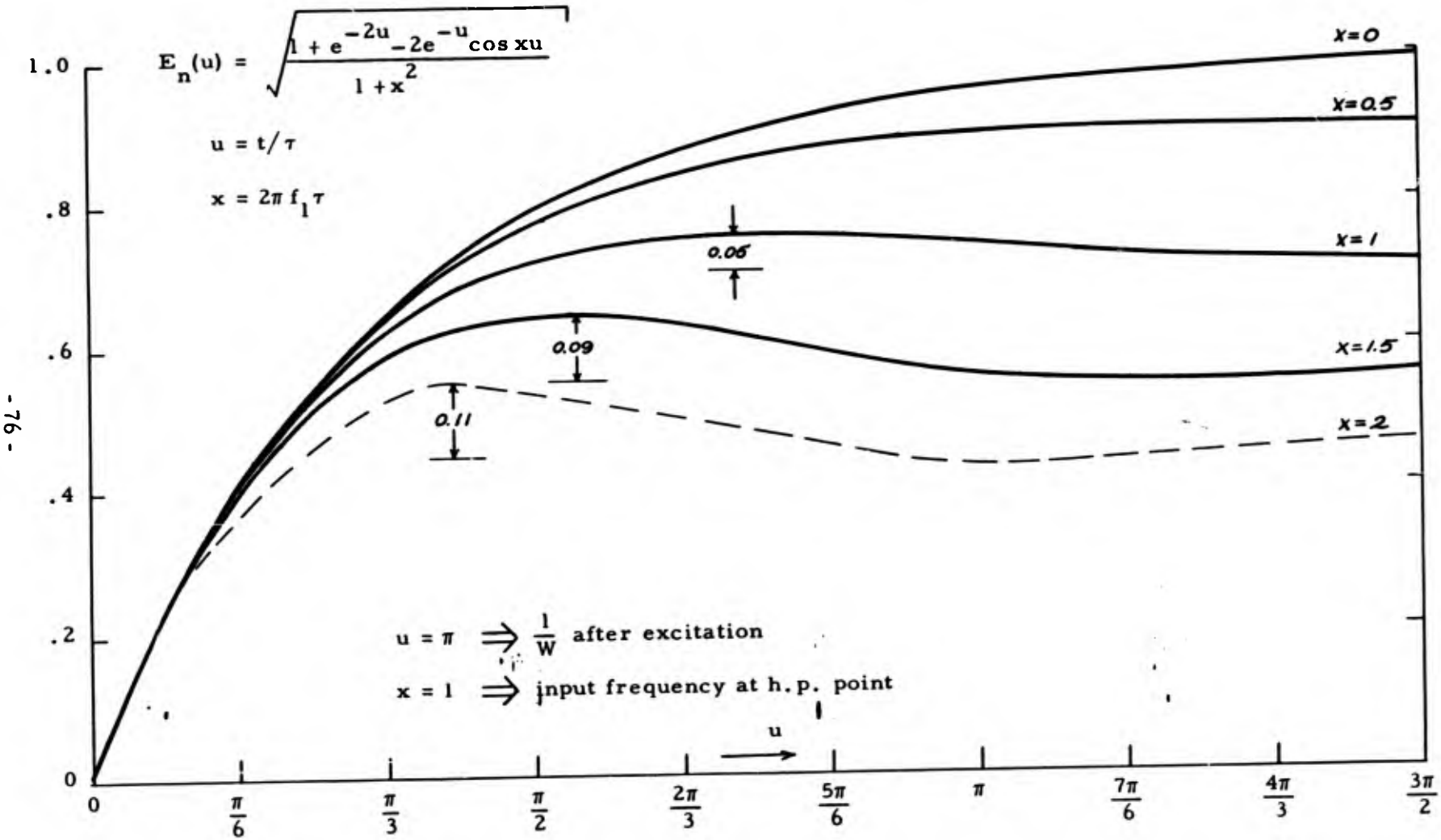


Fig. III-4. Envelope of Sine Wave Step Response of One-Pole Filter

$$E_n(u) = \left| -\frac{e^{-u}}{1+ix} + \frac{e^{ixu}}{(1+ix)(1+ix^2)} - \frac{2}{\sqrt{3}} \frac{e^{-\frac{u}{2}}\left\{\sin\frac{\sqrt{3}u}{2} + ix\sin\left(\frac{\sqrt{3}u}{2} - \frac{\pi}{3}\right)\right\}}{1+ix-x^2} \right|$$

$$\lim_{u\to\infty} E_n(u) = \frac{1}{\sqrt{1+x^6}}$$

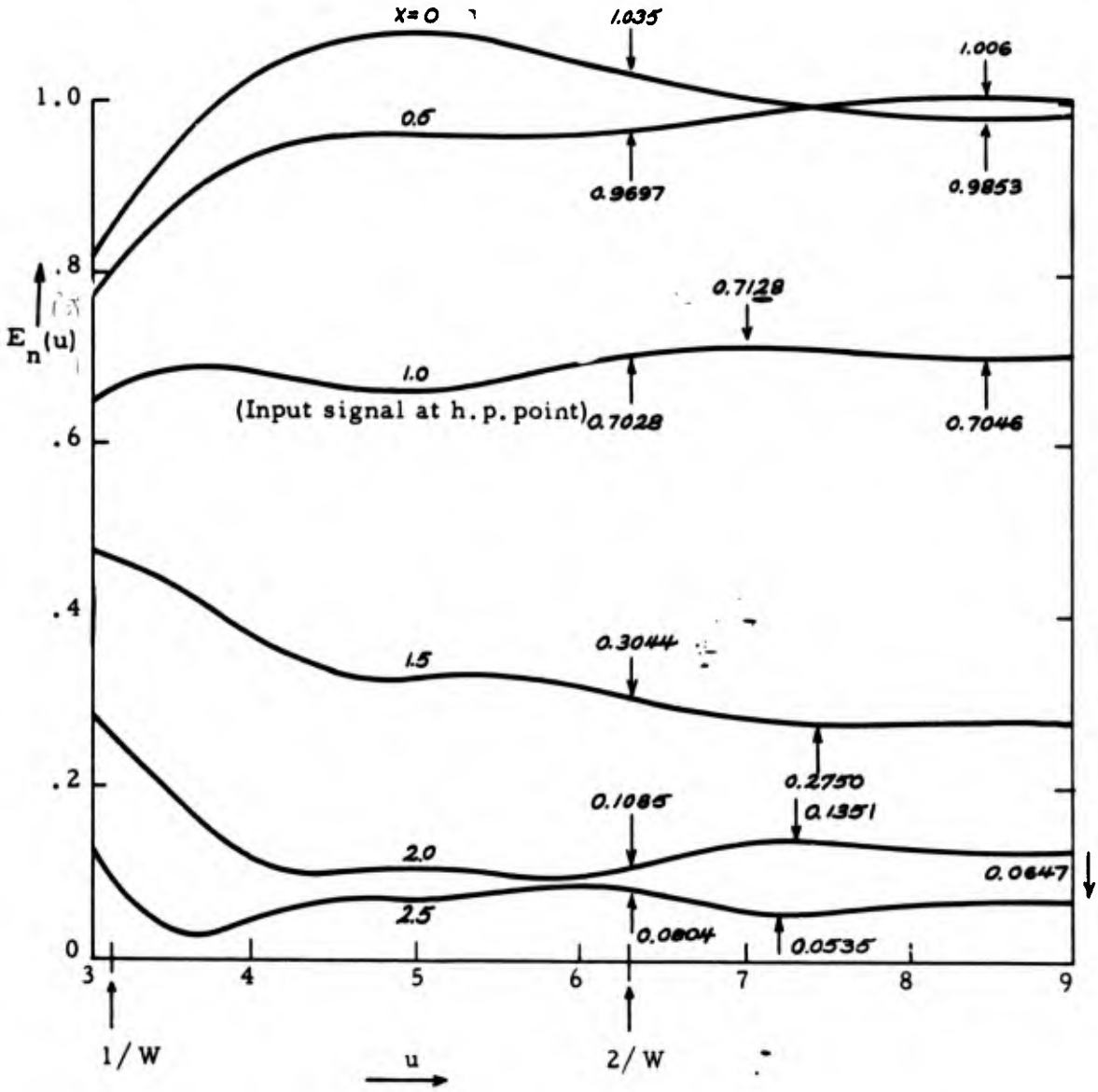


Fig. III-5. Envelope of Sine Wave Step Response for Detuned Three-Pole Butterworth Filter

The delay line bit rate required is determined by its length, τ, and the number of samples, N_s, that must be stored (parallel storage of samples is assumed in one delay line per bit of quantization):

$$B = \frac{N_s}{\tau} \tag{4}$$

Equations 1 through 4 may be solved to determine the resolution of the analyzer:

$$F = \frac{k(f_s)}{IB} \tag{5}$$

Equation 5 indicates that, for maximum resolution (i.e., min F) we should use minimum values of k and f_s, and maximum values of I and B.

To allow for use of 1 mc logic and delay lines, we should use $B_{max} = 10^6$. For greater system flexibility k_{min} should be 2.

The minimum value of f_s is derived in Section 4 of this appendix, and may be expressed as:

$$f_s \geq 2f_o + 6F \tag{6}$$

where f_o is highest signal frequency in the analysis band, W. For a 3 kc channel speech processing project, f_o = 3.3 kcps, and the resolution desired will not be less than F = 75 cps. Therefore the sampling frequency should be no less than about 7 kc. The above constraints determine the attainable resolution of the analyzer:

$$F_{min} \tilde{=} \frac{100}{I} \tag{7}$$

The upper limit on I is determined by the allowable skew, or analysis time:

$$T_a = N_f \tau \tag{8}$$

where N_f is the number of analysis frequencies (equivalent to the number of filters in a parallel filter bank analyzer). If the spacing between analysis frequencies is equal to the filter bandwidth, then

$$N_f = \frac{f_o - f_m}{F} = \frac{W}{F} \tag{9}$$

where f_m is the minimum frequency in the analysis band, W. For the voice identification project, W = 3 kc. The analysis time, T_a, can be determined from Equations 3, 7, 8, and 9:

$$T_a = \frac{W}{100\, f_s} I^2 \simeq \frac{I^2}{100} \tag{10}$$

For simplicity of logic design I shall be an integer. If analysis time is limited to 20 ms (50 spectra per second),

$$I^2 \leq 4 \tag{11}$$

or I = 2 or 1, for maximum resolution, and the analysis time is 5 ms. or 20 ms.

The maximum resolution can now be obtained from Equation 7:

$$F_{min} = 100 \text{ cps for } I = 1 \text{ and } T_a = 4.2 \text{ ms.} \tag{12}$$

$$F_{min} = 50 \text{ cps for } I = 2 \text{ and } T_a = 16.8 \text{ ms.} \tag{13}$$

where we have specified:

$k = 2$

$B = 10^2$

$f_o = 3.3$ kc

$f_s = 7$ kc

$f_o - f_m = 3$ kc

$T_s \leq 20$ ms

I is an integer

If the maximum resolution is not required, or if we can limit the analysis to a single pole filter and lower accuracy, we can reduce the analysis time and increase the sampling rate f_s by decreasing k to 1.5 or increasing F. To make the possible compromises more apparent, Equation 5 may be solved for f_s and Equations 3, 5, and 8 solved for T_a:

$$f_s = \sqrt{\frac{IFB}{k}} \tag{14}$$

$$N_f T = T_a = N_f \sqrt{\frac{kI}{FB}} \tag{15}$$

The DELTIC unit is essentially specified by Equation 14 or 15, and, since I is an integer and $B = 10^6$, there are only a few solutions for each choice of F, some of which are listed in Table 1 for F = 50, 75 and 100 cps. As indicated in Equation 13 there is only one solution for F = 50 cps for which $T_a \leq 20$ ms. (i.e., I = 2). Also, for I = 1, the max. resolution of 100 cps (and $f_s \geq 7$ kc) is verified by the value of $f_s = 6$ kc given for F = 75 cps and I = 1.

TABLE III-1. Possible Choices for a DELTIC Spectrum Analyzer

R_s	F(cps) (k=2)	I	f_s(kc)	τ	N_f	T_a	F (k = 1.5)
140	50	2	7.07	282	60	16.8	35
160	75	1	6.1	163	40	6.25	50
120	75	2	8.5	228	40	9.12	50
100	75	3	10.6	282	40	11.25	50
80	75	4	12.2	330	40	13.3	50
140	100	1	7.07	141	30	4,2	70
100	100	2	10	192	30	5.7	70
80	100	3	12.2	250	30	7.5	70
70	100	4	14	282	30	8.4	70

The number, I, of samples taken per recirculation of the delay line should be small for the following reasons:

1) to minimize the flip flop storage requirements – $N_q(I-1)$ flip flops are required for intermediate storage of samples, where N_q is the number of delay lines.

2) The delay T_a increases with $\sqrt{I}$.

3) The sampling frequency, and therefore speed required of the A/D converter also increases with $\sqrt{I}$.

On the other hand it is desirable to have a high sampling frequency, f_s, to allow for analysis of wider bandwidth signals, and a larger value of, I, for increased flexibility once the delay line length is fixed.

The speed up ratio, R_s, of the audio waveform accomplished by the DELTIC unit is determined by the delay line bit rate (assumed to be 1 mc.) and the sampling frequency, f_s:

$$R_s = \frac{1000}{f_s} \tag{16}$$

where f_s is in kc. The sweep width required of the spectrum analyzer is equal to the signal bandwidth times R_s.

Actually the compromise selected for implementation on this voice identification project involved the following quantities:

$R_s = 64$

$I = 2$

$f_s = 7900$ cps

$\tau = 127\ \mu$ sec

$N_f = 40$

$T_a = 5.1$

$N_q = 6$ bits

The attainable resolution with this device is approximately 50 cps for a single-pole filter, but to allow for smoothing the analyzer output before locating spectral peaks, an analyzing filter bandwidth of 250 cycles has been incorporated.

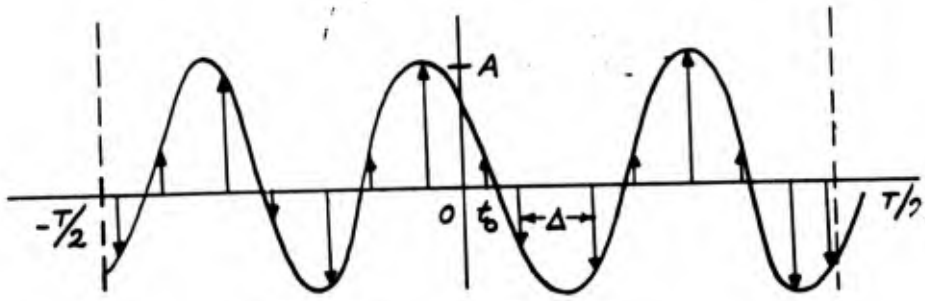
3. DERIVATION OF MINIMUM SAMPLING RATE, f_s

Earlier in Section 2 of this appendix, an expression was given relating the minimum sampling rate, f_s, to the maximum frequency, f_o, in the analysis band, and the analyzing filter bandwidth, F.* The remainder of this appendix is devoted to a derivation of this expression.

---

*From Reference 6.

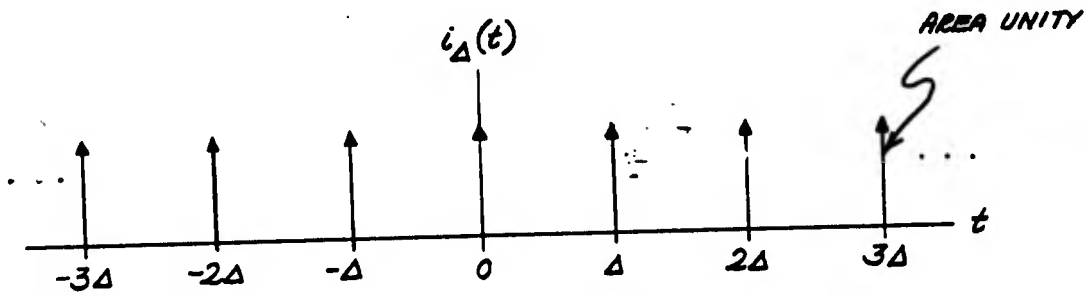
Given a sine wave of amplitude A, frequency f_1, phase θ (at t = 0), and of duration T. Sample it every Δ seconds. What is the resulting spectrum?



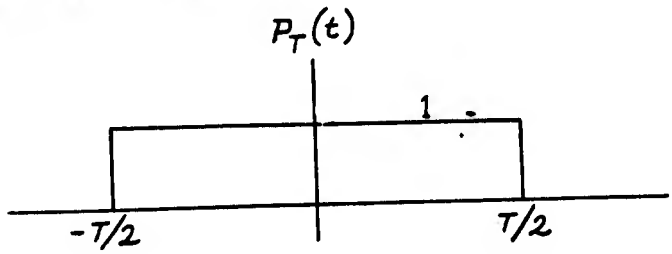
The finite impulse train can be expressed as

$$Ai_{\Delta}(t-t_0)\, P_T(t) \cos(2\pi f_1 t + \theta) \tag{17}$$

where $i_{\Delta}(t)$ is an infinite train of impulses,



and $p_T(t)$ is a finite pulse of duration T.



There is no loss of generality in the definition of zero time (t = 0), because we have allowed an arbitrary θ and t_0.

First the spectrum of

$$Ap_T(t)\cos(2\pi f_1 t + \theta)$$

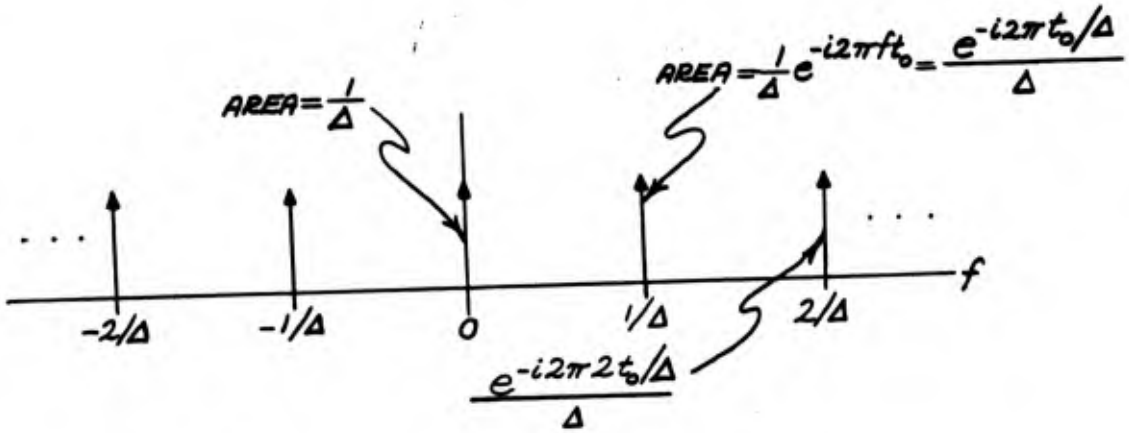
is the convolution of T sinc Tf and

$$\frac{A}{2}\left[e^{i\theta}\delta(f-f_1) + e^{-i\theta}\delta(f+f_1)\right]$$

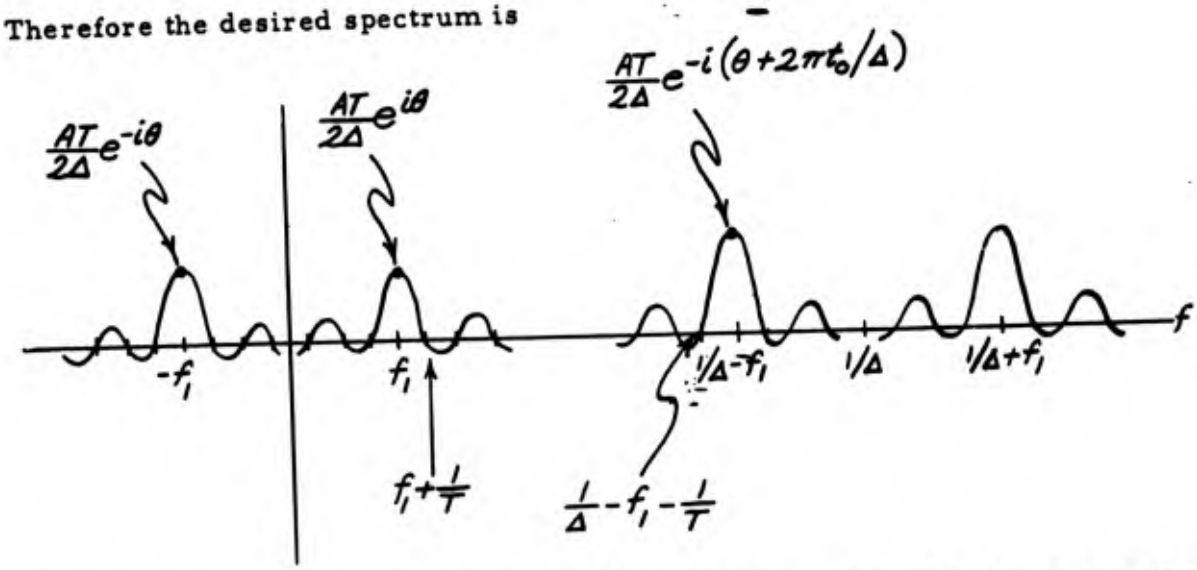
which is

$$\frac{AT}{2}\left[e^{i\theta}\,\text{sinc}\,T(f-f_1) + e^{-i\theta}\,\text{sinc}\,T(f+f_1)\right] \tag{18}$$

Next, the desired spectrum is the convolution of this spectrum with the spectrum of the infinite impulse train, which is



Therefore the desired spectrum is



Thus there is a component of power near the fundamental of the incoming wave, and is of bandwidth roughly $\frac{1}{T}$ cps. The next component is at $\frac{1}{\Delta} - f_1$ cps, and has bandwidth $\frac{1}{T}$ cps also. In order that these two components do not overlap significantly we must have

$$f_1 + \frac{3}{T} \leq \frac{1}{\Delta} - f_1 - \frac{3}{T} . \tag{19}$$

the factor of 3 being allowed for the "tails" of the spectrum. That is,

$$\Delta \leq \frac{1}{2f_1 + \frac{6}{T}} \tag{20}$$

is required for unambiguous determination of frequency content.

For a given desired frequency resolution of F cps, and filter of same bandwidth, there is no point in making T much larger than $1/F$, because the spectrum components would be too narrow to resolve by this filter. Therefore, we may choose $T = 1/F$.

If we are interested in analyzing input frequencies f_1 up to f_o cps, say, the tightest bound on Δ becomes

$$\Delta \leq \frac{1}{2f_o + 6F} \equiv \Delta_o; \text{ or, } f_s \geq 2f_o + 6F \tag{21}$$

This is the same expression as given in Equation 6.

For very small f_1, the "tails" of the negative frequency component overlap the positive component, thereby corrupting the values of the spectrum, the exact amount depending on θ. In order for this not to be a significant effect, we require that

$$f_1 - \frac{3}{T} \geq 0$$

or

$$f_1 \geq 3F.$$

Thus spectrum scanning starting above 3F is not affected by the relative phase of the sine wave.

Thus for frequencies f_1 such that

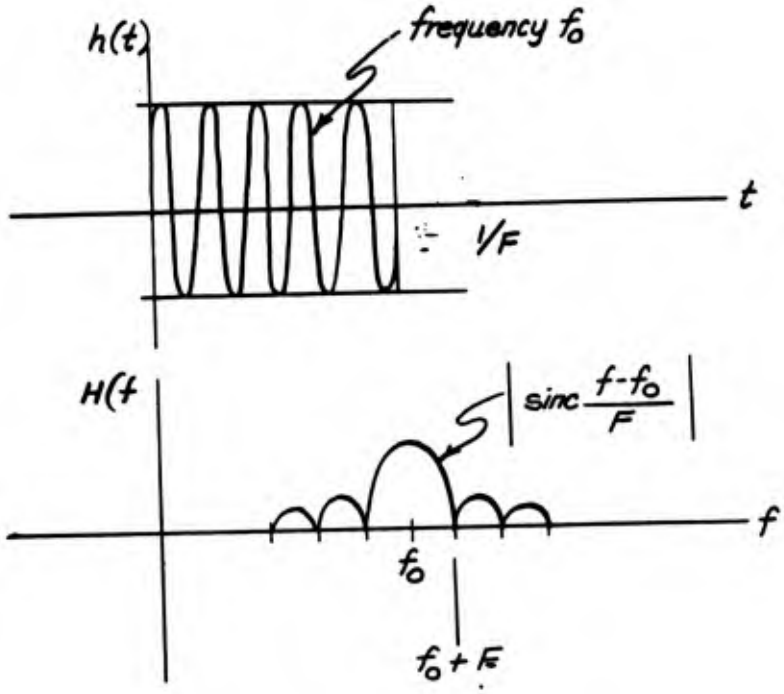
$$3F \leq f_1 \leq f_o ,$$

samples taken oftener than f_{so} samples per second, where

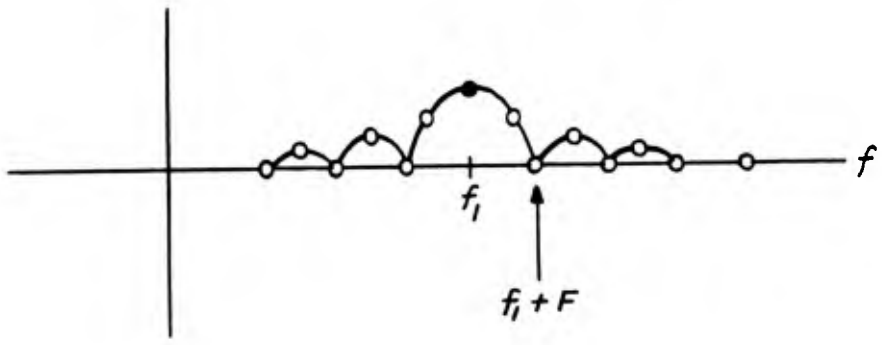
$$f_{so} = 2f_o + 6F,$$

guarantee no corrupting influence on the value of the spectrum estimation obtained.

To see how an actual narrow band filter of bandwidth F cps would respond to such an impulsive signal, consider the filter with impulse response and spectrum as follows:



Now suppose f_o is jumped in steps of F cps. Then samples of the filter output envelope at time 1/F after excitation would appear as below:

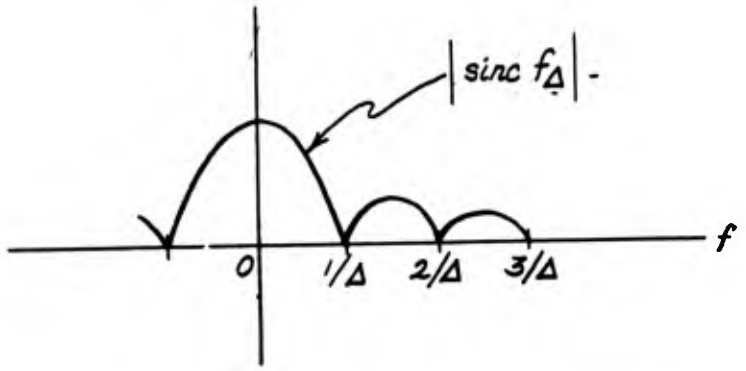


The solid dots correspond to the case where f_o jumps exactly f_1 on one trial, whereas the hollow dots correspond to an intermediate jump. The apparent frequency extent in the two cases is F and 2 F, respectively. Thus, there is a random "smearing", depending on the relative values of f_o and f_1. This has been eliminated in the analyzer by jumping in approximately $\frac{F}{3}$ cps each time.

If the impulsive waveform is passed through a boxcar circuit of duration Δ, the spectrum depicted above is changed by the factor

$$\text{sinc}\,\frac{f_o}{\Delta} = \text{sinc}\,\frac{f_o}{2f_o + 6F} \simeq \text{sinc}\,\frac{1}{2} = \frac{2}{\pi}\ ,$$

as indicated in the diagram.

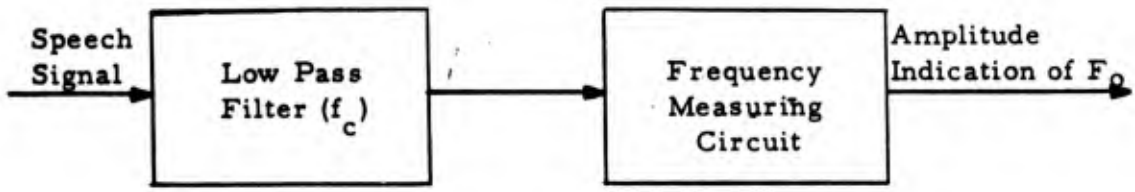


In general, this is a known factor for any frequency, namely sinc $\frac{f_1}{\Delta}$, and can be compensated for.

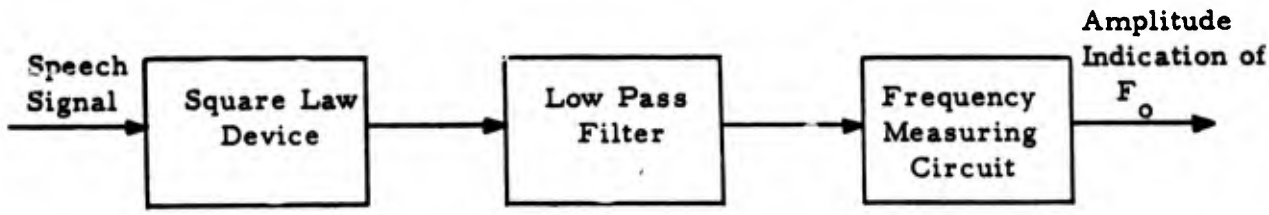
APPENDIX IV. PITCH EXTRACTION IN THE PRESENCE OF NOISE AND SPEECH BANDWIDTH TRUNCATION

The operational goal of a pitch extractor is to derive an indication of the fundamental frequency of the speech waveform, through an examination of whatever portion of this signal is available. This indication is required during voiced speech intervals, within which the speech waveform exhibits quasi-periodicity.

A rudimentary conventional frequency domain pitch extractor performs this operation by first low pass filtering the speech waveform and converting the zero crossings of the resulting waveform to an amplitude indication in a frequency measuring circuit (Figure IV-1(a)). Under certain conditions this method can yield a reliable pitch indication. However, for practical application this rudimentary method has several drawbacks. First it is clear that if the pitch fundamental is not present in the speech signal available for processing (as a result of bandpass filtering in a microphone, for instance), then a high pitch indication will result. A basic approach to the removal of this limitation is to pass the available speech signal through a nonlinear, no-memory device (such as a rectifier), which produces an output with fundamental periodicity determined by the minimum spacing between spectral lines in the spectrum of the input. For voiced speech signals this spacing is the pitch fundamental, F_o. Thus, as indicated in Figure 1(b) an amplitude indication of pitch may be obtained by low pass filtering the nonlinear device output and utilizing a frequency measuring circuit. While this improved method restores a missing pitch fundamental, it is critically dependent on the presence of at least two adjacent harmonics in the input speech signal. Moreover, even when this condition is satisfied there remain the problems of (a) determining the proper low pass filter cutoff and (b) distinguishing between the doubled



(a) Rudimentary Frequency Domain Extractor



(b) Improved Frequency Domain Extractor

Fig. IV-1. Rudimentary and Improved Frequency Domain Pitch Extractors.

fundamental frequency and the probably smaller amplitude F_o component when the pitch fundamental happens to be present at the input. Both of these problems can only be solved for the methods depicted in Figure 1 through acquisition of "a priori" information concerning pitch, i.e. through adjustment of the low pass filter cutoff frequency for each speaker. This debilitating feature of the basic frequency domain extractors, coupled with their susceptibility to the presence of noise in the input speech signal, indicate that a more sophisticated approach to the problem is required.

Several investigations of methods of improving the performance of pitch extractors have been conducted in the past few years. There are apparently two guiding philosophies behind the majority of improved approaches which have been suggested: (1) utilize only that portion of the frequency spectrum which provides useful and relatively noise-free signals for processing, and (2) utilize several pitch extractors, each of which is designed to work well for a small range of pitch values, or under certain conditions, and switch between the outputs of these several devices to utilize the best estimate.

Methods embodying these philosophies have been examined by several investigators, both by construction of special purpose devices and by computer simulation. In the remainder of this appendix, two somewhat different methods are described, and a means of combining them outlined. The combination has been implemented with an extractor designed and constructed during the course of this project.

FORMANT CHANNEL FILTERING

Consider first the characteristics of voiced speech signals which might indicate that certain portions of their spectra would be more useful than others, particularly when noise is present. It is clear that the spectral lines located near formant positions offer greater potential for estimation of spacing between spectral lines, than do the lower amplitude portions of the spectrum.

Further, if a nonlinear operation is employed to generate difference frequencies, then it is necessary that any spectrum selection operations must involve filtering with resolution less than the highest pitch frequency to be encountered. These considerations suggest that a potentially worthwhile approach to pitch extraction when noise is present but the pitch fundamental is absent, is to pass the input speech signals into several extraction channels, each of which involves bandpass filtering the signal at a region in the frequency spectrum corresponds to high amplitude spectral content. Ideally, one of these pitch extraction channels might be created for each of the formants, or spectral concentrations of energy generated in voiced sounds. Thus, we may regard these as "formant channels". As indicated in Figure IV-2, one way of utilizing the signals generated by the bandpass filters is to extract a pitch estimate within each channel, and through logical operations select one of these estimates for the pitch indication. One approach to the logical operations is to select that estimate which corresponds to the lowest pitch, with certain reservations. The minimum selection is based on the presumption that the minimum spacing between spectral lines in any formant channel is the pitch frequency. The basic reservation to this minimum selection is that spurious estimates resulting from low input signal levels be rejected. The rejection can be performed by inhibiting the output of any formant channel for which evidence is available that the input signals are in doubt. A straightforward method of implementing this inhibition is to envelope detect the output of the formant filter, and inhibit the output if this detected signal level falls below a threshold value.

In order to realize the potential improvement afforded by formant filtering, it is necessary that some means be available for tuning the bandpass filters to spectral regions encompassing concentrations of signal energy. There are several levels of sophistication with which this filtering control can be exercised. The simplest method is to use a single channel, with a fixed bandpass filter encompassing a frequency range within which formants generally

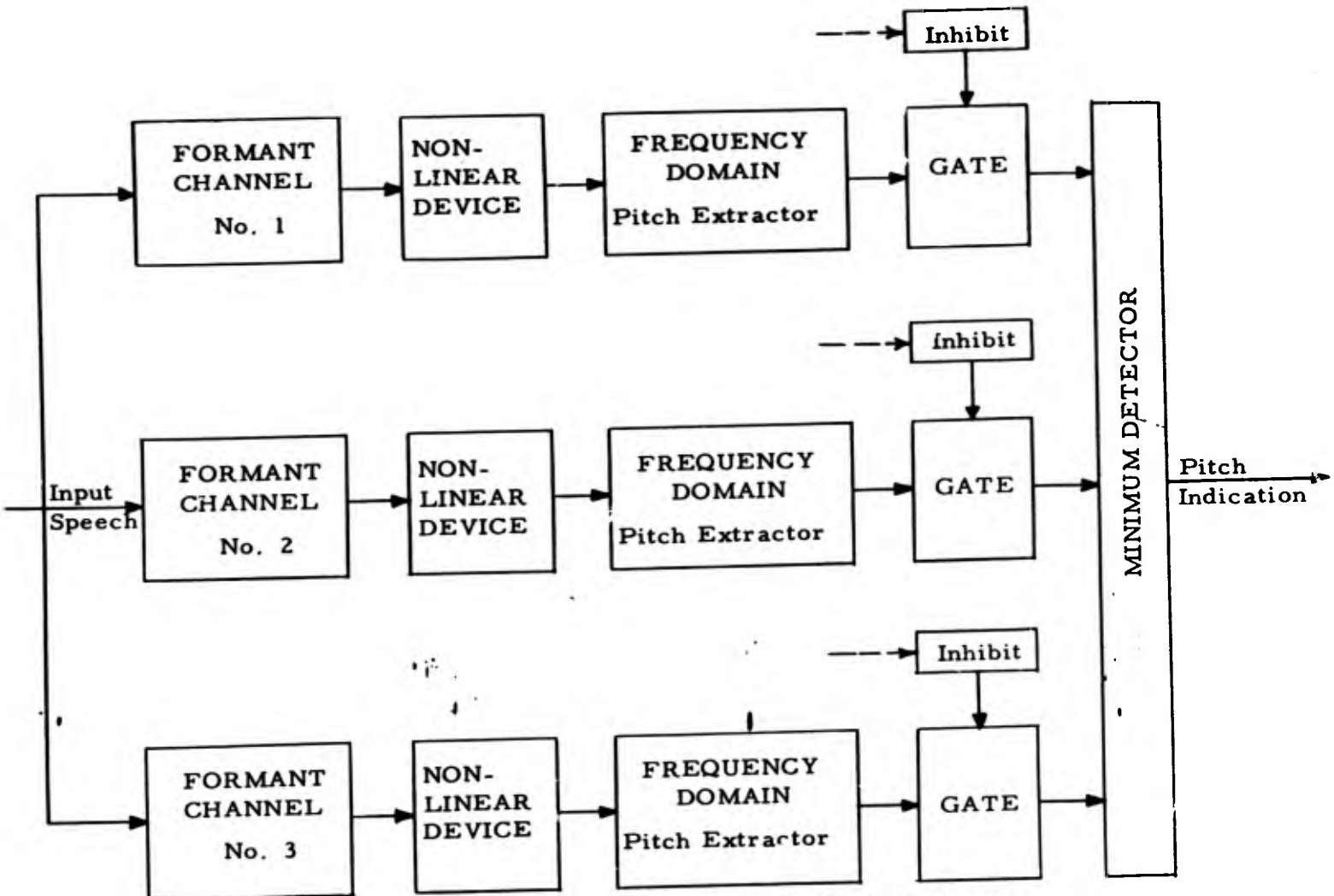


Fig. IV-2. Parallel Frequency Domain Pitch Extractors

occur. This frequency range would be determined primarily by the available bandwidth of input speech signals. For speech confined to the 3 Kc channel between 300 cps and 3300 cps, for instance, we have found that a 2 Kc filter between 300 cps and 2300 cps usually provides sufficient information for reliable pitch extraction under a wide range of broadband noise conditions.

The next level of sophistication which could be incorporated in the band-pass filtering approach is the utilization of several (no more than 3) fixed-frequency filters, each of which is tuned to a formant range. Although we have not yet performed extensive tests with this arrangement, it appears that this method will afford improved noise immunity.

Perhaps the most sophisticated means of controlling the formant filters would be to tune them to current estimates of formant positions. These estimates could be provided directly by a formant extractor, based for example on spectral peak locations. Although comprehensive tests have not yet been performed, a peak picker has been designed and constructed for use with an 18-channel filter bank to obtain estimates of formant positions. The output of the peak-picking unit may be used to drive a voltage controlled oscillator to change the center frequencies of the formant filters, through the use of a pulse spacing-to-amplitude conversion device.

A PITCH FREQUENCY RESOLVING METHOD

As pointed out earlier, one of the problems associated with a frequency domain pitch extractor is that the appropriate frequency range over which the initial low pass filtering should be performed is either not known in advance, or may require manual adjustment if other steps are not taken. If, for instance, the pitch fundamental is present at the input, but the second harmonic is suppressed (by falling outside the first formant range in the sound | i |, say), then the pitch extractor may produce an output corresponding to twice the pitch fundamental. Or, more directly, the problem of distinguishing between the second

harmonic of a low pitch fundamental and a high pitch fundamental still remains for each of the parallel pitch extractors involved in, for instance the method illustrated in Figure IV-2.

Fant[11] has suggested an approach to this problem which is similar to the logic involved in the output of the parallel formant channel extractors of Figure IV-2, and in fact carries this idea one step further. Specifically, for each pitch extractor, the output of the nonlinear device is passed to three parallel "resolving" filters, each of which spans somewhat less than one octave, and which together span the entire range of permissable pitch frequencies (taken here as 75-350 cps). As indicated in Figure IV-3, the output of each resolving filter is then passed to a frequency measuring device. The outputs of these three resolving channels may then be examined and processed to produce a single estimate of pitch.

Since each of the resolving filters spans less than an octave, in no case will the pitch fundamental and a harmonic lie within the pass band of any single filter. Further, if a waveform exhibiting the pitch fundamental is present at the input to the resolving channels, then it will lie within the passband of one of the filters. Thus, if the processing which precedes the resolving filters has accomplished its appointed task of converting the input speech to a waveform with fundamental frequency equal to the pitch frequency, then at least one of the resolving channels should produce the proper pitch indication. Selection of the proper channel output can be accomplished in the same way as suggested before, namely minimum selection with reservations.

To extract pitch on this project, the pitch frequency resolving method has been implemented. This technique was found to produce a high immunity to broadband noise at speech signal-to-noise ratios above 4 db, even when the most crude "formant filtering" is employed (namely, bandpass filtering between 300 cps and 2000 cps). Sample outputs of the three resolving filters are

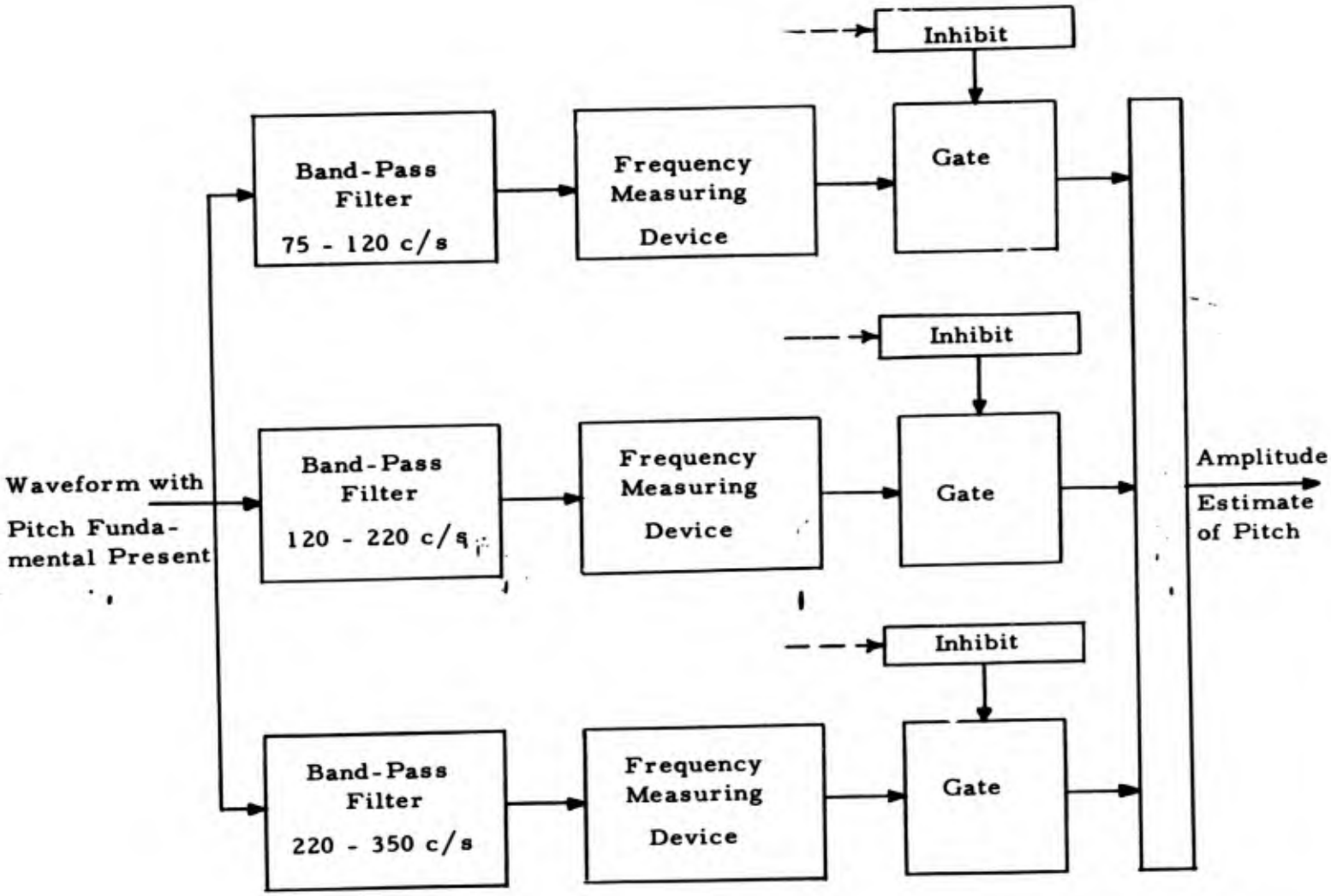


Fig. IV-3. Pitch Resolving Technique

shown in Figure IV-4 for three signal-to-noise ratios. The pitch extractor input for these curves was a slowly ascending pitch synthetically generated waveform corresponding to the utterance /a/. The pitch varied between 75 cps and 300 cps. These curves indicate that with appropriate switching between channels, a reliable pitch indication can be obtained.

Most of the implementation problems associated with this type of extractor arise from this switching task. Direct inhibition of a resolving channel output through comparison of the input envelope amplitude with a threshold generates spurious outputs at the ends of voicing intervals. These transient problems have been overcome through the introduction of appropriate delays in the inhibiting circuits. These delays introduce a 60 msec lag between onset of voicing and registration of a pitch indication, but this lag is commensurate with the time it takes for the frequency meter to respond*.

The curves in Figure IV-5 indicate the final output of the extractor for the same input as used for generating the curves in Figure IV-4.

This pitch extractor can easily accomodate any method of improving rejection of other types of noise at its front end. In particular, it is contemplated that the effects of constant frequency interference will be investigated next with this device.

---

* For the purpose of clue measurement, this rather long averaging interval provides some smoothing at the expense of losing the pitch indication for the first one or two samples in a voiced interval.

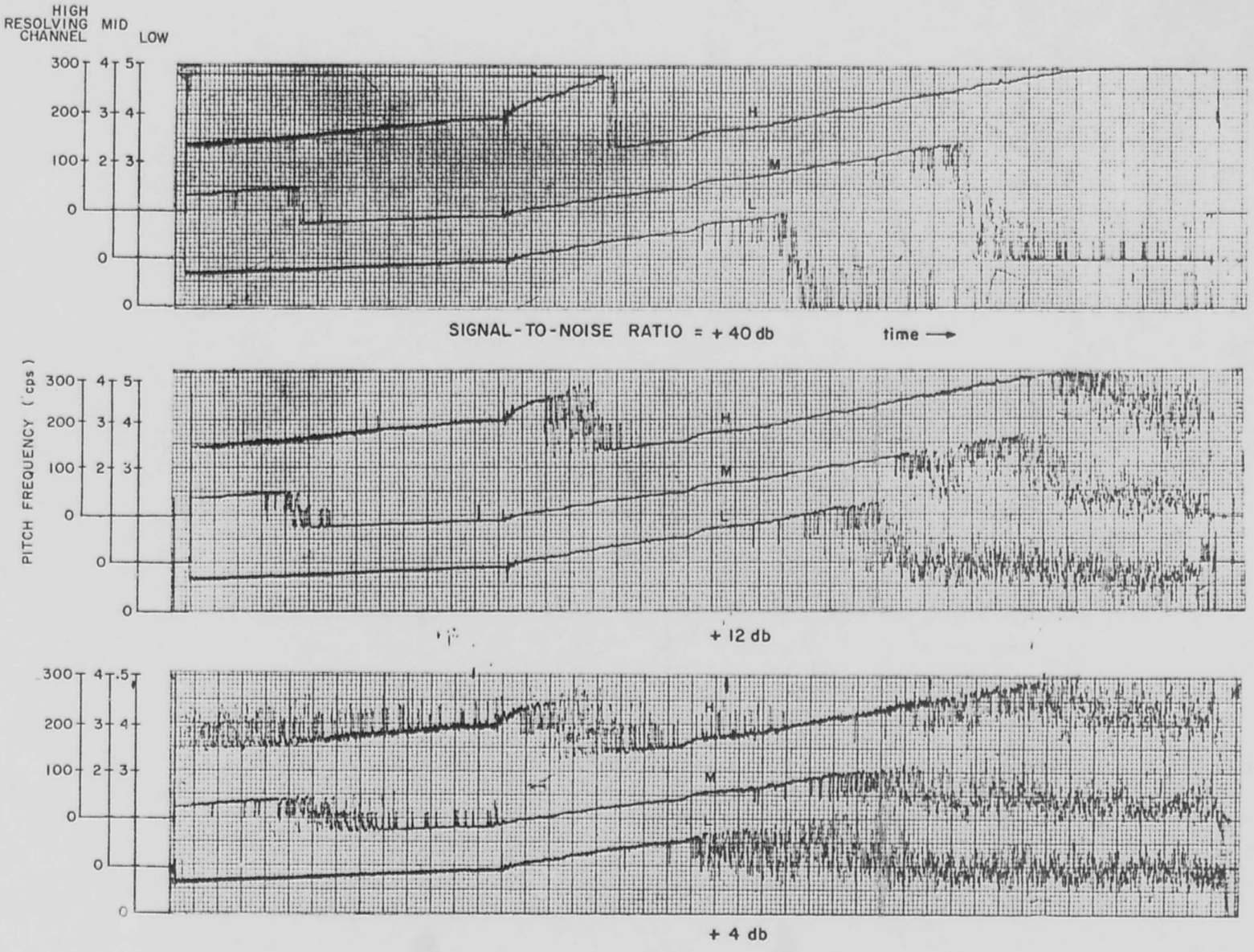


Fig. IV-4. Pitch Resolving Channel Outputs for a Slowly Ascending Pitch, Synthetic Speech Sample Input for Three Signal-to-Noise Ratios, with No Formant Channel Filtering

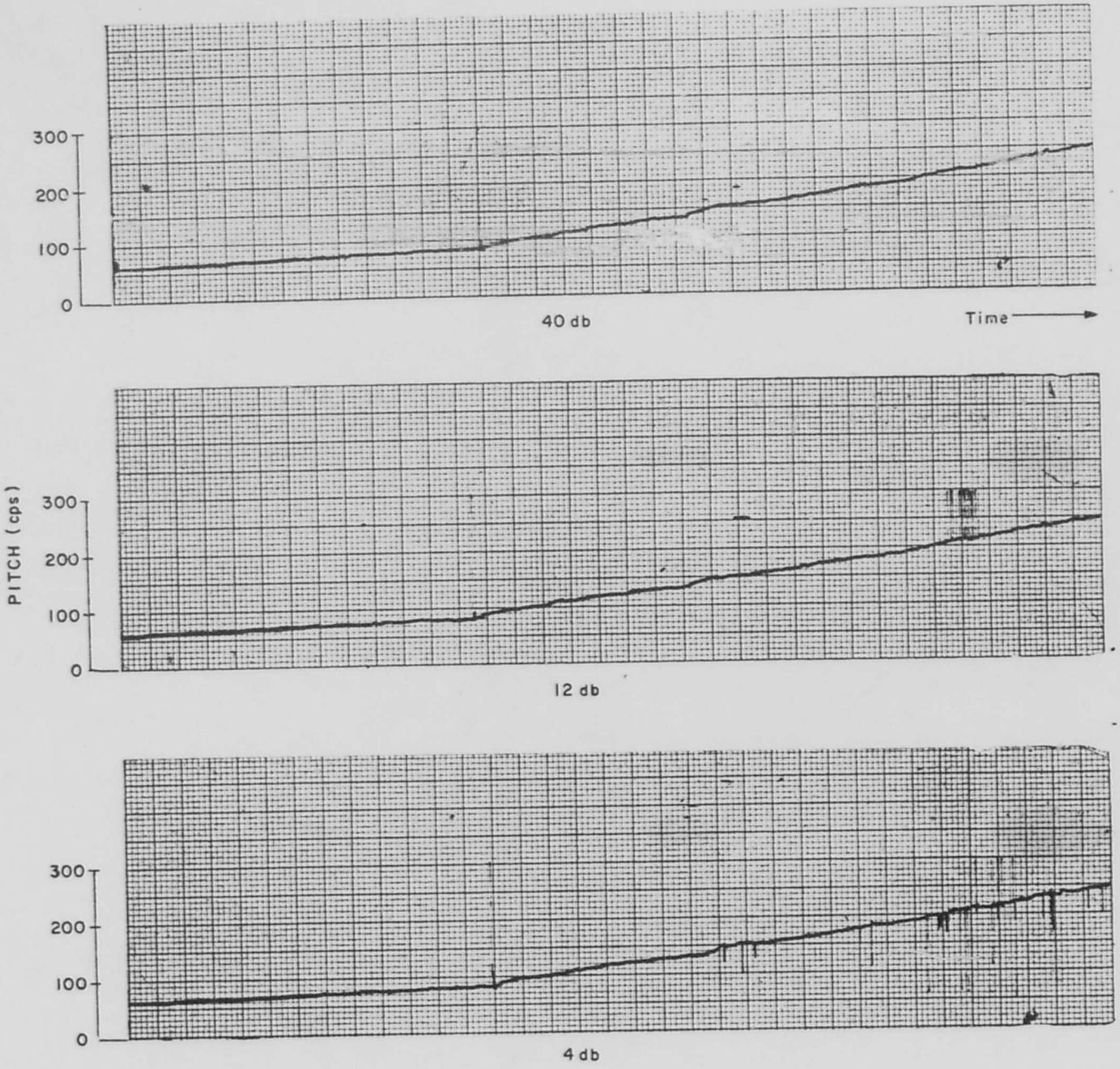


Fig. IV-5. Pitch Extractor Output for a Slowly Ascending Pitch, Synthetic Speech Sample for Three Signal-to-Noise Ratios, with No Formant Filtering

REFERENCES

1. Final Report on contract AF30(602)-2499, Voice Identification General Criteria, RADC-TDR-62-278, 16 May 1962.

2. Kersta, L.G., "Voiceprint - Identification Infallibility", paper presented at the 64-th Meeting of the Acoustical Society of America, November 1962.

3. Pruzansky, S. "Pattern Matching Procedure for Automatic Talker Recognition", JASA, Vol. 35, No. 3, March 1963, pp. 354-358.

4. Smith, J.E. Keith, "A Decision - Theoretic Speaker Recognizer", presented at the 64-th Meeting of the Acoustical Society of America, 8 November 1962.

5. Weiss, M.R., and Harris, C.M., "Computer Technique for High-Speed Extraction of Speech Parameters", JASA, Vol. 35, No. 2, February 1963, pp. 207-214.

6. G.E. Valley and H. Wallman, Vacuum Tube Amplifiers, MIT Lab. Series, McGraw-Hill, 1948.

7. I. Pollack, J.M. Pickett, and W.H. Sumby, "On the Identification of Speakers by Voice", JASA, Vol. 26, No. 3, pp. 403-406, May 1954.

8. R.W. Peters, "Studies in Extra-Messages: Listener Identification of Speakers' Voices Under Conditions of Certain Restrictions Imposed Upon the Voice Signal". Joint Project NM 001 064.01, Report No. 30, Pensacola, Fla.; Ohio State Univ. Res. Found. and Naval School of Aviation Medicine, 1954.

9. R.W. Peters, "Studies in Extra-Messages: The Effect of Various Modifications of the Voice Signal Upon the Ability of Listeners to Identify Speakers' Voices", Joint Project Report No. 61, Bureau of Medicine and Surgery Project NM 001 104500, U.S. Naval School of Aviation Medicine, 1956.

10. Final Report on contract AF19(628)-1604, Pattern Recognition Research, AFCRL-63-548, 14 June 1963

11. Fant, C.G.M., "Speech Analysis and Synthesis", Summary Report, contract AF61(052)-342, January 31, 1961.