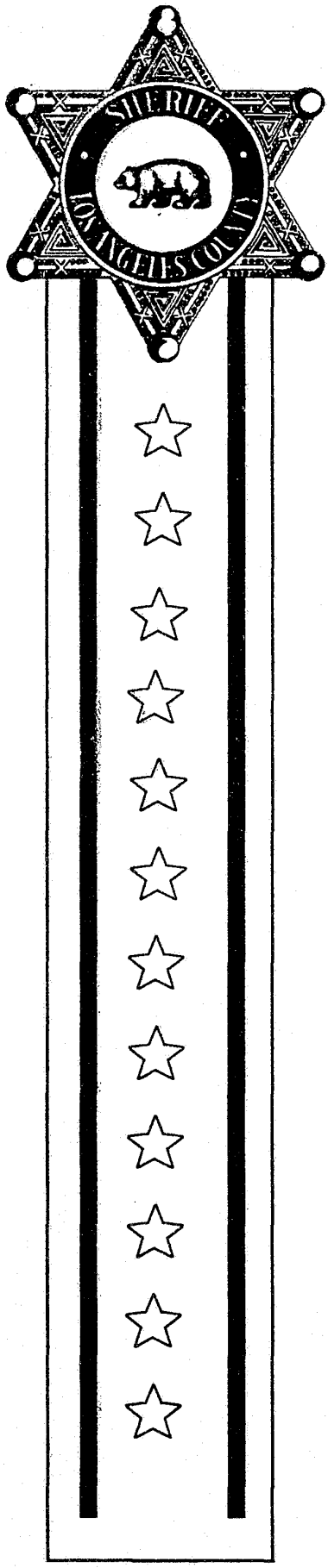


PROJECT C. A. V. I. S.

(COMPUTER ASSISTED VOICE IDENTIFICATION SYSTEM)

FINAL REPORT

NATIONAL INSTITUTE OF JUSTICE

GRANT NO. 85-IJ-CX-0024

OCTOBER 1989

LOS ANGELES COUNTY
SHERIFF'S DEPARTMENT
SHERMAN BLOCK, SHERIFF

FINAL REPORT

PROJECT C.A.V.I.S.

(COMPUTER ASSISTED VOICE IDENTIFICATION SYSTEM)

NATIONAL INSTITUTE OF JUSTICE

GRANT NO. 85-IJ-CX-0024

LOS ANGELES COUNTY SHERIFF'S DEPARTMENT

SHERMAN BLOCK, SHERIFF

OCTOBER 1989

This final report has been prepared by the research staff named below.

Hirotaka Nakasone, Ph.D.
Craig Melvin, Sgt.

Los Angeles County Sheriff's Department
October, 1989

COMPUTER ASSISTED VOICE IDENTIFICATION SYSTEM (C.A.V.I.S.)

FINAL REPORT

GRANT NO. 85-IJ-CX-0024

LOS ANGELES COUNTY SHERIFF'S DEPARTMENT

OCTOBER 1989

ABSTRACT

Project C.A.V.I.S. is a scientific research effort to develop a computer based system to assist forensic voice examiners in their task to identify or eliminate suspected voices associated with criminal activity. September 30, 1989 marks the culmination of the four year research effort in which a forensic audio work station was developed with capabilities to analyze voices and other recorded forensic audio events.

The major goal of this project is to develop a system that is capable of dealing with transmission-independent, text-independent voice data, and rendering objective decisions. Throughout this project, our main target has been to develop a 'man-machine' interactive system of voice identification, as an investigative tool, and eventually as a court room tool. Numerous speech parameters were extracted and tried - some of them kept evolving for improvement. The research revealed that high identification performance rates can be accomplished by using a combined set of speaker specific parameters.

This report describes our work on voice data processing techniques, procedures of parameter extraction, strategies used in the speaker specific parameter selection, performance rates in voice identification and verification processes, and implications for future application.

ACKNOWLEDGMENTS

The Los Angeles County Sheriff's Department is deeply appreciative of the financial support and encouragement expressed by the Project's contributors who made the research effort possible.

The National Institute of Justice provided grant funding in the amount of $220,880 to support the project during the first two years and provided a grant extension in the amount of $185,200 for the remaining two years.

The Project was also assisted by the United States Secret Service who provided $60,000 in funding and the loan of three microcomputer systems.

Two private organizations also provided assistance to the project: The Margaret W. and Herbert Hoover, Jr. foundation provided $22,500 in funding and the Ralph T. Weiler foundation provided $1,000 in funding to Project C.A.V.I.S..

The Los Angeles Sheriff's Department provided soft match funding of $547,994 over the four year period.

We would like to express our sincere appreciation to the following scientists who served as advisory committee members providing technical information and insightful comments during the first phase of this project: Dr. Glenn Bowie, Dr. George Papcun, Dr. Michael J. Saks, and Lt. Lonnie Smrkovski. In particular, we wish to mention that some of the prototype software algorithms developed by Dr. Bowie for us have become a significant part of various aspects of this research project.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

1 INTRODUCTION

1.1 Introduction

This is the final report on Project C.A.V.I.S., Computer Assisted Voice Identification System, a research effort funded primarily by the National Institute of Justice under grant No. 85-IJ-CX-0024. The report presents the original project goals, scope, experimental procedures in speech signal processing, speech parameter extraction, voice identification and verification operations, and implications for future applications as a forensic investigative tool.

The voice has long been used as a means to identify criminals. Currently, the Los Angeles County Sheriff's Department uses the combined method of aural and spectrographic analysis for voice identification. The need and importance of developing an objective, expedient and reliable technique of speaker identification is increasing.

In addition to the Los Angeles County Sheriff's Department and a few other local law enforcement agencies, The Federal Bureau of Investigations (Koenig, 1986) has been providing speaker identification services but limiting its use as an investigative tool. Speaker or voice identification, by the aural and spectrographic method, continues to be controversial regarding its reliability and acceptability as a court room tool. The main source of the controversy is related to the subjectivity in the decisions rendered by a human examiner. Another inherent problem associated with the spectrographic method is that it is very time consuming and cumbersome.

Unlike the rigorous research effort in the area of speech recognition, there seems to be only a handful of research groups that are engaged in speaker recognition in general fields. Speaker recognition in commercial applications, such as security access control, has shown to provide a high verification performance as high as 99.9% (Naik and Doddington, 1987). In these cases the speaker is considered to be cooperative as he or she utters prescribed phrases and the system commonly uses a fixed type of transmission system for all voice entries.

In contrast, various difficulties are associated with speaker recognition in forensic environments. Criminals are inherently uncooperative. They do not read prescribed phrases, unknown paths and transmission channels are employed in the course of committing the crime, and multiple speakers involved in conversation is common. Under such circumstances, full automation of speaker identification appears inhibitory.

Extra cautions are always inevitable in the variety of real cases in screening, editing, and segmenting the right voice sources. Text-independency is a feature that is of great attraction to forensic use. A drawback of the currently practiced method (aural and spectrographic comparison) is that it requires verbatim texts from all speakers involved. The need for verbatim voice samples generates some constraints, such as painstaking manual word matching. Further, it usually involves lengthy legal procedures to obtain the verbatim voice samples from the suspects and alerts the suspect that he is being

investigated.

In implementing a computer based voice identification system to overcome the above mentioned problem, we are interested in achieving two types of voice identification procedures: speaker identification and speaker verification.

Speaker identification is typically defined as a process in which a voice sample of an unknown speaker is compared with two or more voice samples collected from multiple known speakers, and the one from the known group is chosen whose voice is the closest to that of the unknown[1]. On the other hand, speaker verification is a process in which two voice sources are provided for comparison, and the task is to determine, according to a prescribed criterion value, whether the two voices belong to the same speaker (case of verification), or to different speakers (case of rejection).

1.2 Background and Motivation

The concept of being able to determine whether two recorded voices were uttered by the same speaker is based on the combination of two basic premises. The first being the unlikelihood that the physiology and anatomy of the voice

---

[1] The term 'voice identification' is a generic name which encompasses various aspects in the process of determining the identity of an unknown speaker, given a person's voice samples and voice exemplars collected from one or more known speakers.

production mechanism[1] for any two people would be exactly the same. Secondly, that the manner in which a person has learned to speak is going to be characterized by a multitude of differing external influences[2]. When we combine these two variables of biological and learned speech characteristics, the statistical basis is derived that no two people will exhibit that exact same speech characteristics.

The Los Angeles County Sheriff's Department has been active in the forensic analysis of recorded audio evidence since the early 1970's. Initially, the audio laboratory concentrated its efforts on the intelligibility enhancement of recorded conversations. The sources and quantity of the recordings increased as modern technology provided society with a variety of communication and recording media. Inherently, the laboratory began to provide additional forensic support in the areas of transcript verification, tape authentication and analysis of recorded acoustic events such as explosions, gunshots, aircraft performance and voice identification.

As mentioned, the method of voice identification currently being used by the laboratory at LASD is the combination of critical aural listening and the comparison of audio spectrograms. This procedure is encumbered by the requirement to

---

[1] The voice mechanism consists of the physiological and anatomical parts, beginning with the vocal cords, the resonance system (pharynx, vocal cavity, nasal cavity), and articulators (teeth, lips, tongue, and jaw).

[2] The manner in which a person learns to speak is influenced by his environment, which consists of his parents' way of speaking, his peers that he grew up with, and differing locales where he may have lived.

have the exact phrases available to compare and the lengthy and tedious procedures to compare and analyze the spectrograms.

The realization soon came that a system was needed which could aid the voice examiner in arriving at his decision. Ideally, the system would be able to do this without having to have the same text spoken, be able to work with varying transmission media, provide objective probabilities and still be a time saving procedure. Although private industry is making great advancements in the application of speech and automation, they have not focused on the unique application of voice identification to the forensic environment. Thus, the Los Angeles County Sheriff's Department assumed the leadership role by establishing its own research effort.

1.3 Need For Computerization

The development and technique of using an audio spectrograph to identify voices arose during the second world war. In the early 1970's these procedures were tested and refined. In an attempt to automate the process, the obvious transition to make was to incorporate the fast processing and analysis capabilities of computers. The research staff chose to use microcomputers in the development and final configuration of the C.A.V.I.S. System. This approach allowed for tremendous costs savings over mini or main frame computers and provided ease in making the workstation multi-tasking.

1.4 Comparison Of Voice Identification Techniques

To familiarize the reader with the currently used method of spectrographic analysis (commonly known as voice print), a brief summary follows:

As previously mentioned, in order to utilize the technique of voice print analysis, it is essential that the two recorded voices to be compared contain similar texts, which enables verbatim pattern matching comparisons. An instrument called a sound audio spectrograph is used to produce the voice prints (See Figure 1.1).

Each print reveals an individual speaker's speech characteristics of a word or phrase in the frequency, intensity, and time domains (See Figure 1.2). A voice examiner will analyze the prints paying attention to timing and frequency relationships. He will also perform a critical listening comparison of the two voices paying attention to tonal quality, pitch rates, articulation, and any signs of pathologies. The examiner, relying on his expertise, then forms his opinion as to whether the voices belong to the same speaker. This opinion is based primarily on the examiner's subjective expert judgment. An excellent summary of the theories, methodology and historical reviews on forensic voice identification is found in a book by Tosi(1979).

Usually, an examiner will offer an opinion in one of the following manners:

Identification

No Decision

Elimination

The examiner then follows his opinion by giving an indication as to how confident he is regarding his decision. This confidence level may be assigned as one of the following:

Low

Moderate

High

Very High

It would be difficult for an examiner to offer greater degrees of diversity in his decision. Examiners in the past have been asked during testimony to assign a percentage level to their confidence. Indeed, how would an examiner be able to distinguish between a psychological confidence of 81% versus an 84%, specially, if he were asked to do the same exam again a year later ? The methodology applied in Project C.A.V.I.S. will be discussed in great detail in Chapter 3 and 4, but a brief overview is offered here.

Unlike the spectrographic method, the C.A.V.I.S. approach will be able to analyze and compare voices with different recorded text, hence, text-independence. C.A.V.I.S. focuses more on the tonal activity of the speaker and microscopically characterizes the manner in which he controls his glottal wavelets. As an example, with C.A.V.I.S., an individuals pitch

is not characterized by simply the mean of his pitch, but rather the total distribution of his pitch production is characterized and reduced to three statistical parameters. Once the speech characteristics for the two comparison samples have been extracted, an assignment of a "Proximity Index" is made which indicates the degree of similarity between the two samples. C.A.V.I.S. dynamically determines which speech features are best to use for a given comparison. The "Proximity Index" is derived from the distribution of the general population obtained to date during the research project.

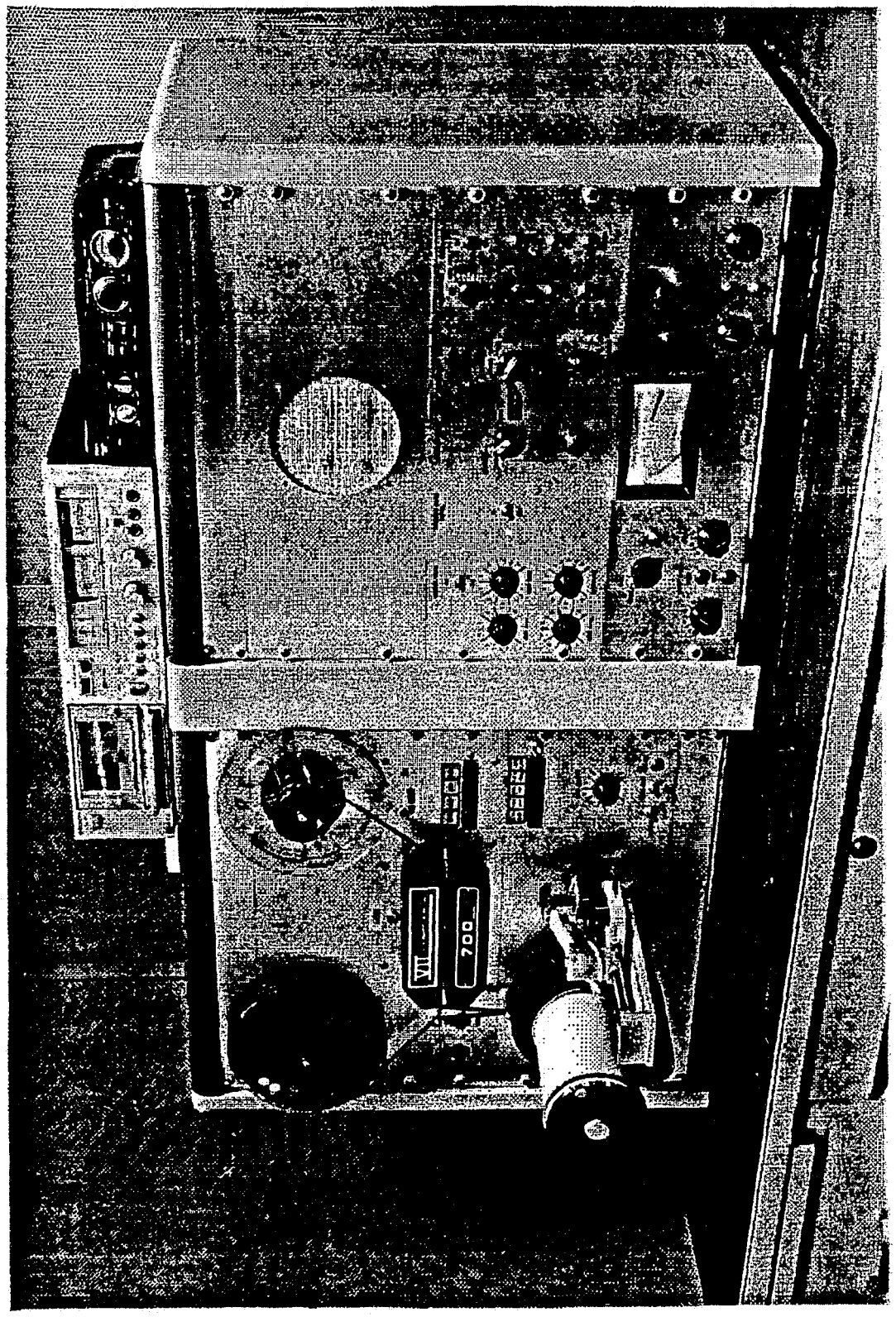


Figure 1.1 Photograph of a Sound Spectrograph.

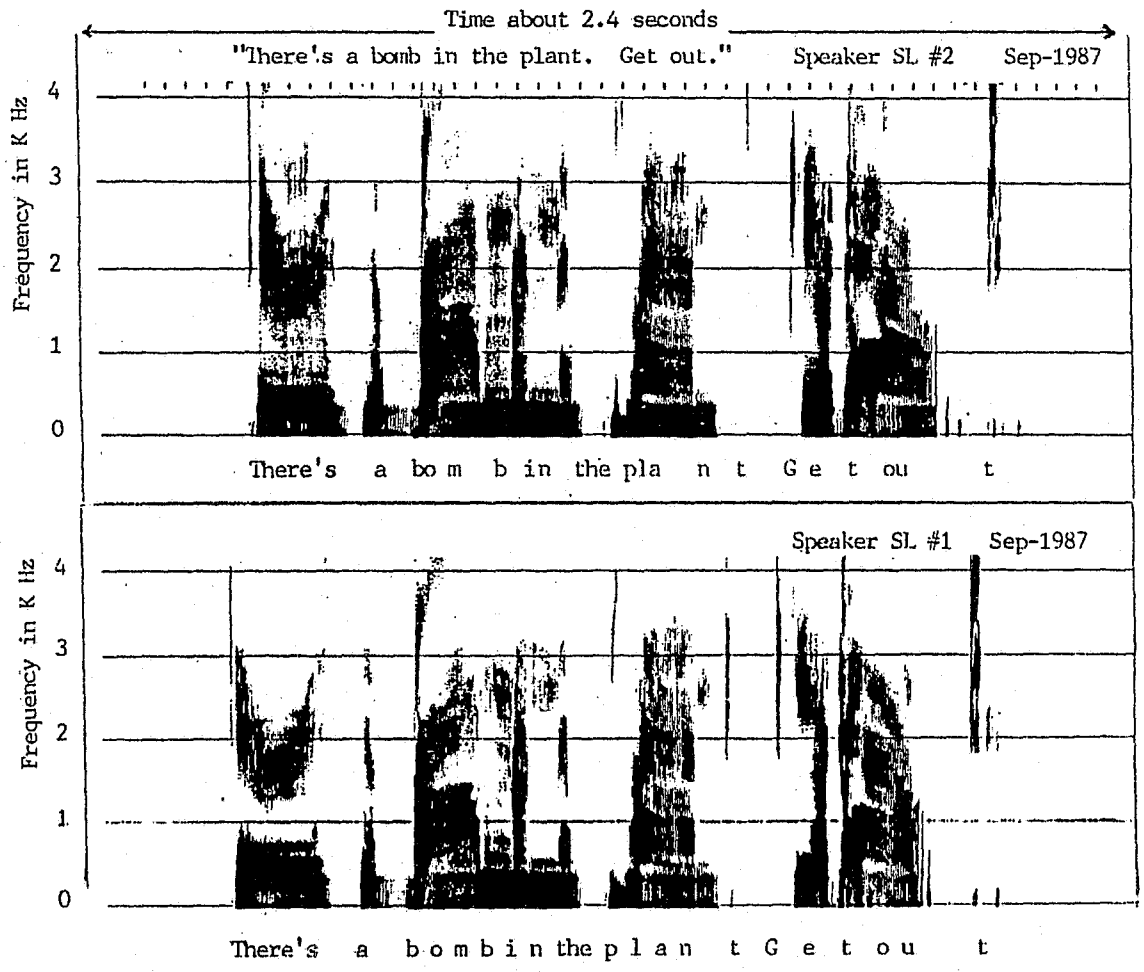


Figure 1.2 Sound spectrograms prepared from the same speaker.

1.5 Comparison To Other Systems

The uniqueness of the C.A.V.I.S. methodology is that it focuses on the inherent problems and nature of a forensic voice comparison. Other systems currently in place or being developed by private industry do not lend themselves to the police environment. Voice based security systems used for building entry, for example, rely on previously obtained voice samples from a cooperative subject. This type of task lends itself to pattern matching techniques when similar text is available. Additional advantages these systems have is that usually they are performed in controlled environments and again, the subject is cooperative. Police type voice comparisons generally will not have a cooperative subject and the samples could come from a variety of transmission media. In order to obtain an exact exemplar of the question call the investigator would be required to reveal to the suspect that he is being investigated.

Attempts at making a fully automated system for forensic purposes have failed in the past. Voice production is a very complex phenomenon. Unlike fingerprints which are static in nature, voice articulation is very dynamic. All speakers have their own intra-speaker variability which must be considered from a statistical point of view. The development of forensic black box systems attempting to analyze voices without any intervention of an operator is still far in the future. The difficulty stems from this type of system attempting to analyze a targeted voice which has not been screened for environmental or system contamination. The old adage applies, "Garbage in, garbage

out." If the voice samples are not representative of the speaker then analysis should cease.

C.A.V.I.S. is not a real time system. Post processing of the data is its luxury. The examiner/operator monitors and screens the data throughout the entire analysis process and is aware that some forensic cases will not lend themselves to analysis. It will become apparent to the police community that if a suspect makes an obvious attempt to disguise his voice or provides an inadequate amount of sample, that this is no different than the fingerprint examiner having no case to work because the suspect wore gloves or if the prints that were obtained were only partial or smudged.

2 OVERVIEW OF C.A.V.I.S.

2.1 Interactive Design

Recognizing that the C.A.V.I.S. System is a tool to be used by an examiner establishes the premise that the examiner and not the machine is in charge.

It has been proven through our experience as well as other reported studies on automatic speaker recognitions that some amount of human intervention should be retained to ensure adequate performance (Federicao et al, 1987; Chen and Lin, 1987).

With C.A.V.I.S. the interaction of the examiner begins with an aural assessment of the data available. Each of the following interactive steps are controlled and activated from a C.A.V.I.S. menu screen. Using an optical mouse, the examiner places a cursor over the desired function and presses a button on the mouse to begin that process.

The examiner must first determine if their is sufficient quality and quantity of speech available from each sample. Basically, the sample must be representative of the speaker and be within an acceptable signal to noise ratio. Comparison samples with dramatically different speaking modes should be avoided.

At present, disguised voice can be detected by the trained voice examiner whereas we do not have sufficient information to implement his knowledge into a computer algorithm. At the front end (before the computer process even begins), the operator must decide the degree of disguise. If it was determined to be excessive, then further analysis will be abandoned or at least he

will adjust the identification criterion properly to avoid erroneous results.

C.A.V.I.S. requires a minimum speech sample to consist of 10 seconds of voiced utterances. Presumably, this length will approach a phonetic balance. The examiner is provided with C.A.V.I.S. editing software to create a compressed speech sample. (This and other software will be detailed later.) The examiner calibrates the system and confirms whether proper digitizing of the sample has been performed.

We are aware of the popular and precisely defined cepstrum technique, originally invented by Noll(1967), and applied successfully for speaker verification research for commercial application by Furui(1981a,1981b), which is designed to provide estimates of the pitch period. This algorithm could have been a convenient tool to make our system more automated. But due to the wide windowing of this algorithm, the pitch detected are gross estimates, and does not capture the fine dynamic variations of acoustic events of each wavelet.

We concluded that the cepstral technique may not be applied to an on-going contextually unbounded speech process but, only to a short phrase unit, or a steady state phenomena, such as during a sustained vowel.

We did not want to stake our conviction that speaker-dependent characteristics information is entrapped within a single wavelet. Hence, we devised an "interactive-technique" to detect and define pitches. The pitch detection task was performed by an operatior making use of an optical mouse, acting

on the graphically displayed signal and assisted by immediate audio playback, when needed. Therefore, this technique is tedious, but ensures the highest possible accuracy.

C.A.V.I.S. speech parameters are derived from both time and frequency domain analysis techniques. In the time domain the examiner plays another role in confirming whether suspected samples have been automatically discarded by the software.

The remaining speech parameters can be automatically extracted and a master data file created for the sample, but again, the operator has the option to monitor the progress at every stage and halt the process if a malady should occur.

For his review, the examiner is presented with a graphical representation of the speech parameters derived from the speaker's available samples.

The examiner then submits the parameters representing the speech samples to an identification program. The C.A.V.I.S. "Proximity Index" finally establishes a rating of the similarity between the samples based upon the distribution of the stored general population.

2.2 Hardware Integration

Figure 2.2-1 is a photograph of the present state of hardware which makes up the C.A.V.I.S. work station. The evolution of the C.A.V.I.S. workstation was in three phases. At the beginning of the effort the research staff relied upon the three IBM AT microcomputers loaned by the U.S. Secret Service to become proficient in the C language programming environment. Due

to administrative complexities, it was not until well into the first year of the grant that equipment was in place to begin the recording and computer entry of the voice samples.

The second stage began when the recording and initial research hardware was in place. Figure 2.2-2 is a diagram showing the equipment configuration used to record the voice samples. For the project's first recording session the subjects entered the sound booth where they were recorded reading prepared texts and spontaneously describing projected slides.

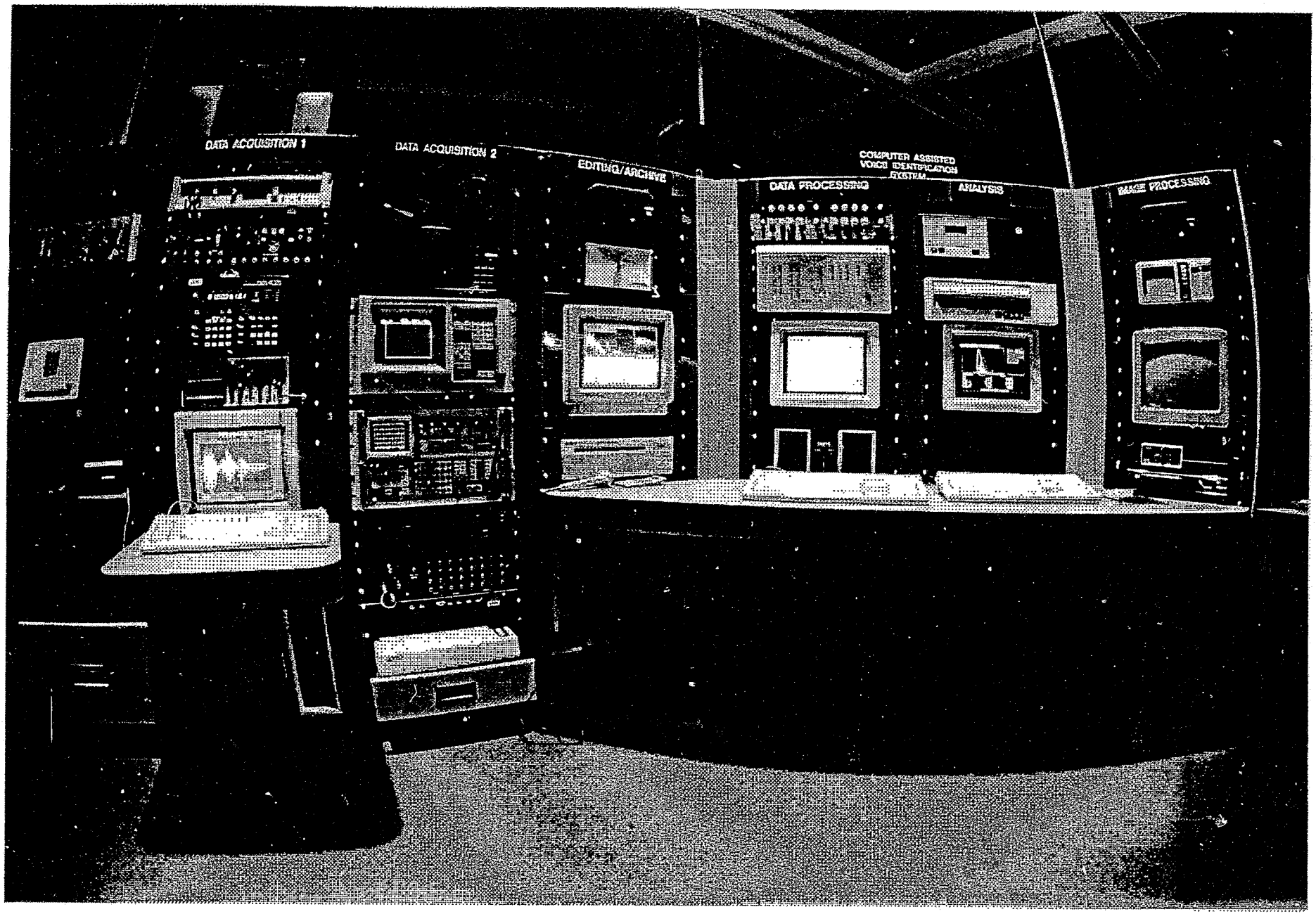


Figure 2.2–1 Photograph of the C.A.V.I.S. work station for forensic voice identification and acoustics analysis developed during this project.

A 1/4 inch Fostex model 80 eight track tape recorder was used to archive the hundreds of voice samples acquired. The recorder's time code based autolocating system was used extensively to automate the procedure. Each speaker was assigned an Ampex Grand Master 457 audio tape. A reference time code was prerecorded on channel eight and each sample for a given speaker started at a specific time on the tape. The researchers fabricated a dual tone pulse generating system. Once the recorder went into automatic record mode, the operator would press a button and a pulse was placed on track seven. A delayed pulse was then heard by the speaker to signal him to begin. The pulse on track seven was later used to automatically start the digitizing process for each sample. Track five contained a narration channel for the operator's use.

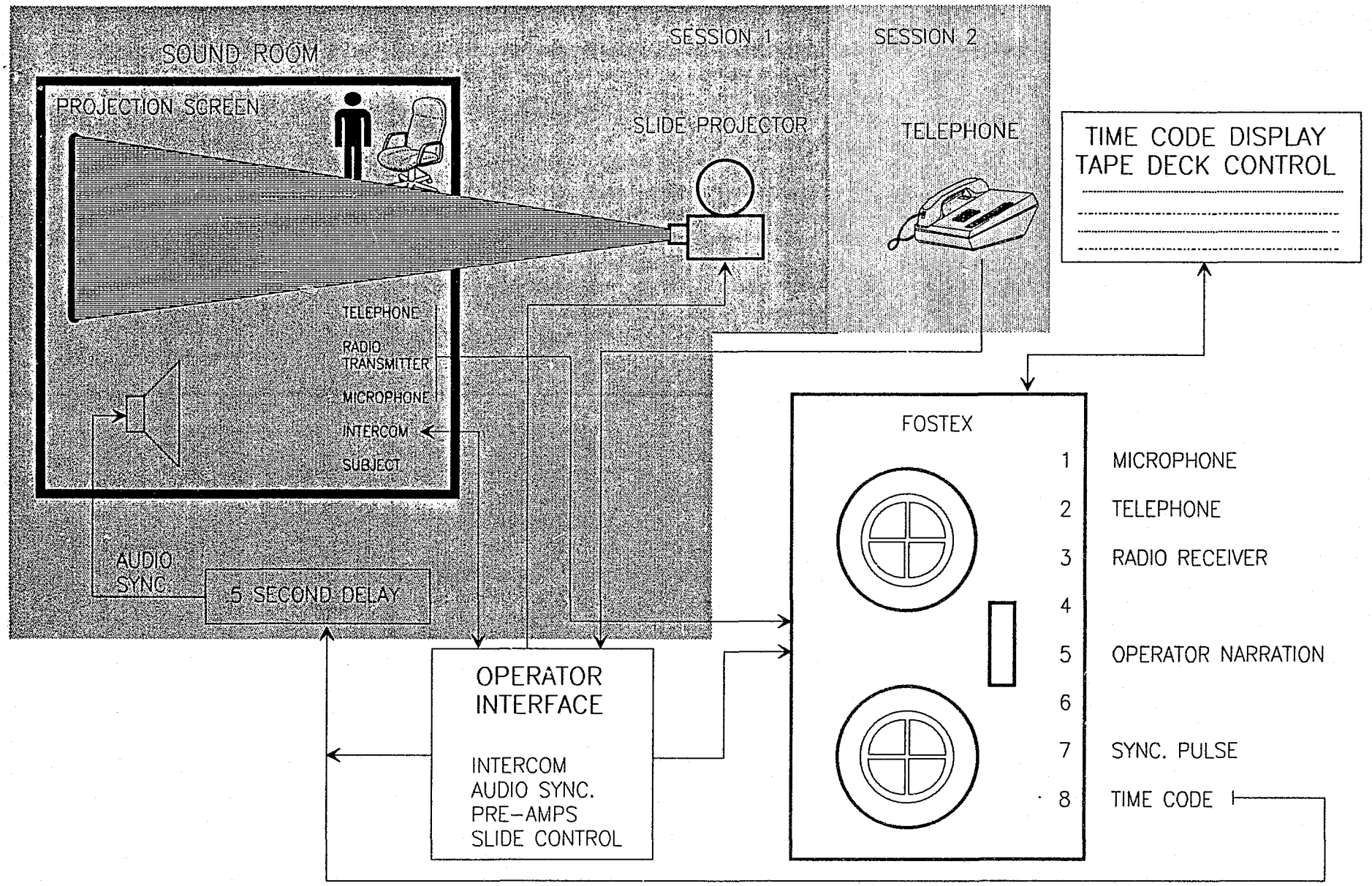


Figure 2.2-2 A diagram showing the equipment configuration used to record the voice samples.

In phase 1, each speaker was simultaneously recorded over three transmission media. The goal was to acquire data which could later be used to verify the effectiveness of the transmission line cancellation algorithm. Track one recorded the output of a Bruel & Kjaer type 2230 sound level meter equipped with a 4155 type 1/2 inch microphone. Track two recorded the intercepted outside phone line transmission and track three recorded a police body wire transmission of the sample.

Subsequent recording sessions in phase 2 deleted the use of the sound level meter and body wire and the subjects telephoned the laboratory from varying remote locations. The resulting analog tapes of each session were cataloged and color coded throughout each phase of the procedure.

Once the analog tapes were acquired, the next step was to enter the voice samples into the computer data base. A Compaq 286 microcomputer equipped with a 70 megabyte hard disk became the heart of the digitizing system. To archive the massive amount of digital data, an Optimum 12 inch optical laser storage device was interfaced to the system. The laser has two Giga bytes of storage capacity per platter. The computer is also equipped with a Tecmar 60 megabyte tape backup system which was used to backup the digital data.

The acquisition system utilizes a Data Translation model 2801A 12 bit digitizing board to input and playback the audio samples. The researchers designed and fabricated an interfacing panel which allowed for the control of data levels, channel selection and auto starting of the digitizer. Anti-aliasing

filters were also incorporated into the interface. Additional equipment making up the data acquisition station includes a Spectral Dynamics model 375 narrow band analyzer with a model 348 waterfall display, a Digital Audio Corporation model PDF2048 programmable digital filter, a Hitachi 40 MHz oscilloscope, a BGW Systems model 85 audio amplifier, and monitor speakers.

A National Instruments GPIB controller board was used to computer interface the narrow band analyzer and the programmable digital filter. The usage of these items will be detailed in chapter 3. The above equipment is housed in three 19 inch Stantron equipment racks.

When the grant was extended for an additional two years in October of 1987, additional equipment was acquired to configure the C.A.V.I.S. workstation to its present state. Three computers can now access the archival data stored on either the laser or a Spark removable and fixed hard disk system. United Systems Company of Tustin, California fabricated and interfaced the computers through a custom hardware switch.

A Compaq 386 microcomputer performs the speech parameter extraction and a Compaq 386/20 concentrates on the voice identification and verification tasks.

A portion of the C.A.V.I.S. methodology makes use of FFT analysis. The calculation of the spectra has evolved through three stages; software calculation, interfacing to the SD375 Real time Analyzer, and finally the use of an Ariel DSP16 digital signal processing board. The speed enhancement using the

processing board greatly aided the research effort.

A conveniently located Kay Elemetrics Sonagraph model 5500-1 enables the real time spectrographic review of the voice samples. While editing a sample the voice examiner now has a familiar display to verify its quality and duration.

The final component of the workstation is a high-resolution color graphics display system. A Chorus 16 bit Frame Grabber board is housed in the Compaq 386/20 to drive the monitor. In addition to its high pixel resolution and continuous color data display capabilities, it can be used to capture and archive suspect photographs through an outboard video camera. This capability also allows for its use in tape authentication and image enhancement work. The work station now incorporates 7 full equipment racks.

2.3 Software

As a part of our "man-machine" interactive system of voice identification, we recognized, among other sophisticated mathematical computation requirements, two essential capabilities which must be realized into the system: The first is related to the interfacing of computers, internal/external peripherals and the various types of memory storage media. The second is related to the interfacing of analytical computation, graphic display, and the digital to analog conversion process.

In order to attain the above capabilities, the researchers at the onset of the project chose to program the C.A.V.I.S. system using the "C" programming language. A review of popular

signal analysis menu type programs did not lend themselves to the requirements of C.A.V.I.S. This decision prompted the team to become proficient in all aspects of "C". The learning curve at first was a barrier, but the resultant flexibility, control and customizing ability attained proved that the right decision had been made.

After working with other products, the Microsoft C Optimizing Compiler was chosen as the C.A.V.I.S. standard. Additional software library routines are used and include Media Cybernetics "Halo 88" graphics, Greenleaf Functions, and specific driver software supplied with peripheral interfacing boards. Hundreds of programs were written by the researchers for interfacing hardware, for acquisition and analysis of data, and for development of meaningful graphic displays on the monitor during the identification process. A list of major program names and the brief functional descriptions is provided in Appendix B.

2.4 Networking

The C.A.V.I.S. system has two levels of networking capabilities. Within the workstation environment the computer systems are capable of sharing databases and programs through a LAN (Local Area Network) card. On a broader range, the researchers tested the feasibility to network the system to other police agencies across the country. Lt. Lonnie Smrkovski of the Michigan State Police Department established an abbreviated workstation at his audio laboratory. Lt. Smrkovski has been a noted voice identification examiner for many years and has a deep

appreciation for our effort.

Using a "Closeup" communication software package, the two laboratories were able to link and share data as well as control analysis hardware. With this capability an agency equipped with an upgraded IBM PC compatible computer could send data to our laboratory via phone lines and analysis results could be returned the same day. Or, a C.A.V.I.S. examiner could travel to a location, assist in getting voice samples, and remotely access the C.A.V.I.S. system using a portable computer.

2.5 Scope

The remainder of this report will concentrate on the procedural details and description of the parameters employed by C.A.V.I.S.. The intention of this report is to offer to the reader a description of the research effort and a summary of its accomplishments. It is not intended to be used as a procedural manual.

3 METHODOLOGY

3.1 Data Acquisition Procedures

3.1.1 Database Size And Type Of Speech

Our experimental voice data used in the first two years of the project was collected from 50 white male speakers recorded through several different transmission systems (telephone, microphone, and RF body transmitter). Due to time restraints in the first period of the grant, only 21 of these speakers were fully processed. A forensic telephone path was created utilizing outside local lines. For each speaker a new telephone connection was established by re-dialing. While in the laboratory's sound booth, each speaker produced two sets of ten speech samples, each 30 seconds long and of different text. This was done by reading randomly selected text material, and also by speaking spontaneously while viewing projected slide pictures.

During phase II, new elements were added. The speaker population was increased to 150 males. The type of voice data was changed from read text to spontaneously produced text. The recording interval was changed from one session only to two sessions separated by a minimum of two months. Every speaker was recorded through an outside telephone line from his work or home to our laboratory telephone set. No restrictions were imposed as to which telephone line each speaker used. The speaker offered spontaneous text while looking at prepared pictures for the first session, and entirely unsolicited text for the second session. For the second session, it was left to each speaker to choose his own topics (five different topics each lasting 30 seconds).

These new elements were added to make our experimental voice data approach a realistic forensic situation. It naturally yielded a more complex and an unstable type of voice data and resulted in the overall lowering of the system performance.

Some phenomena observed can be described as follows. A certain group of speakers were always stable. This was indicated by the small intra-speaker variation measured by parameters from both the frequency and time domains from several samples within and across the two recording sessions. Another group of speakers were shown stable, but only within a single recording session. The smaller group of speakers did not exhibit stability within either session.

The approach taken to handle these variations will be discussed in a later section.

3.1.2 Duration Of Samples

In the initial project proposal it was estimated that a minimum length of ten seconds of compressed speech would be required to constitute a sufficient sample. This minimum length would help insure that the voiced sample would approach a phonetic balance. The recording length of each sample throughout the project was thirty seconds. In the first phase of the project the samples analyzed came from subjects who were reading prepared text.

Once the pauses were removed from these samples, the remaining voiced speech did provide the research staff with compressed samples exceeding the ten second minimum. The samples

from the second phase, however, are samples obtained from thirty seconds of spontaneous speech. As a consequence, the percentage of voiced speech in these thirty second samples as compared to the samples obtained from reading prepared text is appreciably reduced. This reduction is likely due to the subject's thought process inherent to spontaneous speech.

3.1.3 Calibration

The main goal of the calibration procedure is to insure that the signal level of the sample attains sufficient intensity to allow parameters to be extracted, yet safeguard it against 'over driving' the digitizing process. Absolute level calibration in dB was not required because in the forensic environment there are no references, only relative levels. In the first recording session a 94 dB pistonphone calibration was used on the sound level meter for system checking.

In the actual use of C.A.V.I.S. a typical scenario would be to have a questioned call presented on a standard cassette tape. This call is transferred to track one of the Fostex recorder. The Fostex tape would have already been pre-recorded with a time code on track eight. During the transfer, the operator monitors the playback of the original while being attentive to signal quality. The signal level going onto the Fostex should not exceed "0" VU. We took extreme care to obtain optimum audio input levels during the digitization process. A calibration tape of 1000 Hz at "0" VU is played back from the Fostex into the

interface panel of the acquisition system. A program called "Cklevel" is then called and three seconds of this signal is digitized and displayed on the computer monitor in EGA color graphics. The examiner adjusts the input level until it reads full scale on the computer monitor. The oscilloscope level is adjusted to match the computer screen. Once this is completed the real time oscilloscope display represents the level at which the digitizer will see the signal. The examiner while monitoring the oscilloscope then maximizes the level of the sample taking care not to over drive it.

3.1.4 Digitization

The digitizing board used by C.A.V.I.S. allows for 12 bit quantization. This provides sufficient signal to noise ratios when input levels are properly adjusted. The rate of digitizing is 10,240 samples per second. Several factors motivated the use of this sampling rate. The voice frequency range in the forensic situation is limited generally by the response of the telephone line. Dramatic roll off of the signal occurs at approximately 300 and again at 3,000 Hz.. For this reason when using the SD375 analyzer the frequency range chosen was 0 - 4,000 Hz. We matched its sampling rate of 10,240 Hz to maintain system compatibility. Uniquely, this also mathematically sets the FFT analysis bands at an even 10 Hz bandwidth yielding 400 bands within the 4,000 Hz range. A crystal controlled external clock was fabricated to generate this sampling rate which drives the digitizing board. An alternate sampling rate of 20,480 Hz. is also available to the

operator. The maximum rate of the board is 27,000 Hz. For transient analysis work, the SD375 can be set for faster rates. The researchers have also developed a selectable overlapping FFT analysis routine for the DSP board which greatly expands the detail of time varying events.

The maximum duration of a sample is only limited by the amount of available memory on the storage device. Customized software was developed to allow continuous storage of the signal to the capacity of the hard disk.

This capability is also extended to the full capacity of the laser platter. Archived sound files can be played back directly from the laser.

3.2 Voice Data Pre-processings

3.2.1 Determination Of Pre-emphasis Filter Shape

As mentioned, the analysis of a forensic recording which stems from a voice sample made over a transmitter or telephone transmission is plagued with an unknown response curve associated with the media.

The application of a pre-emphasis filter shape was improvised primarily because of the above mentioned adverse effects of the transmission system. As it will be discussed in 3.3.4.3 we use a special type of spectra to represent a speaker in the frequency domain, namely, an intensity deviation spectrum (IDS) for the purpose of neutralizing this influence of the transmission and recording media. An algorithm developed to compute the IDS has been proven to work excellently if energy

associated with the bandwidth is sufficient to generate a measurable amount of deviation or variation around a central value across a set of FFT short-term spectra. When there is not sufficient energy, we experienced unpredictable results. A pre-emphasis procedure which we developed alleviates this situation.

There is a secondary effect of this pre-emphasis technique, which is rather significant in the overall aspect of the project for voice identification. At the perceptual level, i.e., when we listened to the unprocessed speech signals, it was evident that the quality of voice commonly associated with a high fidelity microphone differed from that of the voice usually associated with a typical commercial telephone transmission. This difference was to the degree that the voice recorded through the telephone sounded different from the one recorded through the microphone.

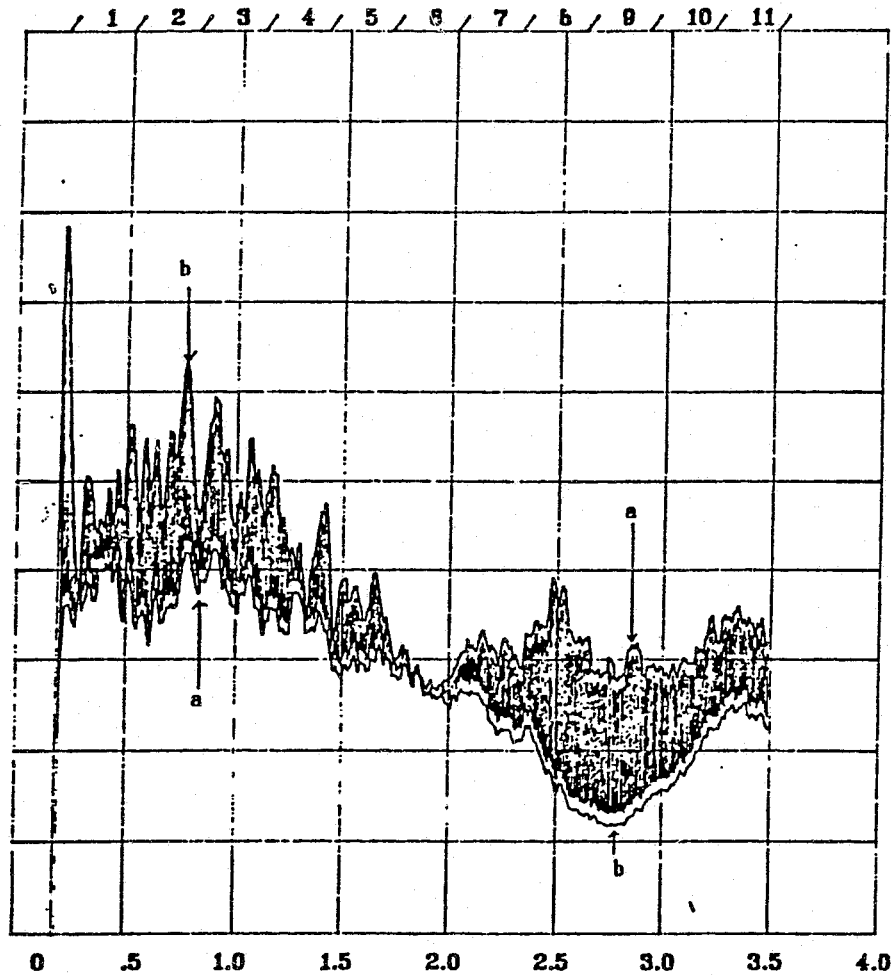
By applying the above pre-emphasis filtering technique, however, the difference of the voice quality due to the transmission systems is diminished. This aspect is extremely important relative to the listening process during the forensic voice identification process.

Figure 3.2-1 illustrates the extreme effect of different transmission response characteristics where the long term spectrum of the same utterance was computed from a telephone sample versus a microphone sample.

Figure 3.2-2 illustrates the effect of the pre-emphasis filtering combined with the effect of the IDS algorithm. Note

that two IDS spectra are generated from the same utterance, one being transmitted through microphone and the other through a telephone are closely matched.

C.A.V.I.S. PROGRAM AUTODISP: Thursday 22-Jan-87 - 8:32:09pm



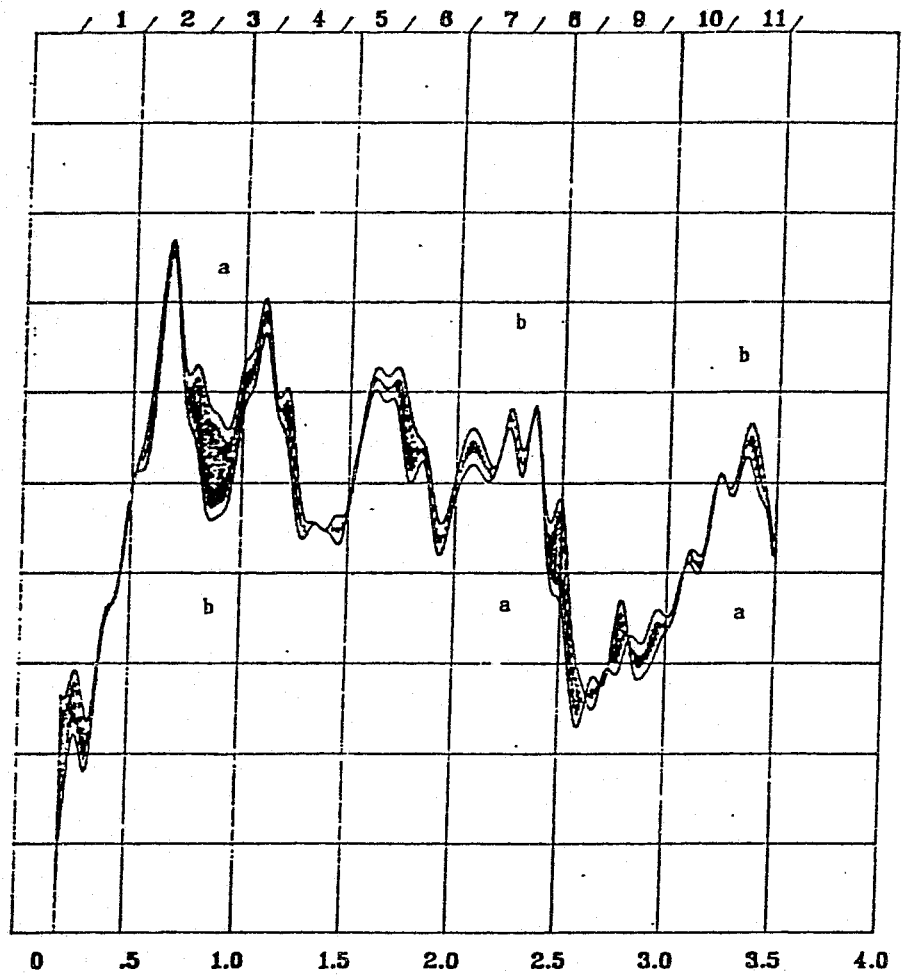
File1: 13m1.lts

File2: 13t1.lts

dif=	11244.29	(.129)
sim=	75846.34	(.871)
tot=	87090.63	(1.000)

Band #	Intv(Hz) Low	High	Corr Coef
1	200	500	-.474
2	500	800	.861
3	800	1100	.878
4	1100	1400	.193
5	1400	1700	.817
6	1700	2000	.890
7	2000	2300	-0.066
8	2300	2600	-.556
9	2600	2900	.745
10	2900	3200	.617
11	3200	3500	.577

Figure 3.2-1 Graphic display illustrating the influence of two transmission systems upon the resulting average power spectra generated from the same utterance. The two spectra were generated from a 10 second long speech samples of the same text recorded by: (a) a microphone and (b) a telephone line. Note the difference in the shaded portion formed by the two spectra.



File1: 13m1.fd3

File2: 13t1.fd3

dif=	3565.10	(0.033)
sim=	104417.05	(.967)
tot=	107982.19	(1.000)

Band #	Intv(Hz) Low	High	Corr Coef
1	200	500	.943
2	500	800	.972
3	800	1100	.936
4	1100	1400	.962
5	1400	1700	.980
6	1700	2000	.949
7	2000	2300	.799
8	2300	2600	.958
9	2600	2900	.360
10	2900	3200	.964
11	3200	3500	.829

Figure 3.2-2 Graphic display illustrating the effects of the IDS spectra in eliminating the influence of two transmission systems upon the resulting average power spectra generated from the same utterance. The two spectra were generated from a 10 second long speech samples of the same text recorded by: (a) a microphone and (b) a telephone line. Note the observed difference between the two spectra. is smaller than the distortion seen in Figure 3.2-1.

Figure 3.3-1 is a schematic diagram of the procedures to determine the individualized filter shape for each text sample of a speaker. A linear phase filter (Digital Audio Corporation, Model PDF 2048) is incorporated into the process to help overcome the attenuation present in the recorded sample. The shape of the filter applied to the data is unique to each sample and determined by an inverted long-term spectrum of the data which is computed by the examiner's operation of the SD375. The long term spectra of the sample is extracted from the SD375 over a GPIB bus. The program "SD_PDF" inverts, normalizes and smooths the spectral shape. The program continues on to compute, from the long-term averaged spectrum, 512 filter convolution coefficients which are sent over the same bus to program the filter. The effect of attenuation balancing is limited to 30 dB.

The benefit of utilizing a filter is demonstrated in Figure 3.4(a-b). It shows a "waterfall" frequency display of successive spectra from a single sample. When viewing the display's upper half, where filtering was used, one can see the improved condition of the spectral information.

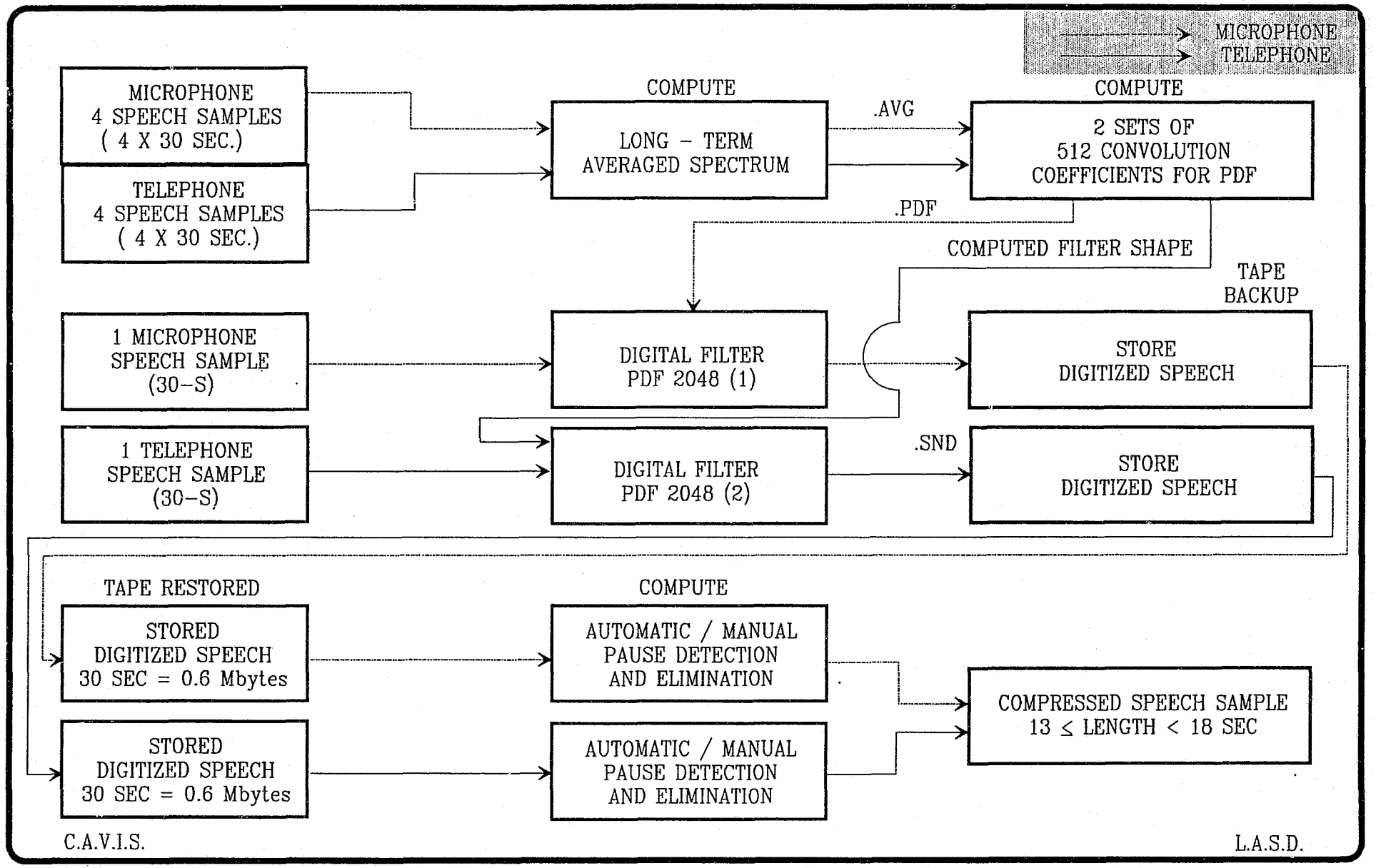


Figure 3.3-1 A schematic diagram of procedures to determine the individualized pre-emphasis filter shape for each sample.

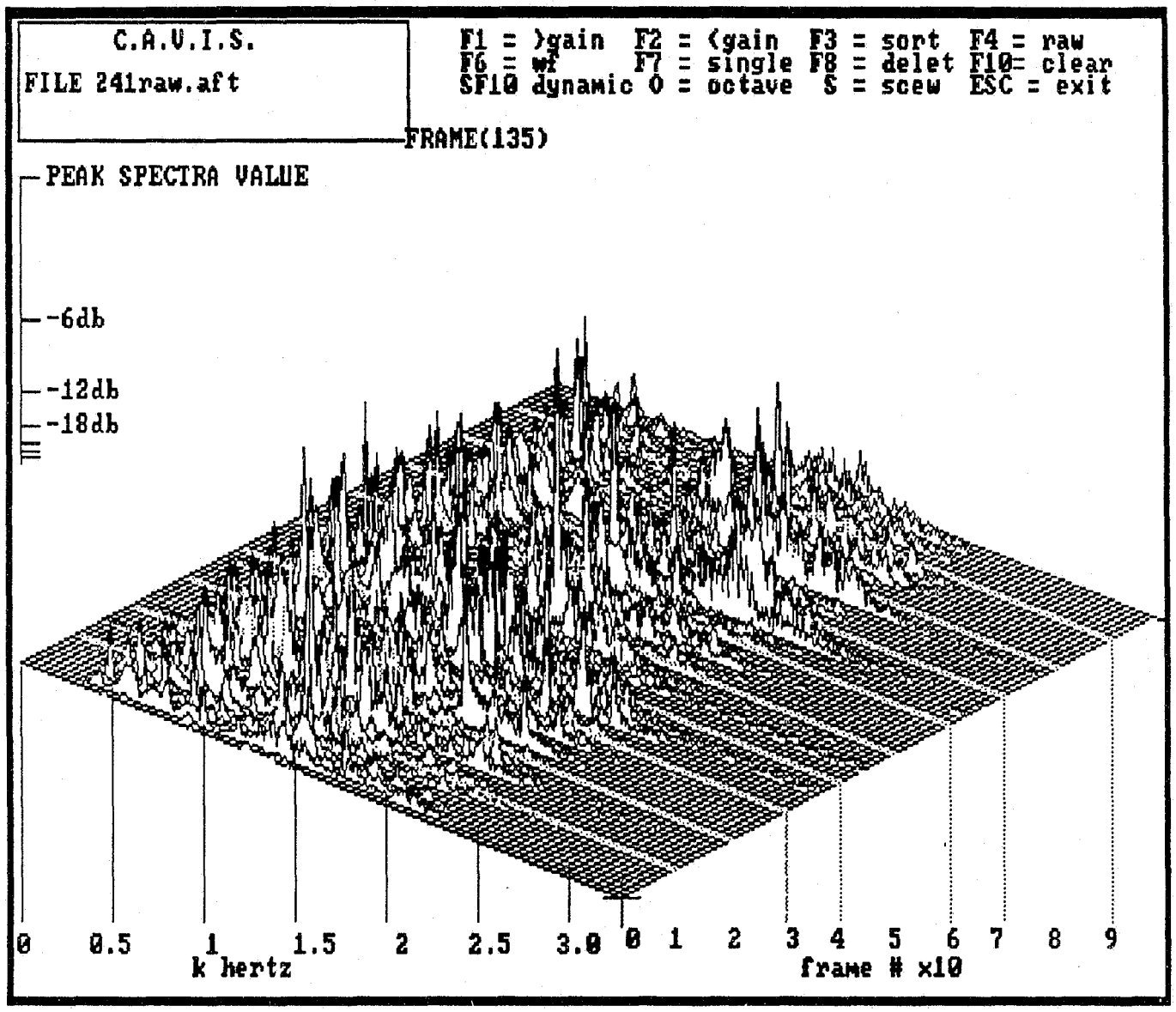


Figure 3.4(a) Waterfall display of successive FFT frames before the application of the individualized filter shape.

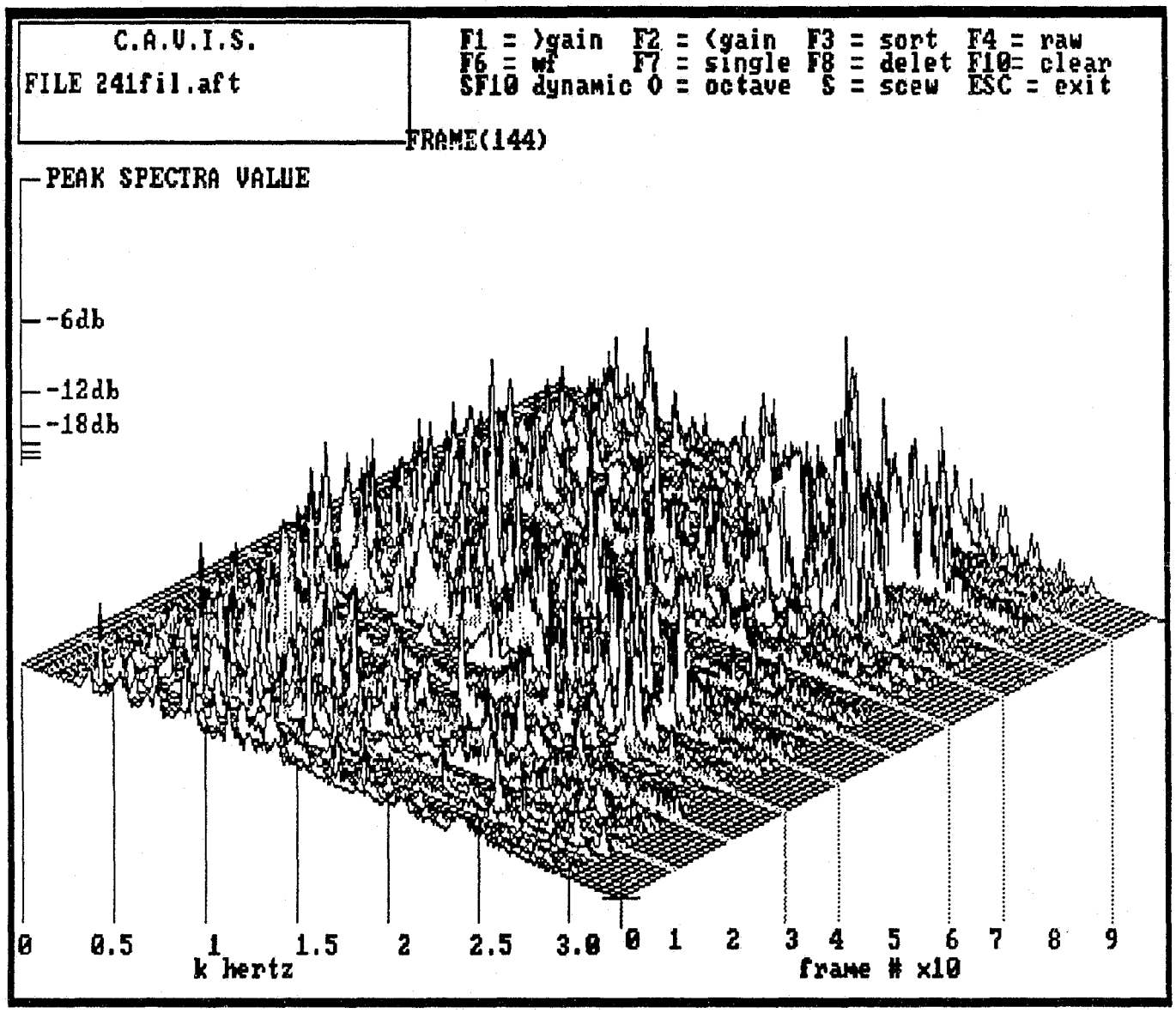


Figure 3.4(b) Waterfall display of successive FFT frames after the application of the individualized filter shape.

3.2.2 Sound File Creation And Storage

The analysis of each voice sample will result in the creation of a number of associated files that begin to reduce the sample to a single parameter file which characterizes the sample. The labeling of each file is found in the file's extension name. The first file in the process is the raw digitized sound file that bears the extension ".snd".

The name of the file is restricted to five characters followed by a 01, 02, 03, 04, 05 depending upon which sample of the event is being processed (example "KNOWN01.SND"). The C.A.V.I.S. mouse driven menu screen is capable of handling two speakers (unknown, known) each having sufficient data to create five samples. The available samples will be used to determine the degree of intra-speaker variability.

Using the time code of the Fostex recorder, the examiner estimates the areas of the recording for each sample that will yield a minimum of ten seconds of voiced speech for the targeted voice. Whatever duration of the recording is necessary to achieve this length will be digitized as one sample.

The analog signal is passed through the pre-emphasis filter, input levels are adjusted and the operator arms the digitizer for manual start. The file is played back through monitor speakers and verified for duration and quality once the digitizing process is completed.

The next step is to remove the areas of the file that do not contain speech. During the research phase of the project where

only one voice was present on the recording, a program was written that scanned and automatically removed the pauses. Figure 3.5(a-b) illustrate the effectiveness of this program. Generally, a subject's conversational speech will contain approximately 40% pause. Our program called "DELPAUSE" uses several levels of pause detection. An initial scanning of the file determines a threshold to be used for detecting the noise floor. The duration of what constitutes a voiced segment as well as a minimum pause is selectable. The uniqueness of the program is in the joining of the segments. A smooth zero crossing is automatically determined. A compressed sound file can be generated at any time from the raw sound file using the output of the program that lists the beginning and ending points of each voiced segment.

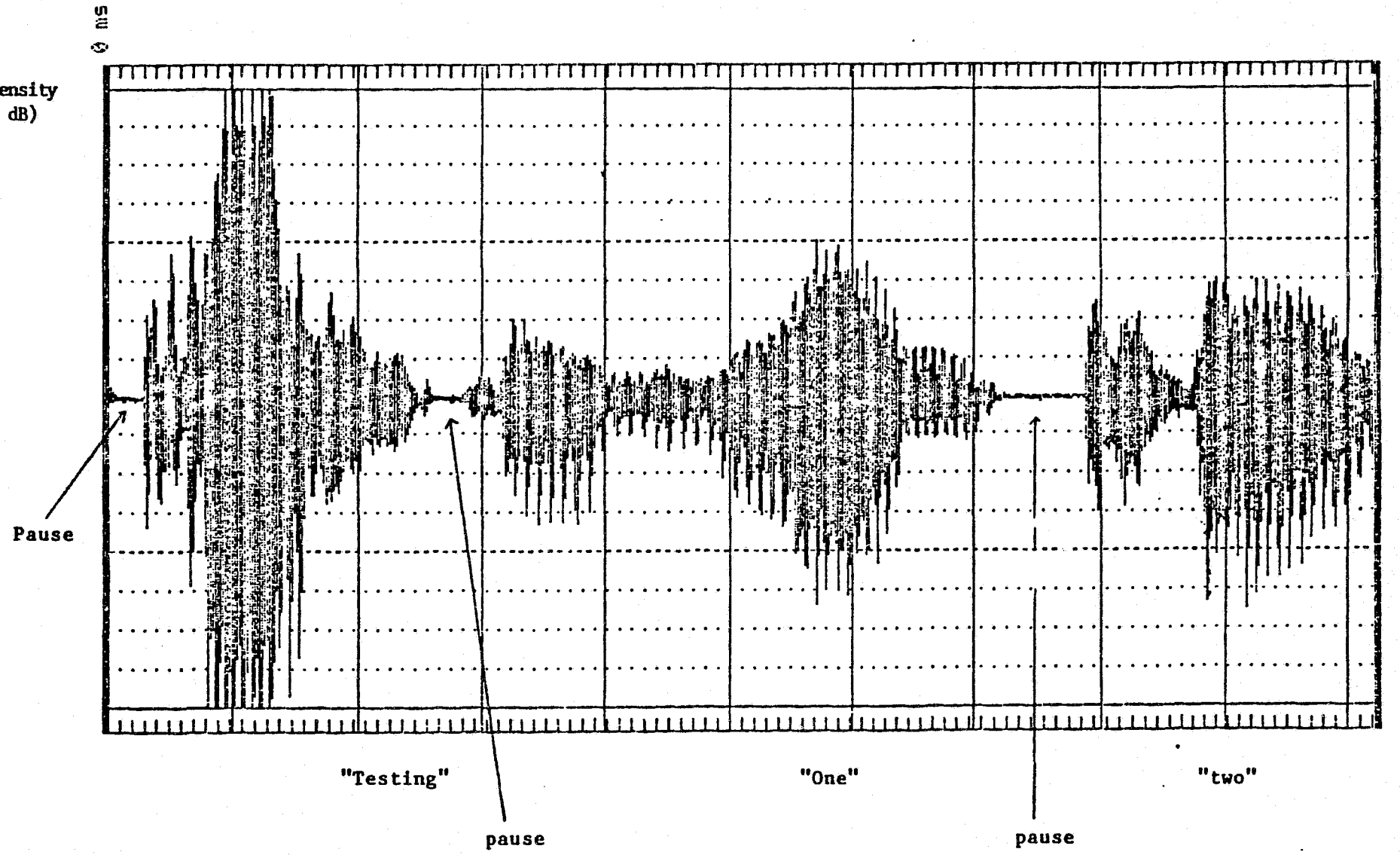


Figure 3.5(a) Graphics of speech signals with pauses.

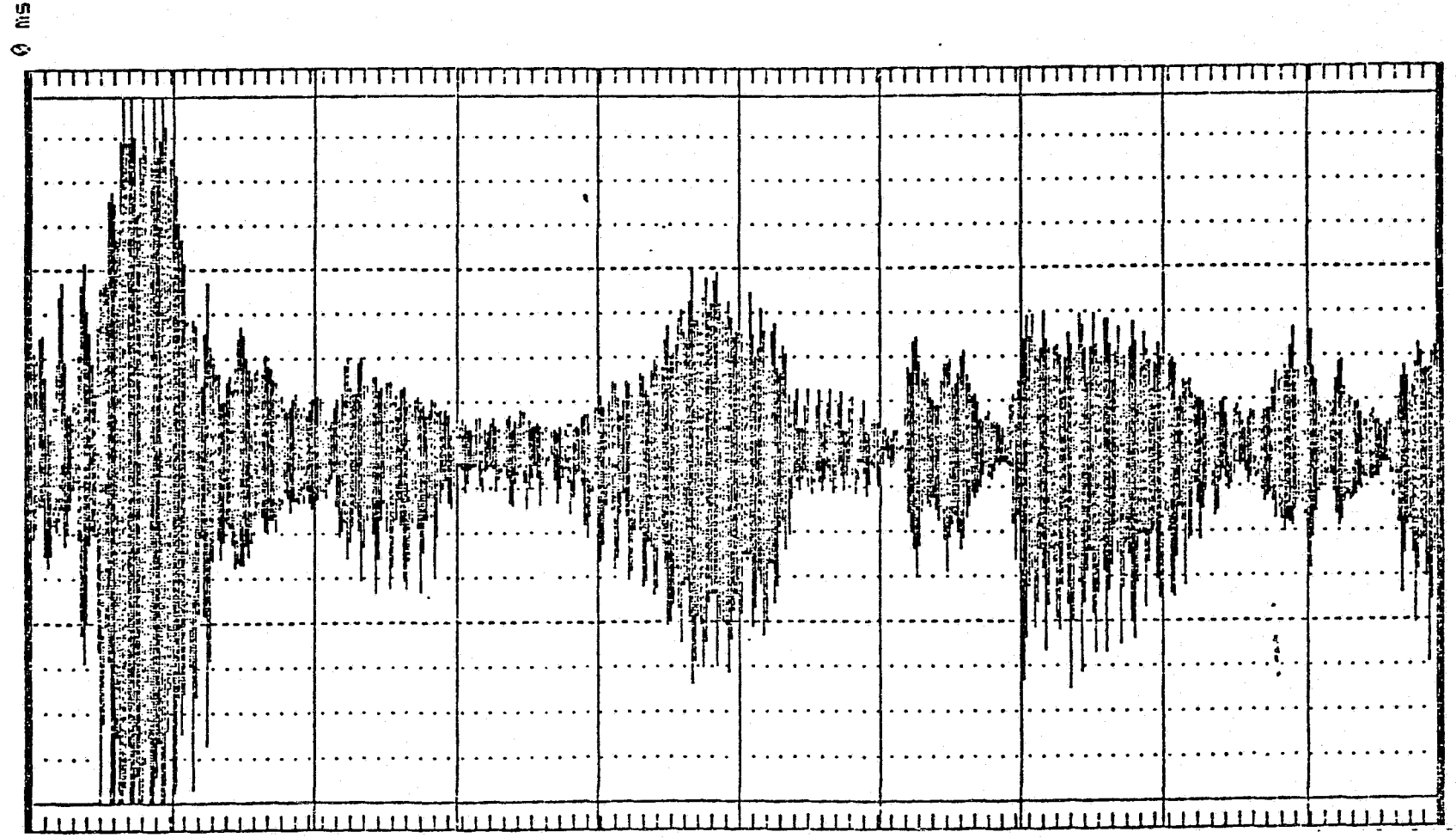


Figure 3.5(b) Graphics of speech signals with pauses removed.

The "EDIT1024" interactive editing program was written to allow the examiner to verify every segment of the sample. Figure 3.6 depicts a typical screen ready for editing. Using the optical mouse the segment to be verified is surrounding by a cursor which expands in even 1024 data sample increments. The selected area can be aurally reviewed by pressing a mouse button. If the segment is desired, a function key is pressed and the beginning and ending points of the segment are stored. The 1024 point multiple segment size allows for ideal fitting of the 1024 point FFT routine used later in the analysis process.

The program "MAKEMCP" uses the stored information from the "EDIT1024" program to generate a compressed sound file bearing the extension ".mcp". It is this file that contains the verified and compressed speech data of the targeted voice.

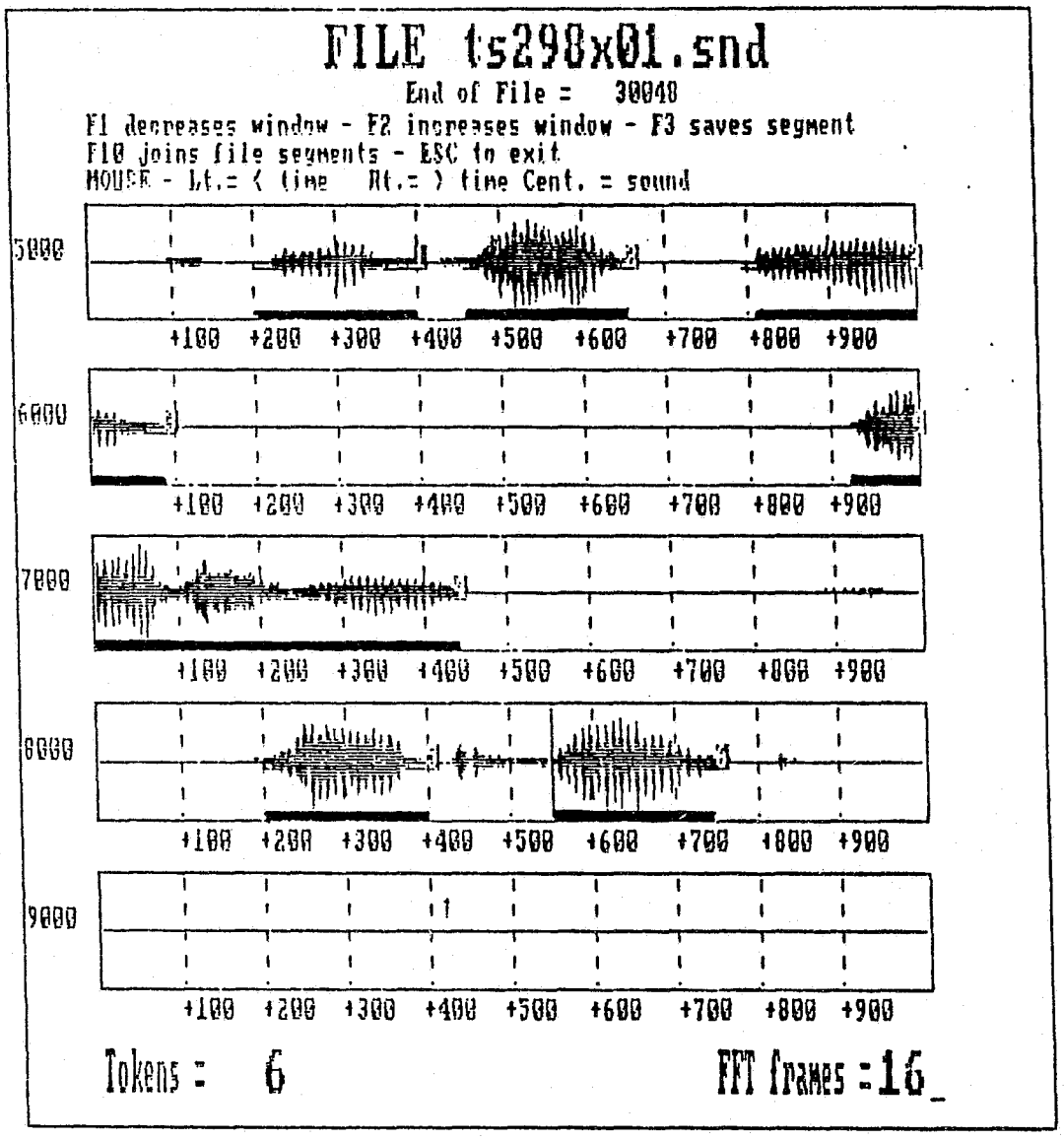


Figure 3.6 Graphic display of computer screen during interactive editing of a sound file.

3.3 Voice Parameter Extractions

3.3.1 Voice Parameters

Ideally, a parameter would vary only a small amount when measured from a single speaker (case where the intra-speaker variability is small), and greater variance would be expected when measured between or among different speakers (case where the inter-speaker variability is large). It has been commonly recognized by many researchers, that one of the crucial keys to successful speaker identification is to search the speech parameters that provide smaller intra-speaker variability, but greater inter-speaker variability. Our original enthusiasm indeed, was centered around the search of such parameters with high inter-speaker variability. During the earlier phase of this project, the original parameter set was found sufficient to achieve high identification rates (reported in 1988 IEEE-ICASSP Proceedings, and also in the previous quarterly reports), but during the later phase when we increased the speaker size and recording sessions, the identification rate was degraded. Such degradation led us to implement further refined schemes in the process of defining parameters.

3.3.2 Time Domain Parameters

3.3.2.1 Parameter Extraction Procedures

Many studies on speaker identification focused predominantly on the extraction of spectral information from the frequency domain by the use of Fourier spectral analysis (Paul et al., Bunge, Markel et al., Federico et al., and Warkentyne et al.).

This type of information is more effective than that extracted from the time domain such as the average pitch and other varieties extracted therefrom. We are in total agreement with the understanding that spectral information is a good speaker discriminator. However, we believe that a computer-based voice identification system based solely on spectral information would fall short in accuracy in the forensic environment.

In an ongoing speech there are many acoustic events that can be resolved only in the time domain: characteristics such as intonation, inflection of pitch, stress, fluctuation of tonal quality, etc. We designed our system to study the detailed microscopic events at the level of single glottal waves and the dynamic phenomena of these events.

Extraction of speech parameters from the time domain begins with the interactive targeting of each glottal wavelet in the sample. In normal speech production, vocal folds are used to modulate an air stream. Such modulation is called one cycle of vocal fold vibration. For an average male adult, that vibration occurs somewhere around 125 times per second. In this report, the term 'wavelet' refers to the waveform that occurs in one of these cycles.

It was found through experimentation that speaker's wavelets have characteristic attributes. Beyond the expected variance due to phoneme production, some speakers consistently generate nicely formed peaks and decay characteristics. Others, perhaps those with raspy or hoarse voices, have poorly defined peaks or second harmonic peaks rising higher than the fundamental (See Figure

3.7-1). This particular behavior discouraged our own efforts to develop an automatic peak (pitch) detection algorithm and we believe also plagues the efforts of other approaches.

The next program called "SMORECT" is run on the compressed sound file to help overcome the difficulty in targeting the glottal peaks. The compressed sound file is rectified and the intensities are squared to accentuate the peaks. This program generates a file with the extension name ".srt".

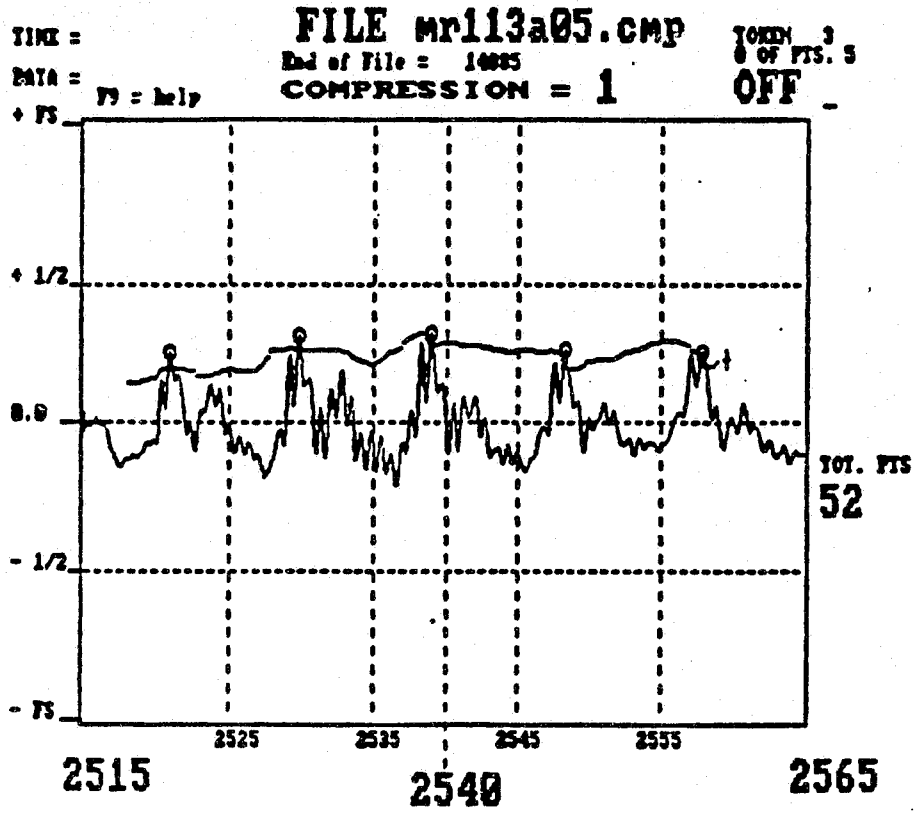
Using the mouse controlled cursor of the "PICKSRT" program, the examiner is now ready to target the peaks. The program loads the smoothed ".srt" file and the ".sts" file from "EDIT1024" and places markers over the data on the screen indicating to the examiner where pauses in the original file were removed. Using the mouse, the trained examiner then draws a threshold line just below the peaks of adjacent wavelets while taking care not to cross over a pause marker (See Figure 3.7-1). The location of each peak becomes part of a stored token in the ".jit" file.

A glottal wavelet is a pulse phenomenon with a fast rise time and slower decay time. Using the established peak of each wavelet, the researchers tried various schemes to automatically locate the beginning and ending point of each pulse. The varying noise floor conditions between samples made these efforts fail because exactness and uniformity could not be maintained. It was therefore decided to define a wavelet as an event beginning at one peak and ending one data point prior to the next successive peak.

The application of this newly defined wavelet added greatly to the performance of the speech parameters. The program "WAVELET" is applied next and generates several files containing computed measurements of each wavelet which ultimately yield the time domain parameters.

At this point, a discussion on how these parameters are treated and how the wavelets are normalized will aid the reader in understanding the parameters.

(a)



(b)

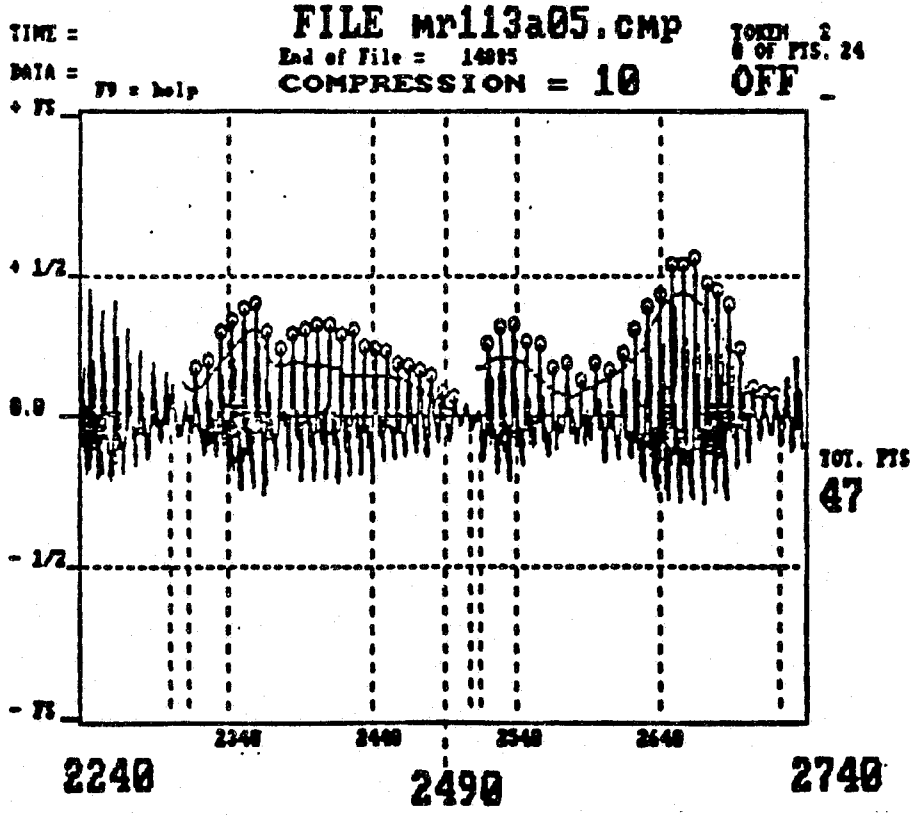


Figure 3.7-1 Graphic display showing the interactive peak detection. (a) is done at the maximum resolution possible which displays every data point sample at 10240 Hz sampling rate. (b) is done at ten time compression, yet maintaining the maximum data point out of a group of ten data points at the same sampling rate.

Figure 3.7-2 exemplifies an output of "WAVELET" which was mentioned earlier. Figure 3.7-3 is a graphic display showing the intermediate time domain data extracted from a set of wavelets. The vertical bars on the display indicate the boundaries of individual tokens targeted by the examiner. Plotted are the computed pitch contours with corresponding plots of peak intensity, total wavelet intensity and auto-correlation values of successive wavelets. These parameters are defined below.

Total Wavelet Intensity

From each speech sample, we segment a series of wavelets. And from each wavelet, its sum of energy is computed. Eventually we have a set of data, each datum being expressed as a total sum of energy of the wavelet intensity. Then, the statistical distribution of this set of data is computed by using the extreme value statistics, which will be described in section 3.3.3.

Variation of Total Wavelet Intensity

The averaged variation of the total wavelet intensity is computed from the same data set mentioned in the previous section, and it is constructed to reflect the fluctuation seen in wavelet to wavelet, or cycle to cycle phenomena of continuous speech production. It's computational expression is given below.

$$\Delta A_i = \left| A_{i+1} - A_i \right|$$

where, A_{i+1} denotes the total intensity value of the i+1 th wavelet, and A_i, of the i th wavelet.

Pitch

Pitch is the reciprocal of period usually expressed in seconds, and used synonymously with fundamental frequency (f_0) in Hz. An f_0 in our process is specifically defined:

$$f_0 = \frac{1}{\Delta T} = \frac{1}{T_{i+1} - T_i}$$

where,

$$\Delta T = \frac{\text{Number of data points between } i^{th} \text{ and } i+1^{th} \text{ peaks}}{\text{Digitizing Sampling Rate (10240 Hz)}}$$

Average Energy Distribution

This parameter is related to the average waveform of successive wavelets. We arbitrarily chose a waveform length to be 256 data long. Each wavelet excised was then adjusted in length by filling in the number of data where prefixed data length was greater than that of a given wavelet. Intensities were also normalized so that all the peaks would start at the same level. These adjusted (warped) data were superimposed and the average values of the data across a set of wavelets were computed, which provided a final form of the averaged wavelet.

Auto-Correlation of Wavelets

Each wavelet excised successively from an ongoing token contains varying wave shapes and also varying numbers of data. The amount of such variations was measured by the use of product moment correlation coefficients. Such a measurement can be considered as a 'jitter' value which is an index of deviation of

one particular phenomenon, or as 'stability' if we are interested in knowing how stable a person's glottal activity is. This parameter, a set of successive correlation coefficients, is computed by

$$r_i = \frac{\text{Sample Covariance}}{S_{xj} S_{yj}}$$

$$= \frac{\frac{\sum_{j-N\cdot(i-1)}^{j-N\cdot i} x_{ij} y_{ij}}{N} - M_{xj} M_{yj}}{S_{xj} S_{yj}}$$

where the 'x' contains a set of n data taken from proceeding wavelet (i-1), and the 'y' contains a set of n data taken from the following wavelet (i).

The measurement appears to be affected primarily by pitch change and is therefore, a good representation of jitter which is the deviation in pitch between successive wavelets.

Average Smoothed Wavelet Shape

The averaged wavelet shape revealed significant speaker distinctive information when measured between data point 30 and 120 within a total of 256 points. By using an eyeball method, we decided to partition the segment into three parts, each part being composed of 30 data points that formed a smoothed decay shape of a wavelet. This program also generates two additional data files which allow the examiner to view each isolated wavelet in its smoothed form as well as having its

energy distribution sorted in descending order. The sorted energy distribution has shown to have speaker dependent properties. After working with the data it was found that speakers that were easily targeted with the PICKSRT program had steep energy distribution curves. Or basically, when they produced their wavelets they exerted most of the energy at the beginning of the pulse. Others generated wavelets that distributed energy throughout the wavelet and had poorly defined peaks.

The program "AVGWAVE" inputs all of the smoothed and sorted wavelets of a sample and gives a graphical display showing their distribution. It also computes the average smoothed and sorted wavelet shape. As mentioned, each wavelet is expanded and represented by 256 points. The program also computes the standard deviation for each of the 256 bands along the horizontal axis which reveals a stability measure for each area of the wavelet.

To summarize for the reader, C.A.V.I.S. uses the following attributes (parameters) from the time domain. Listed on the right side column are abbreviated codes for the parameters.

1. Total Wavelet Intensity	3w1
2. Pitch (Fundamental Frequency, f_0)	3w2
3. Auto-Correlation of Successive Wavelets	3w4
4. Variation of Total Wavelet Intensity	3w5
5. Average Energy Distribution Curve	3w8

6. Average Smoothed Wavelet Shape **wav**

Some samples of the out put data are provided in Appendix C.

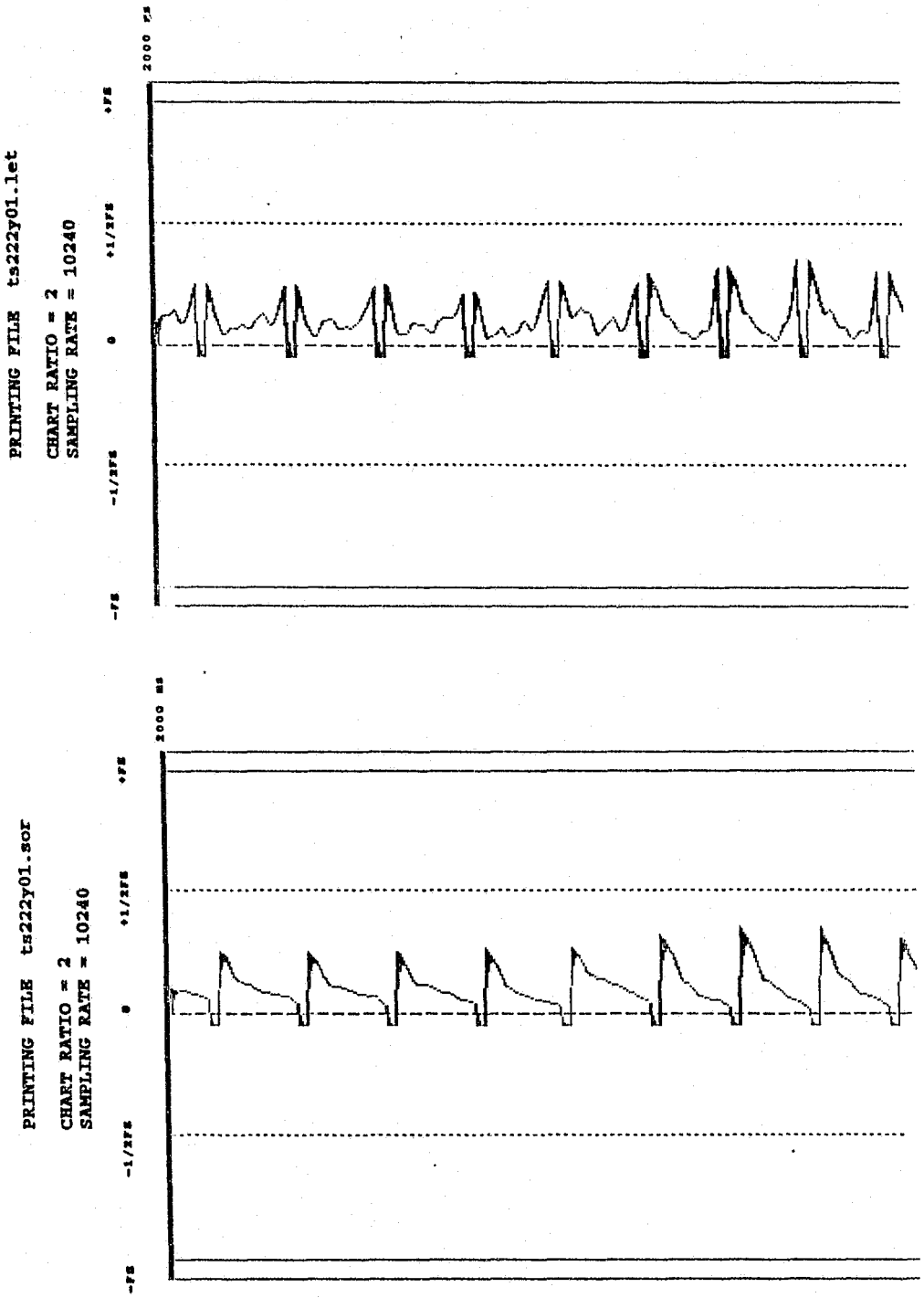


Figure 3.7-2 Graphic display showing a successive series of "wavelets". (a) shows unsorted, and (b) shows sorted "wavelets" by the intensity of each data point. Note that the wavelets are deliberately separated by the inserted negative threshold pulse for later detection.

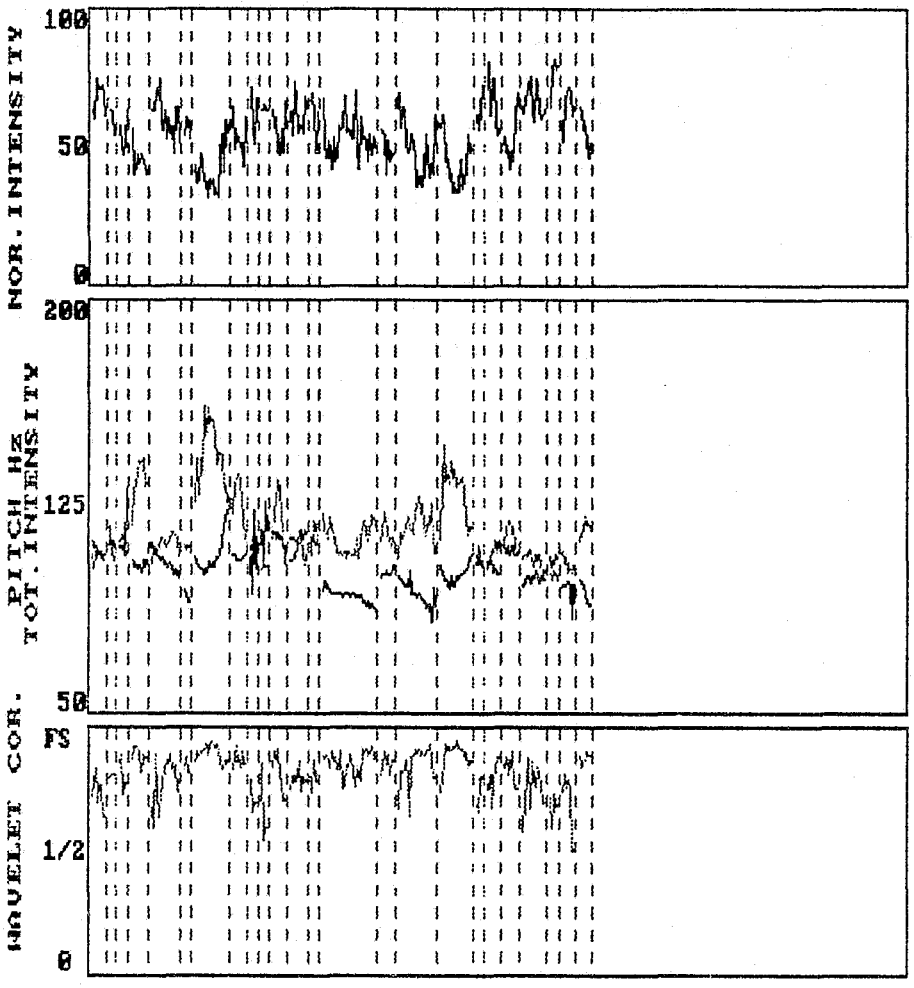


Figure 3.7-3 Graphic display showing the intermediate time domain data extracted from a set of wavelets. Upper window shows a normalized intensity contour, middle window shows a pitch contour (smooth solid curve) and a total intensity contour (light curve), and the bottom window shows a correlation coefficients contour of successive wavelets.

3.3.3 Extreme Value Statistics

3.3.3.1 Extreme Value Statistics

Suggestions of the usage of "extreme value statistics" were made by Dr. Glenn Bowie who has been working as a project technical consultant since the onset of the project in 1985. The intent of this particular statistical approach, as a mathematical tool, was to analyze the dynamics of glottal behaviors during speech utterances, such as the variation of successive cycle to cycle fundamental frequency (f_0), and changing of amplitudes associated with successive cycle to cycle f_0.

Model algorithms were developed by Dr. Bowie to compute, from the above mentioned speech phenomena, three-parameter, and also as an alternative, two-parameter Weibull functions. In the literature on extreme values statistics (Gumbel, 1955; Kinnison, 1985), calculation of two or three parameters of a double exponential function is referred to as the Weibull function. The experimental application of the three-parameter Weibull function to the above data revealed it would reliably represent the data. We, therefore, expanded the application to the data base and confirmed it's utility.

3.3.3.2 Weibull Functions and Time Domain Parameters

Figure 3.8-1 is a sample plot of a three-parameter Weibull distribution function prepared from test f_0 data. The y-axis represents the probability of ordered f_0 's. The estimated probability is shown by a solid line. The x-axis represents f_0 values in Hz after they have been rank ordered in ascending

order. Fitness of the estimated probability and ordered f_0 values is indicated by the correlation coefficient of 0.99 (most of our actual data yielded coefficients > 0.98). Please note the lower part of this figure where E0 (threshold parameter), V0 (characteristics value), and K0 (shape parameter) values are listed. These are the three parameters computed by the Weibull method, which as a set eventually will be used as one speech parameter in the subsequent speaker identification and verification operations.

Due to voluminous amount of data to process, and also boundary constraints imposed by this function to work, we have added sophisticated iterative algorithms to the prototype. Two mathematically imposed constrains are:

(1) $E0 \leq \text{MINIMUM } \{x_0, x_1, x_2, \ldots, x_n\}$, and

(2) $K0 > 1$.

The third constraint is related to the correlation coefficient which measures the fitness of the data $(x_0, x_1, \ldots, x_n)$ relative to the estimated probability. We chose that to be,

(3) $r \geq 0.98$.

Each set of three parameters was computed in a loop using up to four iterations until all three conditions (constraints) listed above were met by trimming outlying data by one standard deviation per iteration at both the low and high extremes. When

one or more of these conditions was not satisfied, the data set was considered bad, prompting us to investigate possible anomalies in the data.

Figure 3.8-2 is a sample plot of a three-parameter Weibull distribution function prepared from a 'Total wavelet Intensity' feature. Computational aspects and algorithms are similar to what has been described for the f_0 feature.

Figure 3.8-3 is a sample plot of a three-parameter Weibull function prepared from successive wavelet correlation coefficients. Since correlation coefficients range in values between -1 and +1, and Weibull statistics rejects negative values, we normalized the coefficients by the following expression:

$$r_{normalized} = (1 - r_{computed}) \times scale.$$

Figure 3.8-4 is a sample plot of the three-parameter Weibull function prepared from normalized variation of the wavelet intensity. Figure 3.8-5 is a sample plot of a three-parameter Weibull function prepared from the succession of the average energy of the wavelet.

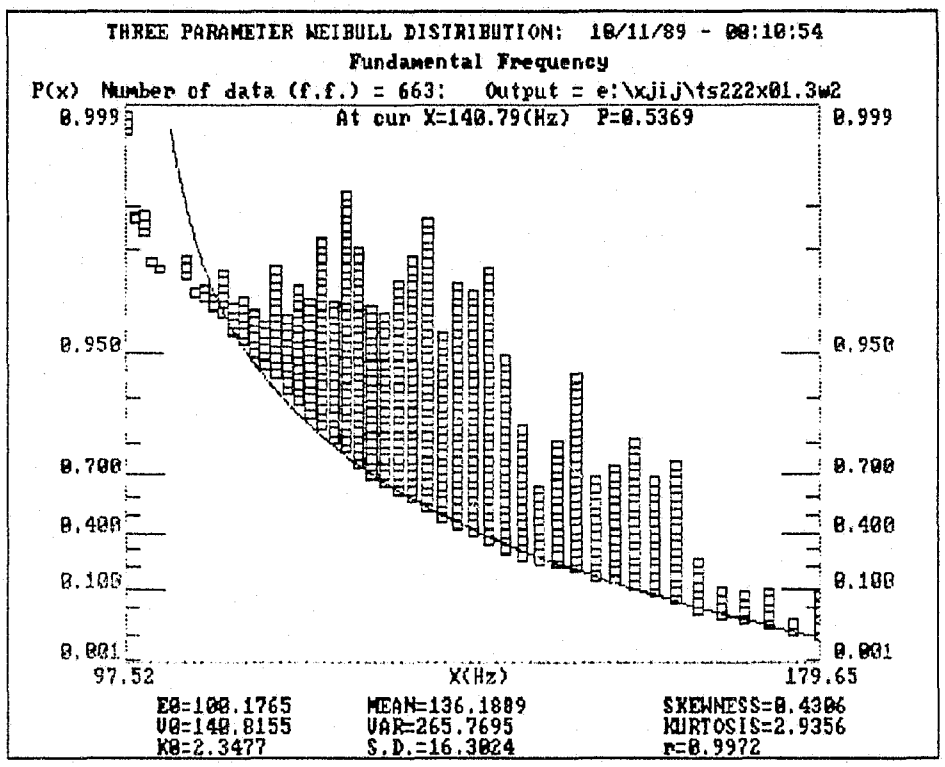


Figure 3.8-1 Plotting of a three-parameter Weibull distribution function of the fundamental frequency (f_o).

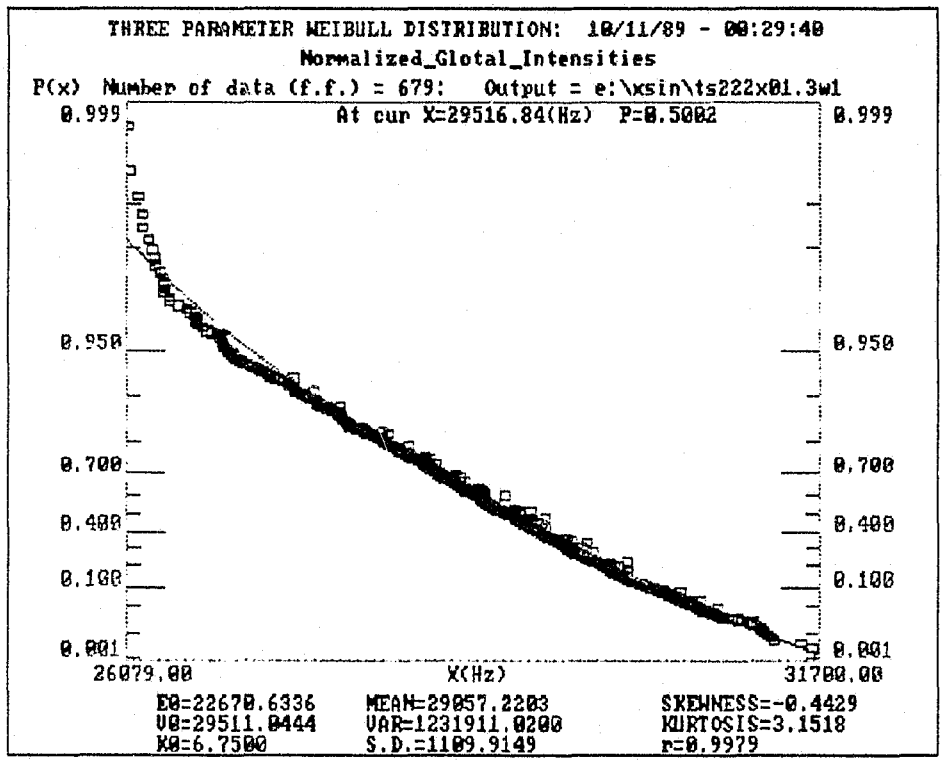


Figure 3.8-2 Plotting of a three-parameter Weibull distribution function prepared from a normalized wavelet (glottal) intensity.

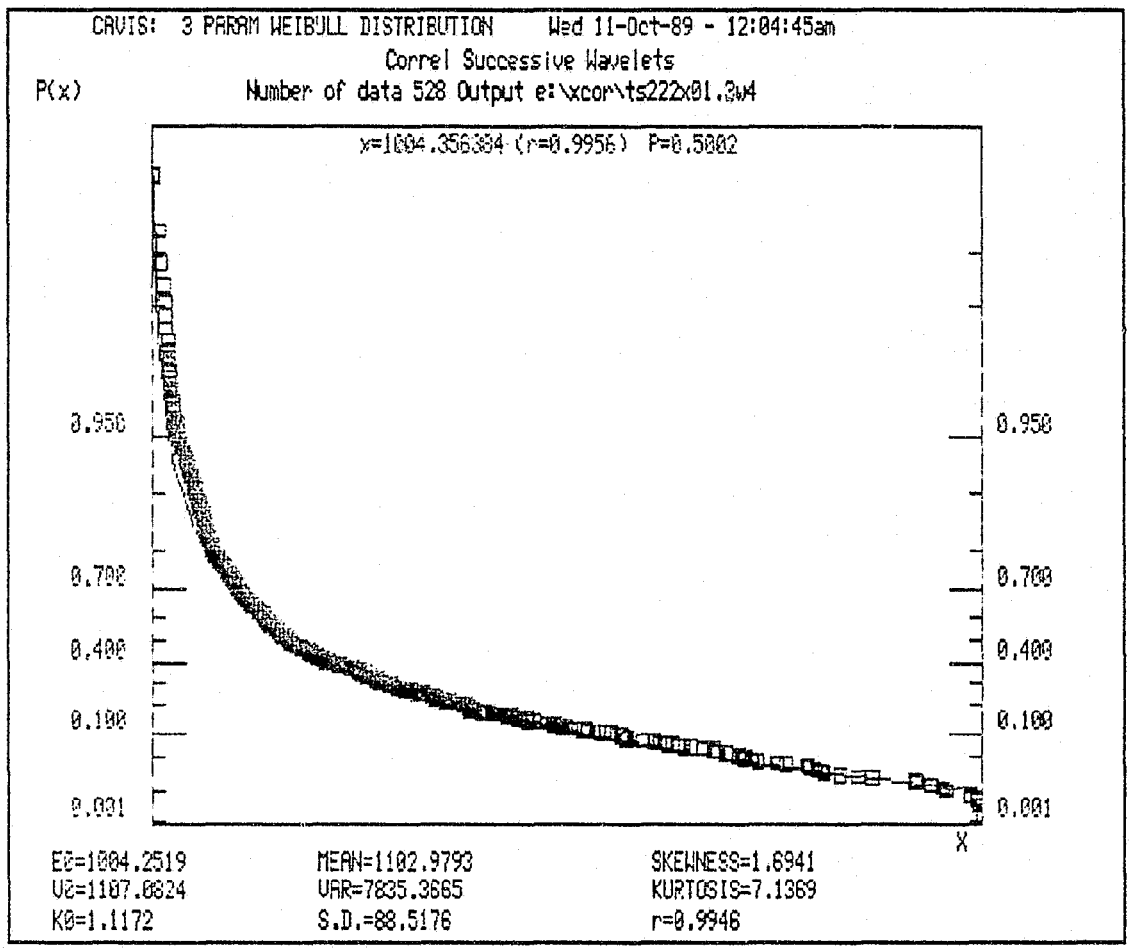


Figure 3.8-3 Plotting of a three-parameter Weibull distribution function of a set of successive wavelet correlation coefficients.

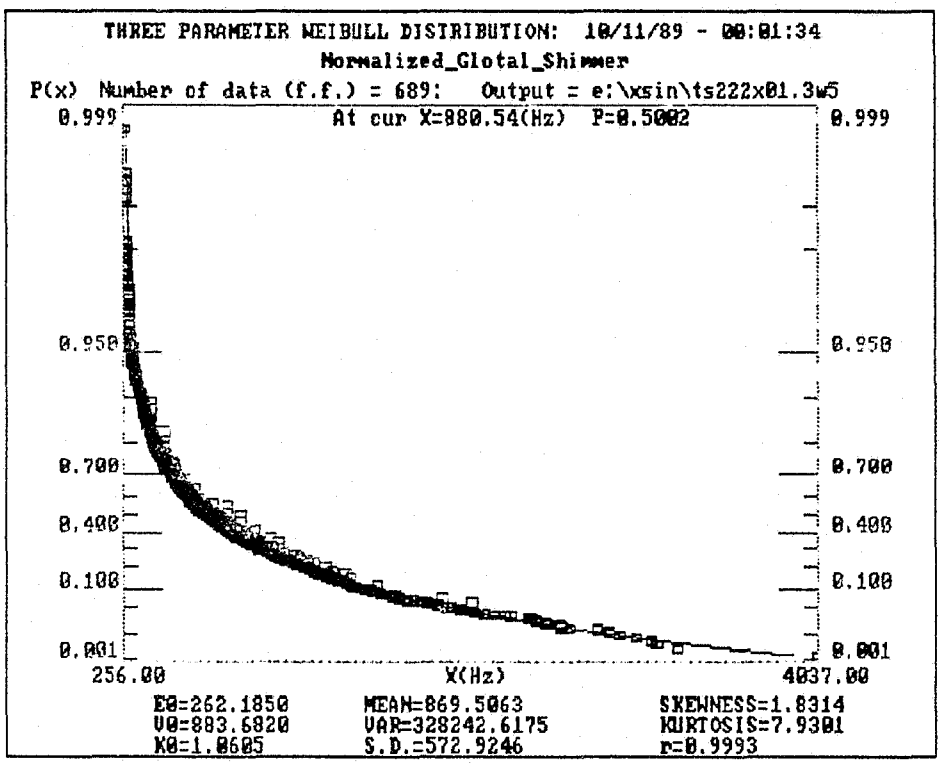


Figure 3.8-4 Plotting of a three-parameter Weibull distribution function of a set of the normalized variation (glottal shimmer) of the wavelet.

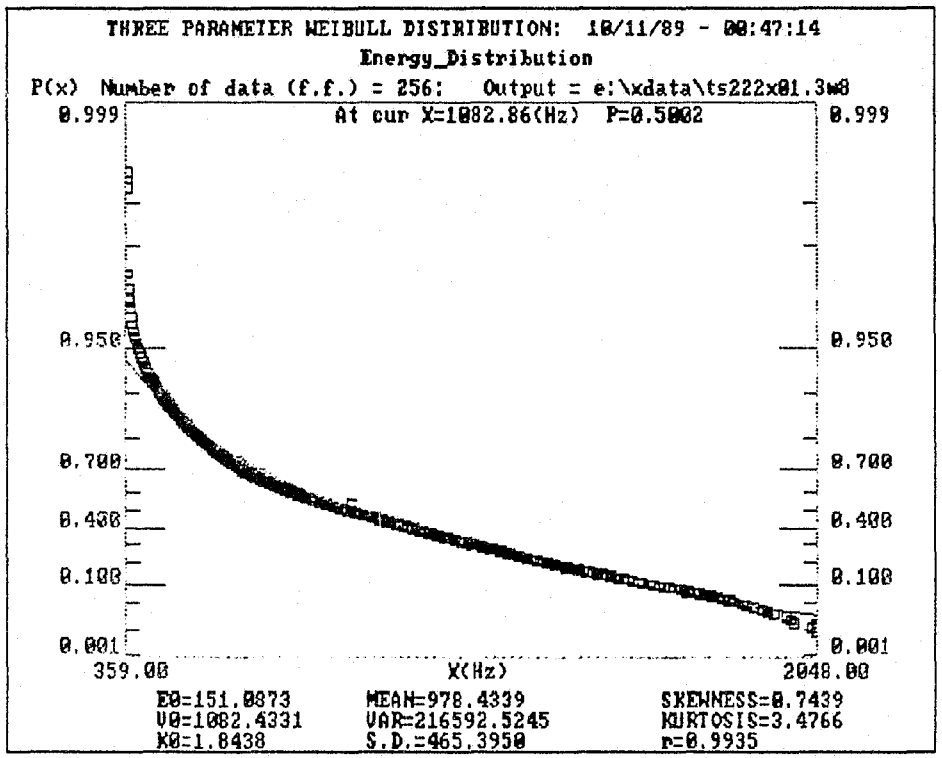
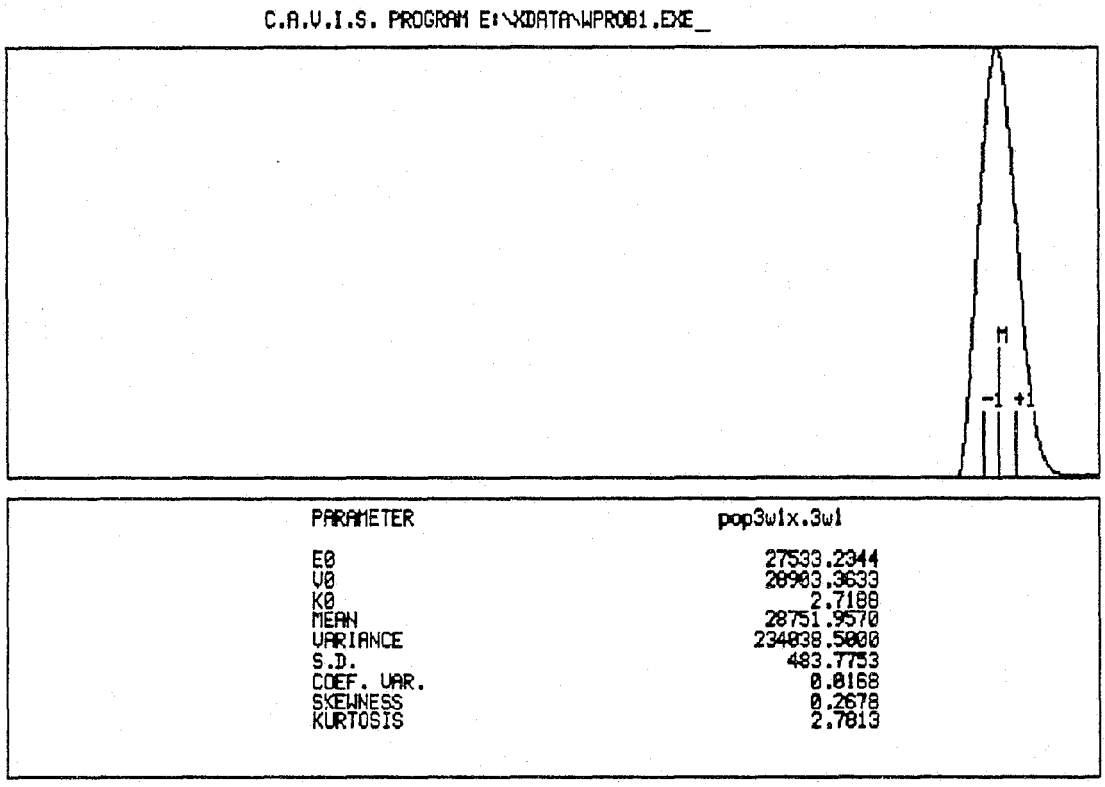


Figure 3.8-5 Plotting of a three-parameter Weibull distribution function prepared from a normalized wavelet (glottal) intensity.

3.3.3.3 Estimate of Population Distributions

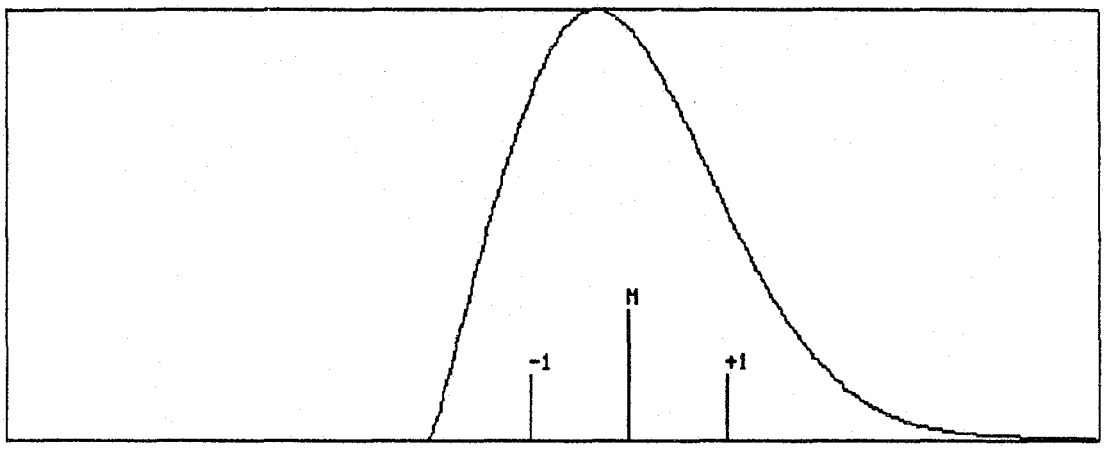
So far we have described how extreme value statistics characterized features extracted from a single speech sample (minimum of about 10 second long) spoken by the individual speaker. We expanded the utility of this statistic to investigate distributions of central tendency (mean value estimated) of features taken from the entire speakers voice data base. Figure 3.8-6 (a-e) illustrate the estimated population probability density functions of the five time domain parameters computed by the Weibull method.



(a) Normalized wavelet intensity

Figure 3.8-6 (a-e) Plottings of the estimated population Weibull probability density functions of (a) normalized wavelet intensity, (b) fundamental frequency, f_0, (c) correlations of successive wavelets, (d) normalized glottal shimmer, and (e) successive averaged wavelet intensity.

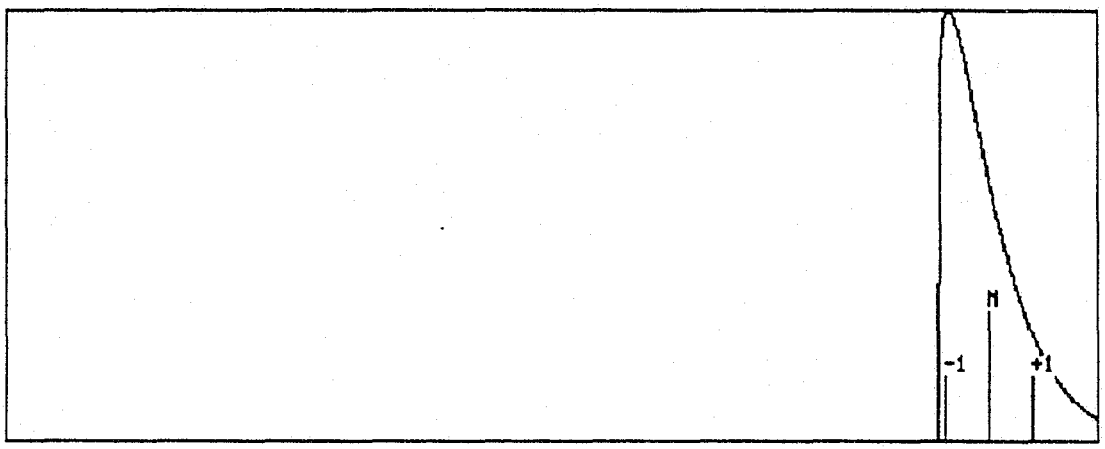
C.A.V.I.S. PROGRAM E:\XDATA\WPROB1.EXE_



PARAMETER	pop3w2x.3w2
E0	78.9154
V0	121.0772
K0	2.1250
MEAN	116.2554
VARIANCE	341.6333
S.D.	18.4833
COEF. VAR.	0.1590
SKEWNESS	0.5521
KURTOSIS	3.1076

(b) Fundamental frequency

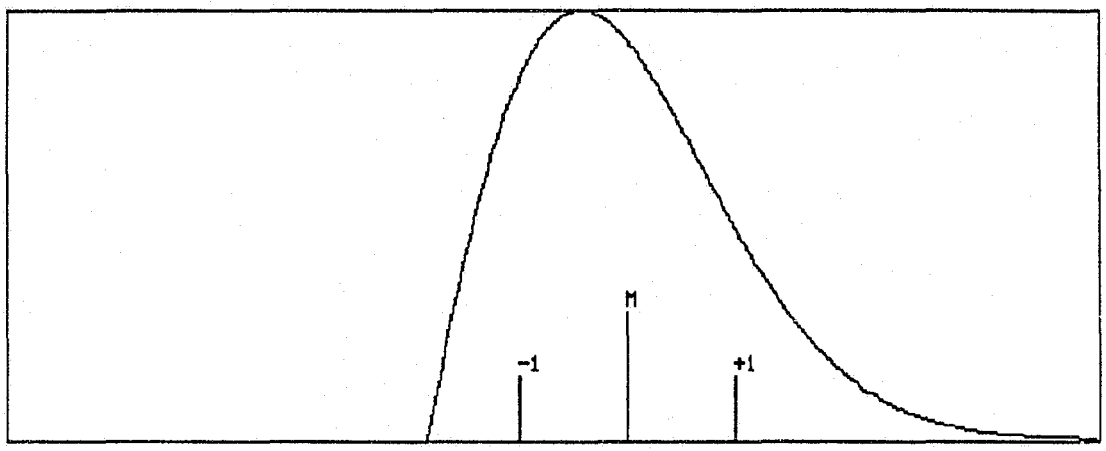
C.A.V.I.S. PROGRAM E:\XDATA\WPROB1.EXE_



PARAMETER	pop3w4x.3w4
E0	1030.5779
V0	1009.9695
K0	1.1797
MEAN	1086.6975
VARIANCE	2279.1953
S.D.	47.7409
COEF. VAR.	0.0439
SKEWNESS	1.5609
KURTOSIS	6.4334

(c) Correlations of successive wavelets

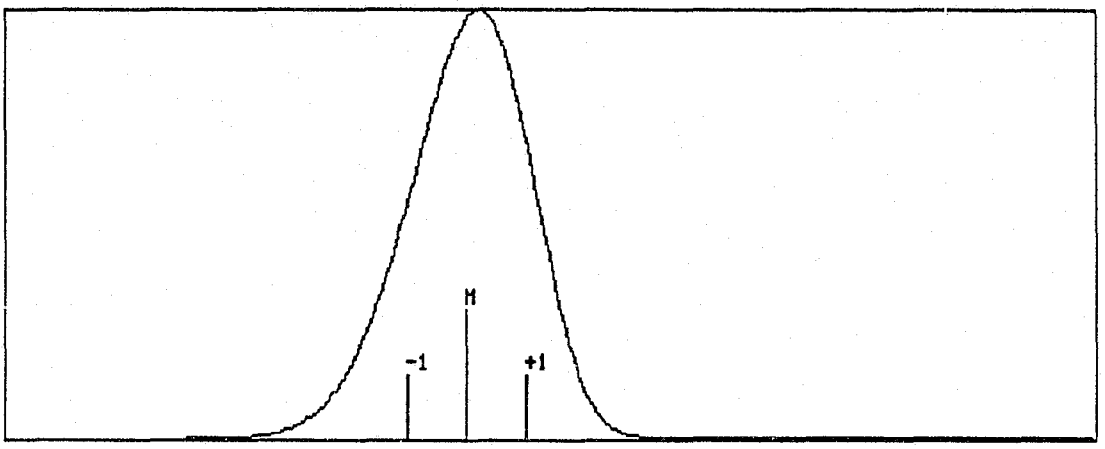
C.A.V.I.S. PROGRAM E:\XDATA\WPROB1.EXE_



PARAMETER	pop3w5x.3w5
E0	585.7995
V0	962.9722
K0	1.9219
MEAN	867.1491
VARIANCE	23239.0137
S.D.	152.4435
COEF. VAR.	0.1758
SKEWNESS	0.6853
KURTOSIS	3.3513

(d) Normalized glottal shimmer

C.A.V.I.S. PROGRAM E:\XDATA\WPROB1.EXE_



PARAMETER	pop3w8x.3w8
E0	353.3977
V0	946.6620
K0	5.4375
MEAN	900.7443
VARIANCE	13480.6748
S.D.	116.1063
COEF. VAR.	0.1289
SKEWNESS	-0.3106
KURTOSIS	2.9461

(e) Successive averaged wavelet intensity

3.3.4 Frequency Domain Parameters

3.3.4.1 Spectral Information

It has been well recognized that vocal tract configuration such as length, shape, and cross-sectional area, differs from speaker to speaker. Spectral information extracted during speech production from these varying sound resonators thus carries distinctive characteristics of the individual speakers. Of course, such spectra also convey information regarding distinctive phonemes, or sound units spoken. In a text-independent system of voice identification, we are interested in the source (speaker) of sound, but not in what particular sounds were uttered. The key to achieve this goal is found in gathering sufficiently long speech samples so that they will be phonetically balanced and ultimately more representative of the speaker.

There are many techniques practically used in general speech research, or in commercial applications to extract frequency information. A few of these are known as FFT (Fast Fourier Transform), LPC (Linear Predictive Coding), Cepstrum (Inverse of FFT) and so forth. These have been popular technics due to their well defined mathematical design and high performance in discriminating speakers under well controlled recording environments. However, under lesser controlled situations, particularly when two voices are recorded through entirely different transmission and/or recording media, spectral frequency information of voices tends to be altered due to the overpowering response characteristics of two different systems.

If we desire to apply the voice identification system to a forensic environment where the voice from the questioned individual and the one from the suspect are commonly recorded through different transmission systems, some methods of compensation for the adverse influence must be incorporated.

In C.A.V.I.S. we attempted to accomplish a method of such compensation by the use of a 'whitening technique' and intensity deviation spectra (IDS). The whitening technique has been discussed in the previous section (Voice Data Pre-Processing). The computational algorithm of an IDS will be detailed in the section to follow.

3.3.4.2 Spectra Matching

Given two spectra, there are many possible ways to compare the closeness of these spectra. We have tested three methods. In the first method we computed a Euclidian distance of 9 dimensions, each dimension having the value associated with the centroid frequency of each band. In the second method we computed a Euclidian distance of 9 dimensions, each dimension having the total sum of energy from each band. In the third method we compared the shape, or envelopes of given spectra by using the use of product moment correlation. It was found that the last method, i.e., shape matching by correlation, is more effective than the measure of the remaining two, and this method is applied in our spectral matching of a pair of IDS spectra.

3.3.4.3 Intensity Deviation Spectrum (IDS)

Every speaker produced five 30 second long speech samples and each IDS was computed from these samples after the pauses are deleted and the set of short-term spectra are generated by the FFT routine. A 1024-point FFT is performed by the use of a Digital Signal Processor Board on the previously prepared time data. From each compressed voice data this FFT operation yielded a set of n short term FFT frames (n = total length / 1024), where an n usually ranges from 100 to 400 frames. The set is referred to as ST_n for n=1,2,...,N.

An intensity deviation spectrum (IDS) is computed from the ST_n discussed above by the following formula.

$$S_j = \frac{\sum_{i=1}^{i=I} |ST_{i,j} - m_j| / m_j}{n_j}$$

Where $ST_{i,j}$ is the value of the j^{th} band of the i^{th} FFT frame that meets $ST_{i,j} > T$ (threshold) condition. 'T' is a value 30 dB below the maximum value measured from a set of the entire set $ST_{i,j}$. 'n_j' is the number of $ST_{i,j}$ values which satisfy $ST_{i,j} > T$ condition, and 'm_j' is the median value of $ST._{,j}$ within the j^{th} band.

The IDS spectrum computed is then subjected to the normalization process across the discrete frequency range from 200 to 2500 Hz.

$$IDS_i = \frac{S_i}{\sum_{j=20}^{j=245} S_j}$$

for i = 0 to 225.

Consequently, the normalized IDS values are made to range between 0 and 1. This normalization is a required procedure to enable subsequent computations.

Figure 3.8-7 shows 10 IDS's from one speaker, 5 from session 1 (TS222x), and 5 from session 2 (TS222Y). Figure 3.8-8 shows 5 IDS's of the speaker from session 2 (TS222y) and 5 IDS's of the speaker from session 2(TS233y).

It will be shown later that each vector contained parameters accompanied by corresponding weighting factors that are speaker specific. The necessity of assigning the weighting factor for each parameter has been confirmed through our tedious laboratory observations, and it's implementation has been carried out during this experiment.

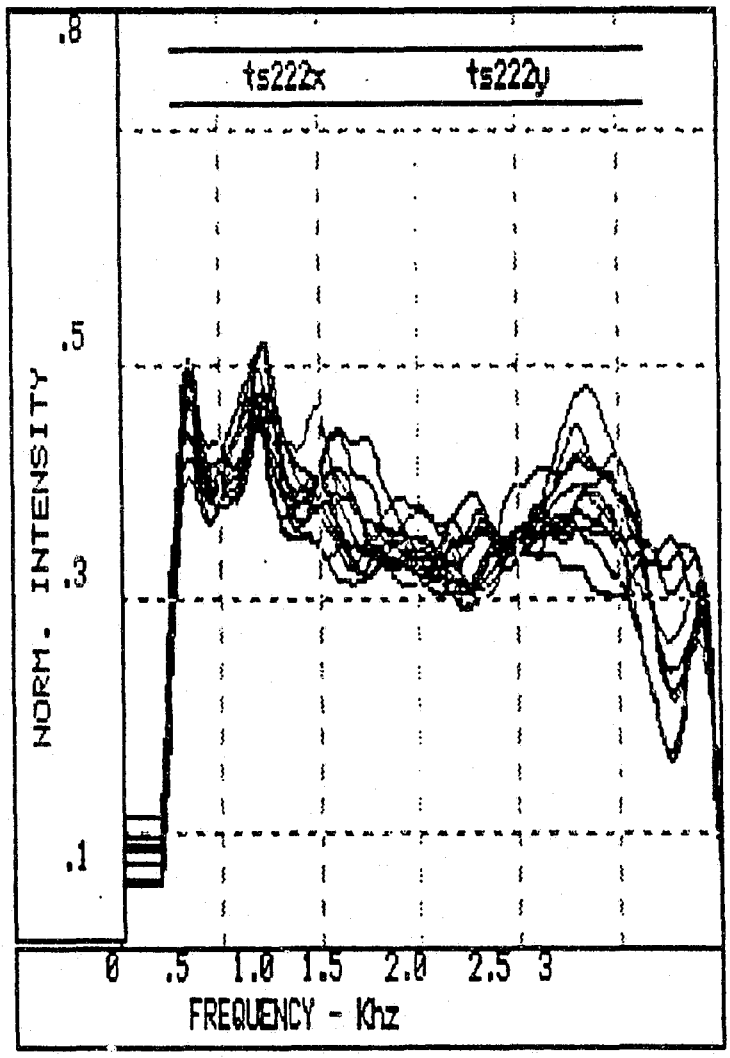


Figure 3.8-7 Plottings of the IDS generated from the same speaker. Of 10 IDS spectra, 5 were made from speaker ts222x recorded in session 1 (in color cyan, or lighter shade), and 5 were made from the same speaker ts222y, but recorded in session 2 (in color red, or darker shade).

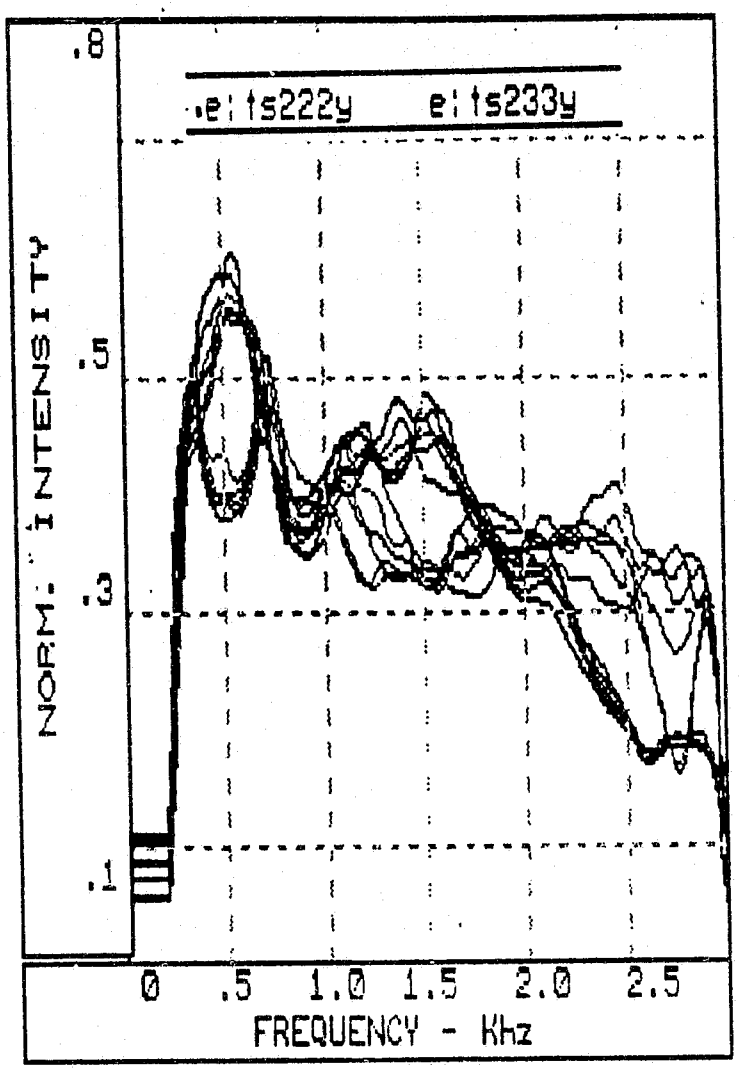


Figure 3.8-8 Plotting of the IDS generated from two different speakers. 5 IDS spectra were made from speaker ts222y, and 5 other IDS spectra were made from the different speaker ts233y, both recorded in session 2.

3.3.5 Combining Time And Frequency Domains

Each sample of a speaker was represented by a vector of m+n dimensions (parameters or features), where m refers to the number of parameters derived from the time domain, and n, to the number derived from the frequency domain. Presently m=5 and n=9 are selected as the optimum parameters.

Although it has been reported by many researchers that spectral information (vocal tract parameters) is more effective for distinguishing speakers than that derived from vocal cord behaviors (time domain parameters), there were cases where spectral information alone would not discriminate speakers. Through our laboratory observation, those who were likely to be misidentified became distinguishable when at least one of the time domain parameters was employed.

Nevertheless, spectral information appears to remain far more powerful in discriminating individuals than any single parameter from the time domain. For this reason, we believe that parameter sets from the frequency domain and from the time domain must be combined to achieve a high performance voice identification system. Prior to actual devising of refinements, we made the following observations.

Parameters From Time Domain:

(1) Some speakers maintained a high degree of stability in their average pitch across the samples in the first recording session, and also in the second session.

(2) Some speakers remain stable in their average pitch, but only within a single recording session. Some of this group

revealed differences in the average pitch as much as 15 to 20 Hz. Usually, the pitch tended to be higher when measured from voice data recorded in the second session than in the first session[1].

(3) Other groups of speakers manifested a high level of variations in their pitch across two sessions, as well as within a single session.

(4) These parameters indicated very interesting behaviors. Taking "variation of total sum of wavelet energy", for example, the parameter distinguishes clearly a certain group of speakers, but does not do well with other groups. In other words, this type of parameter seems capable of separating different specific groups of speakers, but does not seem to give any hint as to identifying or separating other groups of speakers.

Parameters From Frequency Domain:

(5) Particular individuals maintained stable IDS envelopes, throughout all bands (200 to 2450 Hz), or at least within 4 - 6 bands. These speakers were not necessarily the same group of speakers as described in (1).

(6) Only a few speakers exhibited a complete match of IDS envelopes throughout nine bands, either within a session, or

---

[1]In the first session, the speakers produced speech samples spontaneously, but while looking at a picture set provided to them. However, in the second session, they produced speech samples without a picture set. In general, data from the first session showed more stability in terms of rate and pitch change, whereas data from the second session exhibited less stability in rate and pitch.

across two recording sessions.

(7) Scrutinization of 16 test speakers through visual pattern matching of their IDS envelopes revealed that each individual speaker has unique areas of stability. For example, speaker A may have stable bands, between 450 to 700 Hz, 950 to 1200 Hz, and 1200 to 1450 Hz, but unstable bands, between 200 to 450 Hz, 700 to 950 Hz, etc., while speaker B reveals stable bands between 200 to 450 Hz, 450 to 700 Hz, 700 to 950 Hz, but totally random in the remaining bands.

Taking the above observations into consideration, we incorporated refinements in the final stage of the voice vector definition within the experimental design. The first refinement is related to frequency domain parameters, or IDS bands. By the use of a correlation method, the stability measure is calculated for each IDS band. The second refinement is related to the time domain parameters. This refinement involves testing the fitness of a given parameter for a given pair of speakers under comparison. Algorithmic aspects are discussed in sections 4.2 and 4.3.

4 EXPERIMENTAL PROCEDURES

4.1 General Views

Figure 4.1-1 shows the general flow of the experiment used in the voice identification/verification processes. Identification and verification processes were conducted in tandem. The entire process goes automatically by beginning with the speech parameters that have been prepared already in the previous pre-processing stages.

The input database included 49 speakers, yielding 5 text-independent samples per speaker for each of the two recording sessions. Each session was separated by a minimum two month period, and each sample was represented by a vector of 5 time and 9 frequency domain parameters. There are 245 (known speaker samples) x 245 (unknown speaker samples) = 60,025 comparisons to be performed, and the total possible number of trials are 245 per experimental conditions.

Figure 4.1-2, -3, and -4 illustrate examples of the processed input voice data including 5 time domain and 9 frequency domain parameters. Figure 4.1-2 illustrates a sample case where the two given speakers (actually the same speaker, but one was recorded in session 1, and the other recorded in session 2) are displayed concurrently and considered to be a good match. Figure 4.1-3 illustrates a contrary sample case where the two given speakers displayed are considered to be no match. Figure 4.1-4 shows an example case when a "no decision" decision is likely to occur because of the low stability seen in the IDS

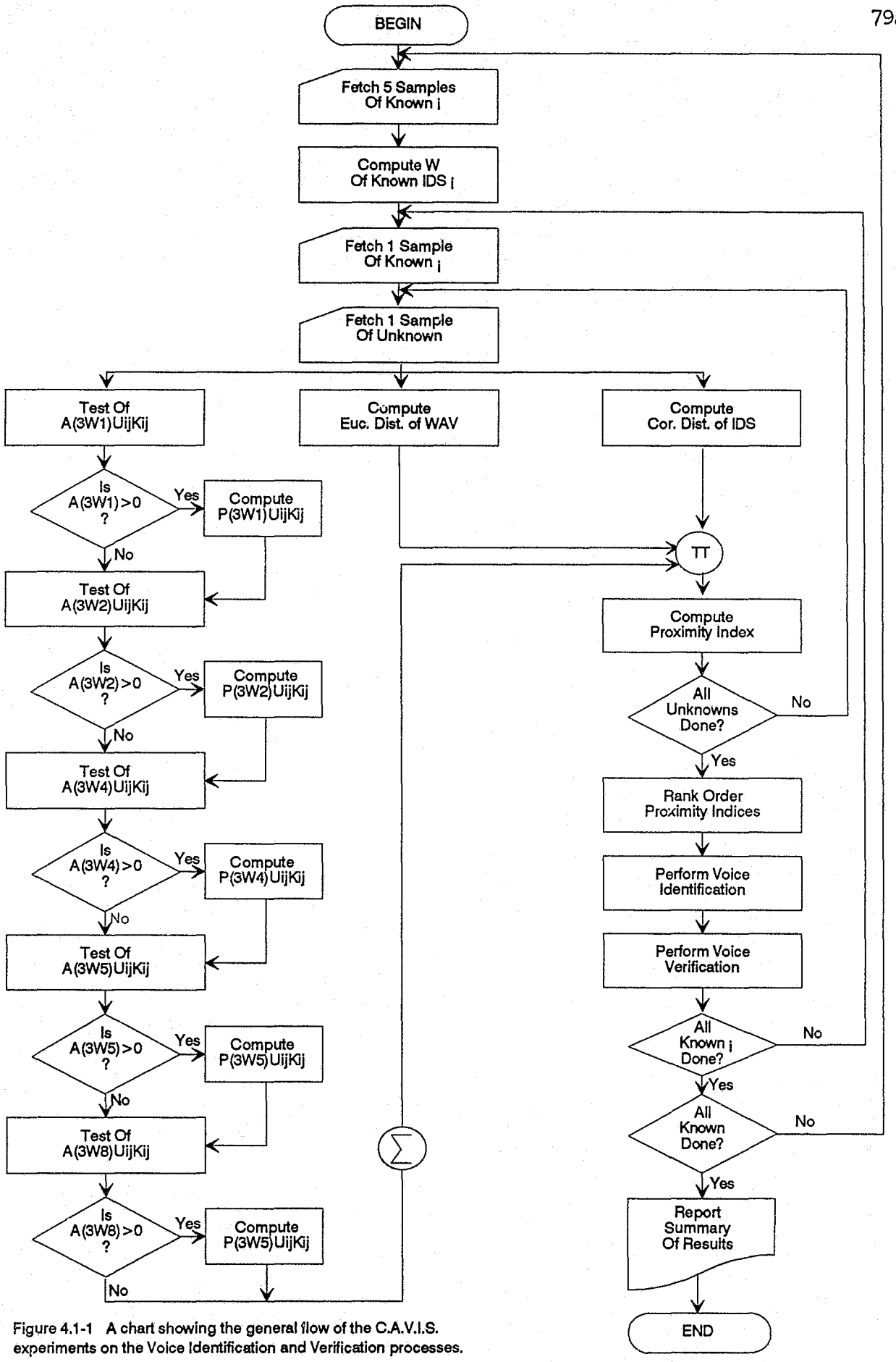


Figure 4.1-1 A chart showing the general flow of the C.A.V.I.S. experiments on the Voice Identification and Verification processes.

spectra of the known speaker samples.

For experimental purpose, we treated the voice data recorded in the first session as unknown speakers, and the data recorded in the second session, as known speakers. The rationale for choosing voice data recorded during the second session as the "known speaker" is as follows. In most real forensic situations, a questioned call recorded earlier in time would belong to a criminal whose identity is unknown, whereas voice exemplars recorded later would belong to a suspect(s) whose identity is usually known. Such an arrangement of known and unknown voices provide advantage in the real forensic world: Generally we do have the liberty of collecting as many voice samples from the known suspects with reasonable variations in speaking rate and mode. It then becomes convenient to investigate the variations of speech samples taken from the known individuals to determine what particular speech parameters best fit to represent that individual. In contrast, we have very little control over the duration, mode, and content of speech, environmental noise, and so forth of the questioned call once the recording has been made. Next, we will discuss the algorithms developed to determine such best fit parameters.

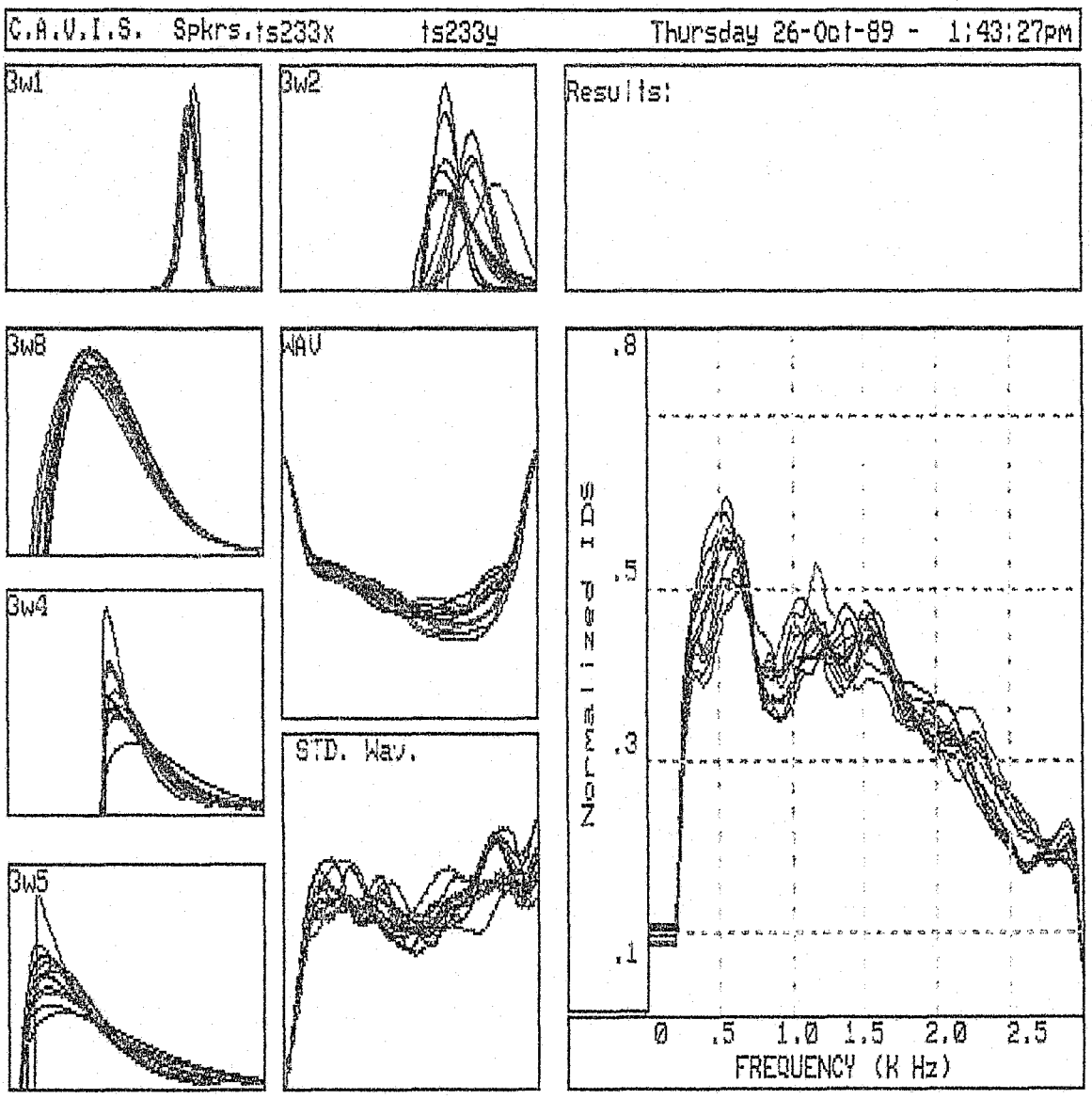


Figure 4.1-2 Graphic display showing the processed parameters: a case of the matching speakers. In each window, the unknown speaker's 5 samples are drawn in color of cyan, and the known's, in color of red. Far right side window contains the corresponding IDS spectra which are partitioned (not visible in the graph) into 9 bands of 250 Hz width.

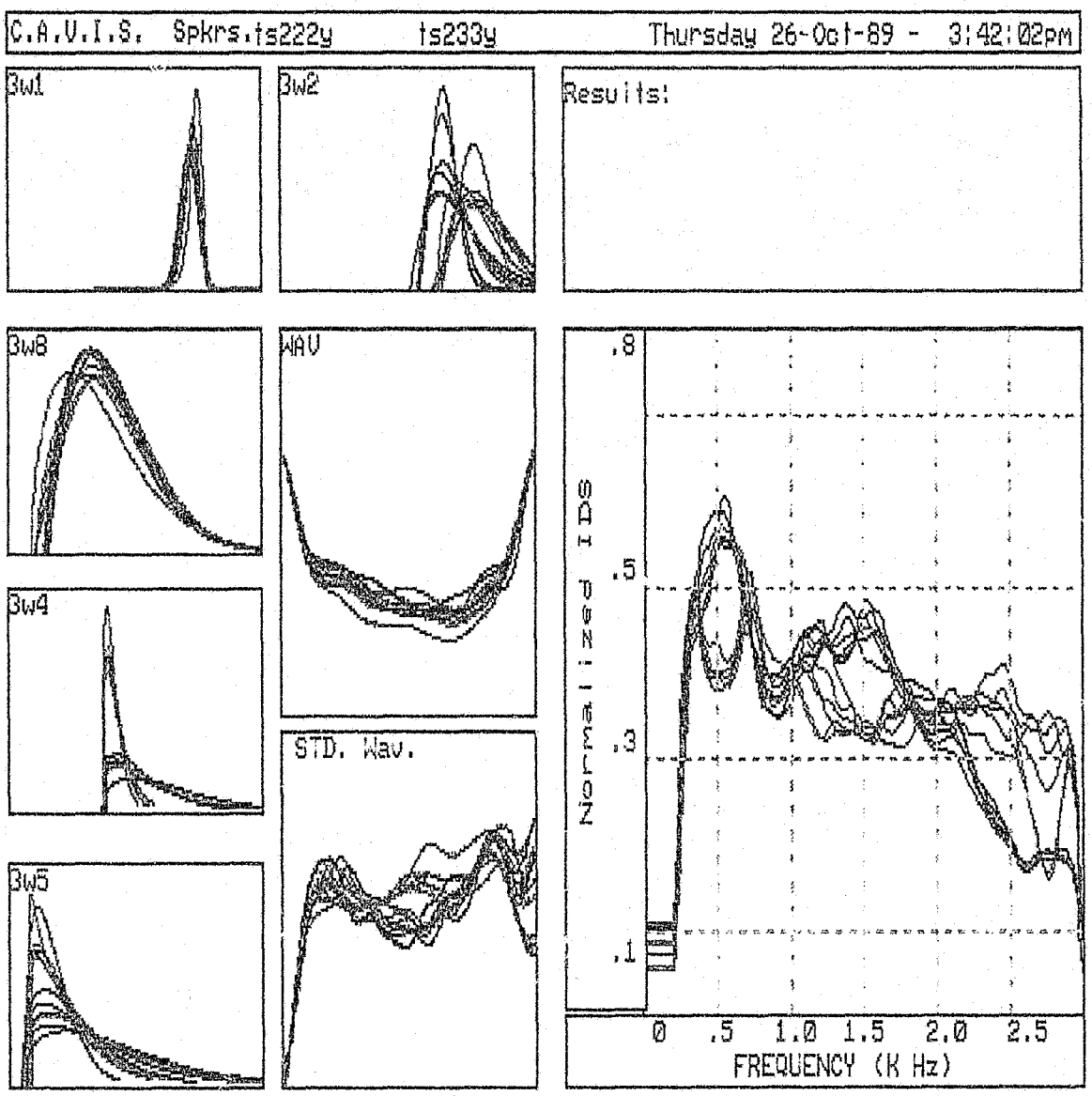


Figure 4.1-3 Graphic display showing the processed parameters: a case of non matching speakers. In each window, the unknown speaker's 5 samples are drawn in color of cyan, and the known's, in color of red. Far right side window contains the corresponding IDS spectra which are partitioned (not visible in the graph) into 9 bands of 250 Hz width. Note the compactness (high stability) of 5 IDS's of each speaker, and clear separation into two groups of the IDS spectra.

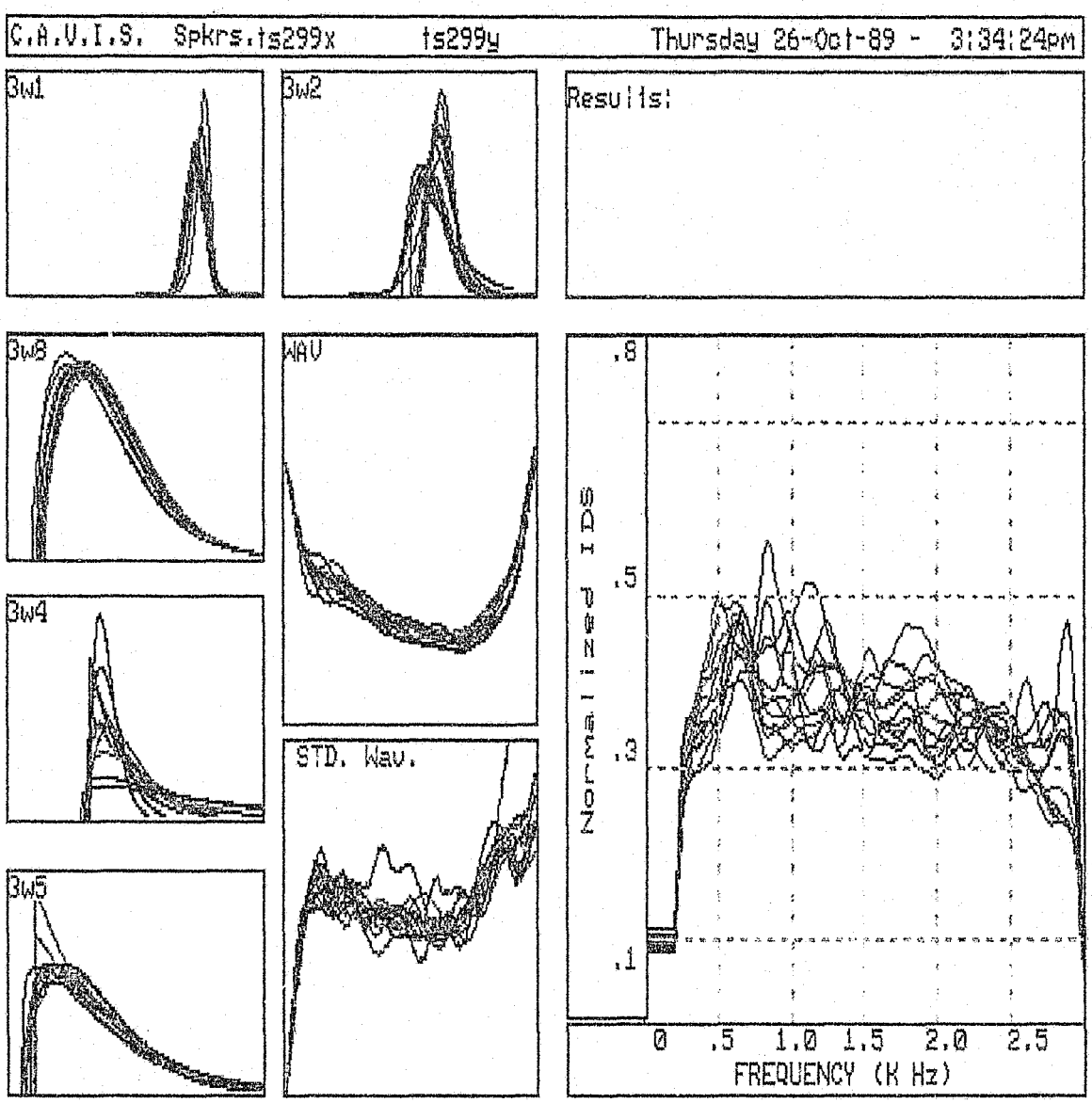


Figure 4.1-4 Graphic display showinbg the processed parameters: a case which the system is likely to deliver a "no decision". Note the unknown speaker ts299y's IDS spectra (drawn in color red) which show only a small amount of the stability throughout the entire band. Because of this instability, despite the fairly good matching results from time domains, the final decision by the system is predicted "no decision".

4.2 IDS Spectra And Weighting Factor

Each IDS ranging from 200 to 3000 Hz were partitioned into 11 equal bands of 250 Hz, each band having 25 discrete frequencies of 10 Hz width. Data above 2450 Hz was discarded, thus retaining 9 bands. From each band we computed w_j (for j=1 to 9) to be applied as the relative strength (weighting) of the j th band of IDS for a specific individual speaker. The following equation was used to compute the weighting.

$$w_j = 1 + \frac{\sum_{i=1}^{i=I-1} \sum_{k=i+1}^{k=I} r_{j,i,k}}{N}$$

where,

I = 5 (number of samples / speaker)

N = I x (I-1) / 2; or, ($_5C_2$)

$r_{j,i,k}$ = correlation coefficients measured, along the j th IDS band, between the i th and the k th samples from a known speaker.

Any w_j values less than 1 are assigned a value of 1 to maintain the weighting factors for positive direction only. In effect, the greater the value of w_j, the more stable and reliable the j th IDS band would be to characterize a known speaker. In other words, w_j can be considered as the measure of the intra-speaker variability, or the average variability (correlation) within an individual. The number of valid w_j's may vary depending on each individual's variability, and in a

case where there are only two or less number of w_j's determined valid, this particular known speaker is to be labeled as unreliable, which will lead to a "no decision" case in the subsequent sections. In essence, C.A.V.I.S. is designed to deliver a "no decision" decision when the given speaker's stability within himself is too low.

At this point after the known's weighting factors, or variability measurements of the frequency parameters have been determined, the unknown's voice samples are read one at a time for comparison. From the experimental unknown speaker's voice sample we assumed no liberty of computing speaker specific weighting factors for the parameters despite the availability of the five samples.

4.3 Tests of Time Domain Parameters

We noted through observation that most time domain parameters discriminate some speakers, but not all of them. In order to determine whether each of the time domain parameters should be included in computation of the probability of match between two items, we devised a procedure as expressed below.

$$A = M_{kEj} \pm \frac{S_{kEj}}{2} \cap M_{uEj} \pm \frac{S_{kEj}}{2}$$

where,

M_{kEj} = Known's mean value of the j th time domain parameter, M_{uEj} = Unknown's mean value of the j th time domain parameter,

S_{kEj} = Standard deviation of the j th time domain parameter that is based upon the all known voice samples, and

A = Area of intersection.

If the resulting value of A is greater than 0, this test fails and the probability of match between the known and unknown is not computed. In case the value of A is equal to 0, i.e., no intersect occurs, then, the given time parameter, E_j, participates in the computation of the probability[1].

The $p(_{k,u})$ is the probability of a match between the two voices based solely on the j th time domain and is expressed as the total area formed by the cross overs of the two probability density functions. A probability of match that is expressed as the intersect of two density curves is computed by the use of two sets of e0, v0, and k0 values, where one comes from the known and the other, comes from the unknown. Expressed simply in a 'C' language function calling convention:

$$p(_{k,u}) = f(e0_k, e0_u, v0_k, v0_u, k0_k, k0_u)$$

where,

$e0_k$ and $e0_u$ are Weibull threshold parameters

$v0_k$ and $v0_u$ are Weibull characteristic values, and

[1]The actual computation is based on the theory of extreme value statistics and the prototype algorithm was provided by Dr. Glenn Bowie. The detail is found in APPENDIX A.

$k0_k$ and $k0_u$ are Weibull shape characteristic values.

The above procedures are performed on each of the five time domain parameters and the final single figure, $E_{k,u}$ generated by these five probabilities is derived by:

$$E_{k,u} = 1 + \frac{\sum_{i=1}^{i=I} (1 - P_{Ei})^2}{I}$$

where P_{Ei} is the probability of match, or intersect area computed by the three parameter Weibull function of the i^{th} time domain parameter, and I is the number of valid parameters which satisfied the test.

The mathematical procedures for the computation of $p(E_i)$ above are provided in Appendix A, and interested readers for further theoretical principles are referred to Kinnison (1985) and Gumbell (1955).

4.4 Euclidian Distance of Wavelets

The averaged wavelet shape was partitioned into three segments, each segment being composed of 30 data points that formed a smoothed decay shape of a wavelet. First, each data was scaled by dividing it by the maximum value of 2048 so that all the resulting data would range from 0 to 1.

The purpose of this scaling was to normalize the range of distance so that it would be suitable as a homogeneous element

within the speaker vector that is comprised of combined parameters from time and frequency domains. Then, the euclidian distance, $d_{k,u}$, was computed between the wavelet shapes (of an unknown and a known speakers) by:

$$d_{k,u} = \frac{1}{3} \sum_{j=1}^{j=3} \sqrt{\frac{1}{30} \sum_{i=j\cdot 30}^{i=(j+1)\cdot 30} (wav_{k,i,j} - wav_{u,i,j})^2}$$

The adjustments (factors 1/3 outside the square root, and 1/30 inside the square root) were to normalize the range of distance so that it would be suitable as a homogeneous element within the speaker vector.

4.5 Computing Correlational Distance

The correlational distance measure of a pair of IDS's for an unknown and a known is computed by:

$$D_{k,u} = \left(\sum_{j=1}^{j=J} 1 - \frac{w_j \cdot r_j}{2} \right) / J$$

where $D_{k,u}$ is adjusted distance based upon r_j, the correlation coefficients, between known and unknown measured on the IDS bands. The weighting factor, w_j, indicates the relative stability (or inversely related to the intra-speaker variability)

of the j^{th} IDS band for the individual speaker used as the known. The value of these variables range:

$-1 <= r_j <= +1,$

$1 <= w_j <= 2,$ and

$0 <= D_{k,u} <= 2.$

The above expression is designed so that the maximum separation between the two items occurs when $D_{k,u} = 2$, and the minimum separation (match) occurs when $D_{k,u} = 0$.

4.6 Proximity Index

The term 'proximity index' is our preferred term over the use of the term 'probability of match' to represent a measure of the similarity between the two voices being compared. In order to avoid possible confusion that the term may cause, a brief explanation is in order.

The probability of match is computed for the time domain parameters on a solid mathematical foundation as mentioned earlier in section 4.2, however, because of the inclusion of the heuristic testing procedures whether probability should be computed or not, and a hybrid application of correlational distance measured from the IDS spectra, we concluded that the term 'proximity index' better fits the design of our system.

The proximity index, $P_{k,u}$, is expressed by:

$$P_{k,u} = D_{k,u} \cdot d_{k,u} \cdot E_{k,u}$$

where $D_{k,u}$ is correlational distance computed between a known an unknown IDS vectors, $d_{k,u}$ is the Euclidian distance computed from the wavelet shapes, and the $E_{k,u}$ is the summation of the squared errors computed from the time domain parameters. The proximity index, $P_{k,u}$ ranges from 0 (for $D_{k,u}=0$, and for any value of $d_{k,u}$ and/or P_{Ei}) to 4 (for $D_{k,u}=2$, $d_{k,u}=1$, and for $P_{Ei}=2$).

The maximum match occurs when $P_{k,u} = 0$, and the minimum match occurs when $P_{k,u} = 4$. If none of the time domain parameters meets the test 'A' condition, the proximity index reflects only the value of $D_{k,u}$, i.e., only the spectral information carried in the IDS plays a part in the voice identification decision process.

When too much variation (or low stability) as illustrated in Figure 4.1-4, is found throughout the entire IDS bands within a known speaker, $P_{k,u}$ is not computed and a "no decision" is rendered. In other words, this speaker's voice exemplars are deemed unstable. In fact, they may, or may not be poor samples in terms of the adequacy of the recording. In any rate, the samples are not allowed to take part in further comparisons as the known speaker. In this case, the minimum required number of valid IDS bands is set to 2.

A vector used to represent a validated known and also an unknown speaker is composed of 14 parameters, and in the process of computing the proximity index each parameter is appropriately weighted, included or excluded, and at the end the similarity

measure will be summarized into a single index.

4.7 Rank Ordering of Proximity Indices

The proximity indices, $P_{k,ui}$, computed between the given known speaker samples and to all unknown speaker samples are rank ordered in ascending order:

$$P_{k,u1} < P_{k,u2} < P_{k,u3} \ldots < P_{k,uN}$$

where $P_{k,u1}$ is the smallest proximity index value (the closest distance), $P_{k,uN}$ is the largest proximity index value (the farthest distance), measured between a given known speaker's sample and any of the samples of N unknown speakers. This ordered set of proximity indices are subsequently used to evaluate whether the given known speaker is correctly identified, or incorrectly identified in the voice identification experiment in the section to follow.

4.8 Voice Identification

Let us denote a known speaker i and his 5 samples by $K_{i,j}$, an unknown speaker i and his 5 samples by $U_{i,j}$, and the rank ordered proximity indices of $K_{i,j}$ to all the unknown speaker samples by $R(K_{ij},U_{ij})$. The identification result is evaluated each time a given known speaker sample is compared with all unknown speaker samples. The result is either a correct or an incorrect identification and is defined by:

$$\text{If Ui } \epsilon \{R_0(K_{ij},U_{..}),R_1(K_{ij},U_{..}),\ldots,R_N(K_{ij},U_{..})\} \text{-> Correct}$$

$$\text{If Ui } \notin \{R_0(K_{ij},U_{..}),R_1(K_{ij},U_{..}),\ldots,R_N(K_{ij},U_{..})\} \text{-> Incorrect}$$

An 'N' that appears in the above expression means the N th ordered (ascending) proximity index between the known sample and the unknown. Identification performance is tested for 15 different levels of N: N=0, 1, 2,..,10, 15, 20, 25, and 30. When N=0, we have the most stringent test. In this case a correct identification occurs only when one of the given unknown's sample yields the smallest proximity index value to that of a known under process, i.e., no other unknown speaker's sample should be closer to that known speaker's sample. When N=1, test becomes less stringent, i.e., a correct identification occurs if the ranking of proximity index of one (or more) of the unknown's sample falls within ranking of 2, and so forth.

Tabulated performance results under each value of N will be presented in the next Chapter.

4.9 Voice Verification

In the process of voice verification, the magnitude of the proximity index, $P_{k,u}$, instead of rank ordering, is applied to determine whether the known speaker is identified (verified) or not identified (rejected). Under this verification process, we set the verification criterion, V_c, which takes a selected proximity index value. The verification decisions are made by

the following simple rule.

If $P_{k,u} \leq V_c$ Verify the given known as the unknown('same')

If $P_{k,u} > V_c$ Reject the given known as the unknown('different')

By the use of 'a priori' information, then, these two responses by the system are checked whether it is 'true' or 'false'. Consequently, there will be four possible outcomes, and these are: verifying the given known speaker as same as the unknown, and actually it it is true (correct identification), verifying the given known speaker as same as the unknown, but actually it is false (incorrect identification), rejecting the given known speaker as different from the unknown, and actually it is true (correct elimination), finally, rejecting the given known speaker as different from the unknown, but actually it is false (incorrect elimination). For the purpose of rating the system performance, these four outcomes were expressed in terms of the four kinds of probability: (1) p(S|s), the conditional probability of correct identification - system announcing 'Same' and actually two given voices are made by the 'same' speaker, (2) p(S|d), the conditional probability of incorrect identification - system announcing 'Same' although actually two given voices are made by 'different' speakers , (3) p(D|d), the conditional probability of correct rejection - system announcing 'Different' and actually two given voices are made by the 'Different' speakers, and (4) p(D|s), the conditional probability of incorrect rejection - system announcing 'Different' although

actually two given voices are made by the 'same' speaker.

Further, in order to determine the general area of the optimum values of the verification criterion 'V_c', it's value is varied in terms of the different proximity index value, $P_{k,u}$. The results of verification performance as a function of different V_c values are summarized in the next Chapter.

5 RESULTS

This chapter reports the results of the voice identification and verification experiments conducted by using "proximity index" which is strategically computed between a pair of vectors, one from a known, and the other, from an unknown. Both experiments were conducted in a closed set trial. The Voice database contained 49 randomly selected speakers, each speaker providing 5 samples of 30 second long contextually unbounded (text-independent) and spontaneous speech materials. These speakers were recorded in two sessions separated by a period of about two months.

The speech samples recorded in the first session are designated as 'unknowns', and ones recorded in the second session, designated as 'knowns'. Each speaker was represented by a vector that is comprised of a set of 14 (5 from time and 9 from IDS) parameters. A proximity index is computed from a pair of these vectors.

Voice Identification

The voice identification process as defined in this project arranges all the unknown speakers on a continuous line after they are rank ordered according to the proximity index values which are computed between the given known and all the unknown speakers. The utility of the voice identification process may be viewed not so much to discriminate the given pair of speakers as to place them into a one dimensional space reflecting their statistical positions relative to others. This type of process

would provide us with an assurance that the system can find the unknown speaker if he is included in the database.

Table 5.1 shows the results of voice identification experiments with 49 known and 49 unknown speakers by using the proximity index. The performance was tested under 15 rank allowances, and for each rank allowance condition, there was a total of 60,025 possible comparisons of the "proximity indices". As shown in the table, the correct identification performance progressively increases as the number of rank allowance increases: 80 % for rank allowance of 0, 85% for rank allowance of 1, 91% for rank allowance of 2, 95% for rank allowance of 7, and reaches 99% range for rank allowance of 15.

It was evident that even if a false identification occurred under the most stringent rank allowance of 0[1], although the table does not show it directly, a correct unknown (or more than one in most cases) was always found very close in line.

It can be equivalently expressed that the system, within the limitation of our present database (49x5=245 unknowns), needs to draw 0.82% of the database (2/245) to achieve 85% correct identification rate, 1.22% of the database (3/245) to achieve 91%, 3.26% of the database (8/245) to achieve 95%, and 6.12% of the database (16/245) to reach 99% correct identification rate.

---

[1]Under this condition, for the unknown to be correctly identified, the proximity index measured between himself and a given known (actually the same as the unknown) must be the smallest value among other proximity indices measured between the unknown and the remaining known speakers.

In short, based on our current database, the system must draw a pool of seven unknown (reference) speakers from the database to be 99% certain that this pool will include the questioned voice who is being sought as the same speaker as the known.

The results indicate that "proximity index" computed from a set of parameters (five from time domain and nine from frequency domain) can distinguish the known and unknown speakers with a high success rate.

Rank Allowance	Voice Identification Rate Hit	Miss	No Dec.	Rate(%)
0	170	46	20	79.5556
1	192	33	20	85.3333
2	200	25	20	88.8889
3	205	20	20	91.1111
4	208	17	20	92.4444
5	209	16	20	92.8889
6	210	15	20	93.3333
7	213	12	20	94.6667
8	216	9	20	96.0000
9	218	7	20	96.8889
10	220	5	20	97.7778
15	223	2	20	99.1111
20	224	1	20	99.5556
25	224	1	20	99.5556
30	224	1	20	99.5556

Table 5.1 Results of Voice Identification Experiments.

Next, we will discuss the results of the voice verification experiments in which the various threshold values along the continuum of the proximity index are applied to test the performance of our voice verification procedures.

Voice Verification

Unlike the voice identification process in which the relative magnitude of the proximity index is the key to arrange speakers in to a one dimensional similarity line, the voice verification process requires the absolute magnitude of some sort of similarity measurements. In specific, we need to have a 'threshold', or 'cut off' value to determine whether a given pair of speakers are verified (accepted as same or match), or not verified (rejected as different, or no match).

The results of the C.A.V.I.S. voice verification experiments are presented by using a classical technique commonly known as the receiver operating characteristic (ROC). The ROC is a graph of the two kinds of probabilities; (1) probability of 'match' decisions when actually it is true, and (2) probability of 'match' when it is false, plotted on a unit square coordinates. The ROC has been a suggested procedure in the field of voice identification (the National Academy of Sciences, 1979; Tosi, 1979) to evaluate the discriminating ability of the human voice examiner, or any other system used for voice identification. Here, we adopt the technique to evaluate the discriminating capability of proximity index applied in the C.A.V.I.S. voice verification procedures with 49 known and 49 unknown speakers.

Table 5.2 shows the results of the voice verification operations expressed in terms of $p(S|s)$, $p(S|d)$, $p(D|d)$, and $p(D|s)$, each computed by the different proximity index values. As it can be seen that $p(S|s)$, the probability 'match' decisions

when it is true, or correct verification, begins to show positive effects (0.0851) at the proximity index value of 0.8, and gradually reaches $p(S|s) = 0.9821$ at the proximity index value of 0.38, and $p(S|s)$ tapers off as the proximity index value gets smaller than 0.38.

run #	Prox. Index Vc	True Ident. p(S|s)	False Ident. p(S|d)	True Elim. p(D|d)	False Elim. p(d|s)	Total Attempts Ident.	Elim.
1	0.8000	0.0851	0.0006	0.9994	0.9149	12945	42180
2	0.7500	0.1041	0.0010	0.9990	0.8959	10367	44758
3	0.7000	0.1292	0.0017	0.9983	0.8708	8094	47031
4	0.6500	0.1638	0.0028	0.9972	0.8362	6044	49081
5	0.6000	0.2173	0.0042	0.9958	0.7827	4192	50933
6	0.5800	0.2494	0.0049	0.9951	0.7506	3492	51633
7	0.5600	0.2833	0.0059	0.9941	0.7167	2880	52245
8	0.5400	0.3277	0.0069	0.9931	0.6723	2319	52806
9	0.5200	0.3838	0.0082	0.9918	0.6162	1795	53330
10	0.5000	0.4428	0.0098	0.9902	0.5572	1346	53779
11	0.4800	0.5045	0.0114	0.9886	0.4955	1007	54118
12	0.4700	0.5586	0.0121	0.9879	0.4414	836	54289
13	0.4600	0.6127	0.0127	0.9873	0.3873	710	54415
14	0.4500	0.6718	0.0134	0.9866	0.3282	585	54540
15	0.4400	0.7340	0.0143	0.9857	0.2660	470	54655
16	0.4300	0.7816	0.0151	0.9849	0.2184	380	54745
17	0.4200	0.8136	0.0161	0.9839	0.1864	295	54830
18	0.4100	0.8318	0.0172	0.9828	0.1682	214	54911
19	0.4000	0.8652	0.0182	0.9818	0.1348	141	54984
20	0.3900	0.9545	0.0189	0.9811	0.0455	88	55037
21	0.3800	0.9821	0.0194	0.9806	0.0179	56	55069

Table 5.2 Results of voice verification experiments. The probabilities are expressed in terms of p(S|s) - the probability of true identification, p(S|d), the probability of false identification, p(D|d), the probability of true elimination, and p(D|s), the probability of false elimination, based on the various proximity index values, V_C, used as verification thresholds.

Figure 5.1 is the ROC curve illustrating the relationship between the two types of probability, one being p(S|s), the probability of the system calling a 'match' when it is actually true, and the other p(S|d), the probability of the system calling a 'match' when it is actually false. The solid curve was prepared from the experiment in which the system was allowed to exercise "no decision" when the IDS bands yielded poor stability. The broken line curve was prepared from the same experiment but with an exception: the system was not allowed to refrain from rendering the decision.

The solid line curve rises sharply to approach p(S|s)=0.9821 which corresponds to p(S|d)= 0.0194, whereas the broken line curve slowly reach p(S|s)=0.9425 which corresponds to p(S|d)=0.0418. The difference shows clearly that the system with "no decision" option allowed performs better in terms of its verification rate than the system without the option of "no decision". Furthermore, Figure 5.1 also indicates that, in order to reduce p(S|d) to 0.0, which is the ideal condition where we have no costly 'false' verification, p(S|s) must be shifted down to 0.0446. A simple way of interpreting these figures may be: In order for the system to maximize its ability (with no decision option given) to correctly verify the criminal's voice to p(S|s) = 0.98, there will be an accompanying stake of incorrectly verifying an innocent individual with p(S|d) = 0.0194.

Figure 5.2 is the ROC curve illustrating the relationship between the two types of probability, one being p(D|d), the probability of the system calling a 'Different' when it is actually true, and the other p(D|s), the probability of the system calling a 'Different' when it is actually false.

We feel that it is safe to say that C.A.V.I.S. is as effective in verifying the speakers as in eliminating the innocent speakers.

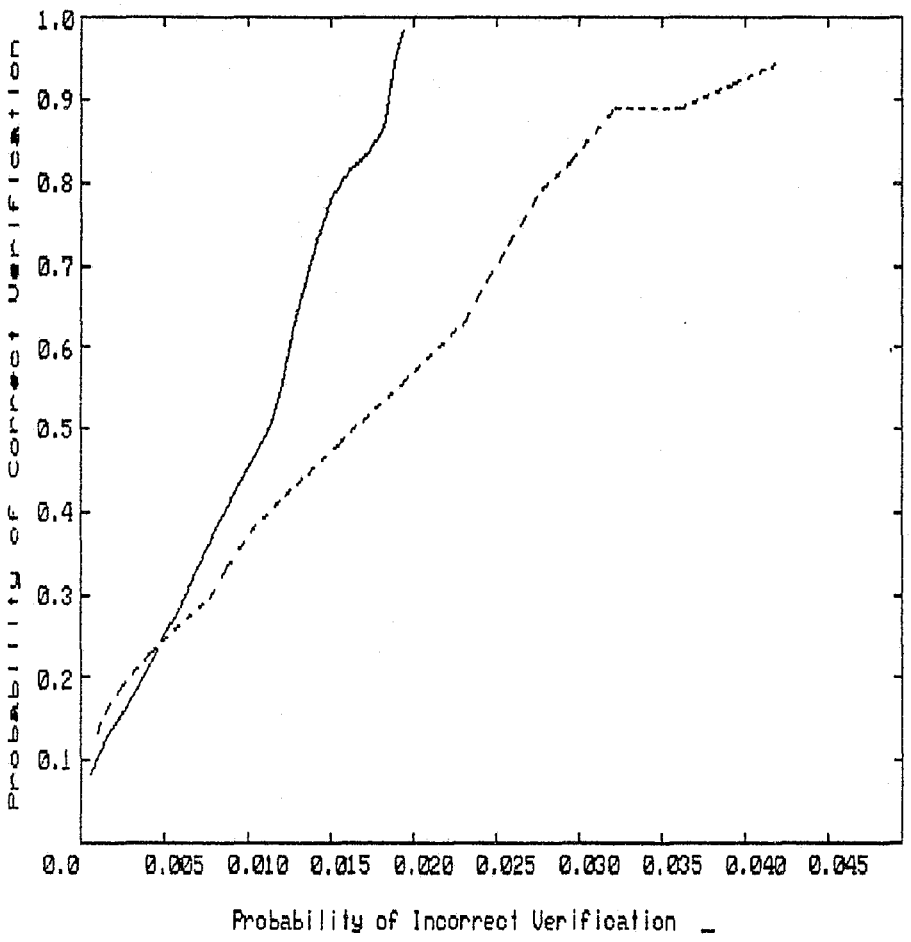


Figure 5.1 Probabilities of correct verification and incorrect verification. The solid line is ROC curve made from the verification experiments with "no decision" allowed. A correct verification occurred when the system declared a known and an unknown speakers are "Same" when actually it is true. An incorrect verification occurred when the system declared a known and an unknown speakers are "Different" when actually they are "different" speakers. The broken line is ROC curve made from the verification experiments with no "no decision" allowed.

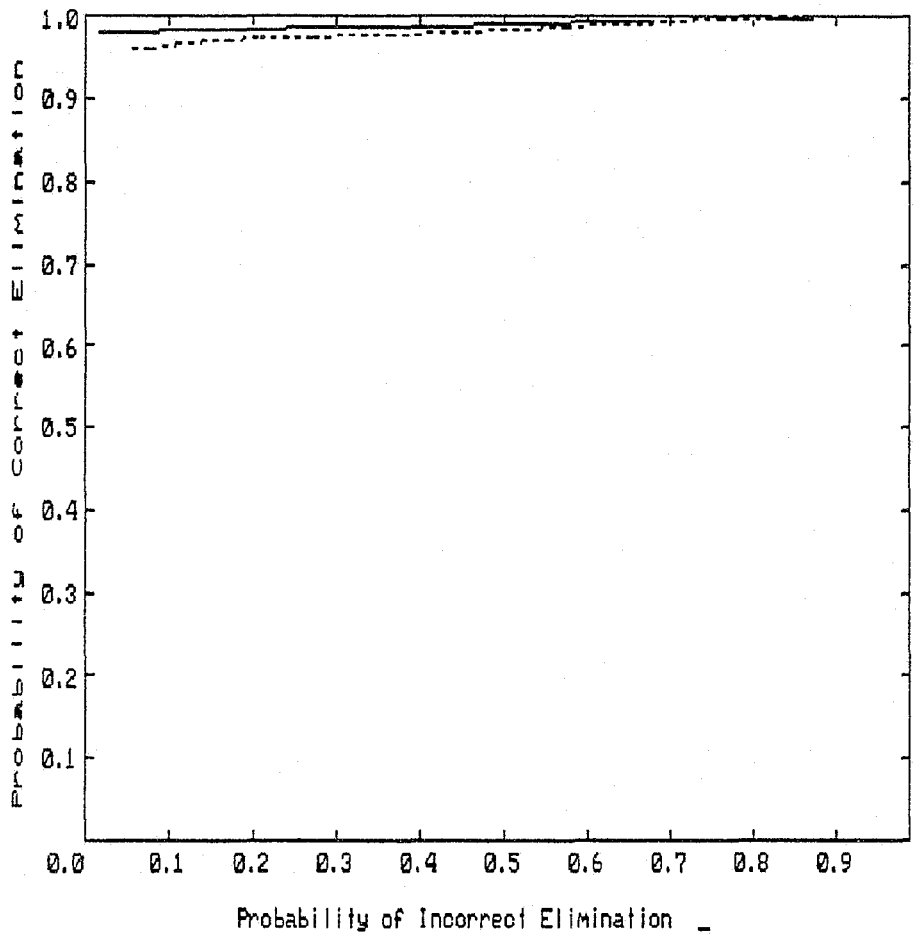


Figure 5.2 Probabilities of correct elimination and incorrect elimination. The solid line is ROC curve made from the verification experiments with "no decision" allowed. A correct elimination occurred when the system declared a known and an unknown speakers are "Different" when actually it is true. An incorrect elimination occurred when the system declared a known and an unknown speakers are "Different" when actually they are "same" speakers. The broken line is ROC curve made from the verification experiments with no "no decision" allowed.

6 CONCLUSIONS

The main goals of this project were threefold: (1) to establish a system that is free from the influence of the transmission and/or recording media, (2) to develop a system that works with text-independent voice data, and (3) to construct a system which can deliver the voice identification decisions objectively.

The problem of the adverse influence of the transmission and/or recording media due to the unknown response characteristics was intensively dealt with in the earlier stage of this project by the use of multiple transmission channels. This problem was approached from three angles. The first was related to the selection of a particular group of parameters that are not associated with a spectral output of speech which is subject to the response characteristics of the media. Pitch or fundamental frequency and the variety of derivatives measurements were selected and found to be the ideal parameters. All the measurements from the time domain described in this report belong to this class of parameters.

Intensity deviation spectrum (IDS) was investigated for its independence from the influence of the transmission medium and also the reliability in distinguishing the speakers. It was concluded that IDS can be made free from the influence and is reliable in distinguishing the speakers. It was found that contextually unbounded and spontaneously generated speech samples

lasting as long as 30 seconds can provide a sufficient amount of information for recognizing the individual's identity.

Although the system, as it stands now, is characterized as being interactive, thus inescapable from inclusion of some amount of subjectivity, there are many objective components found at the various stages.

The objective components in our system can be seen throughout the pre-processing stages, such as the determination of the individualized filter shape for each speech sample unit, analog to digital conversion, pause elimination, generation of spectra by FFT, computation of IDS and the estimates of the probability density functions of time domain parameters, and the process of computing the proximity index.

On the other hand, the major subjectivity in the system exists in three areas: (1) during editing of speech signal to remove pauses, whenever software automatic pause deletion program fails, (2) during the manual targeting of wavelet peaks, which yields the basis for all the time domain parameters, and (3) during the process of estimating three Weibull parameters, e0, v0, and k0, when the automatic process fails, the operator's interactive maneuver is required to optimize the data by deleting the outliers.

Nonetheless, it is important to note that the results are reproducible and procedures are repeatable for they are based on solid computer algorithms: the test re-test reliability is considered high, which is an essential aspect of the objectivity. We, therefore, believe that our third goal

'objective decision' has been fulfilled.

The key factor that emerged during the project to challenge the above mentioned problems simultaneously is the implementation of the strategic optimization of the parameter set for each pair of speakers to be compared. This implies that one particular set of parameters may be the best fit for a particular pair of speakers under comparison, but the same set may not work at all for another pair of speakers. The automated selection process of the optimum set of parameters for each individual pair of the speakers was improvised as described fully in the previous Chapter.

The results presented in the previous chapter clearly show the overall effectiveness of C.A.V.I.S. in distinguishing the speakers (in the voice identification process as well as in the voice verification process). The system in the voice identification process indicated a reasonably high success rate as long as it is allowed to pull out several voices that can be the candidates for the questioned voice. Since there is only one voice we are interested in drawing from the voice database, this type of process may appear to be irrelevant in a real situation. However, the utility of this process lies in the fact that the system can be shown to be sensitive, thus being able to classify a certain group of speakers according to their voice characteristics.

We took very cautious steps in every aspect of the experimentation during this project so that the voice data we analyzed would be as close to a realistic situation as possible.

In that sense, we tend to believe that a certain level of subjective components provided by the knowledgeable operator should participate in the process to ensure that appropriate voice data are analyzed. More importantly, this type of subjectivity is not likely to be the target of psychological bias, on the part of a system user, in reporting the final decision of the voice identification phase.

Toward the end of this project, we grew to sustain a notion of possible existence of the "separator" parameters and the "connector" parameters. Any given parameter from a specific person can be regarded as either "separator", or "connector", but not as both. We acknowledged that this notion is highly speculative, yet it appears to bear a significantly important future research topic in the field of forensic voice identification.

For example, during the process of defining a set of parameters, for a specific speaker, first choose those that indicate a high stability within the individual and discard those which exhibit random measurements: This can be done by taking the statistical measurements of the variability of each parameter within a given speaker. Next, select a set of parameters, from the same speaker, those which separate the speaker from the rest: This is achieved by means of comparing each parameter taken from two speakers against the estimated population distribution of the parameter. Then conduct a test to see whether or not these two speakers fall into the same region bounded by one standard deviation. Within an informally

constructed experimental design, we noted that having two speakers fall in such a bounded region does not necessarily mean that they are matched, but simply implies that the speakers become indistinguishable by that parameter. On the other hand, when two speakers fall farther apart outside of this one standard deviation region, they are separated with a high degree of certainty.

In our research project, the parameters measured from the time domain are treated under this very concept of 'separator' or 'connector' for distinguishing the speakers.

7 FUTURE IMPLICATIONS

Our intent is to contribute to the crime investigation process by using the methodologies and findings which have been integrated into C.A.V.I.S.. We feel strongly about the future contribution of our system for voice identification to the law enforcement community supported by the new techniques as described below.

The schemes and ideas promoted in this project include: speech parameter extraction techniques, the use of speaker-dependent stability measures, the development of a technique for statistical processing of these stability measures, implementation of separator vs. connector concept, and a strategy to reduce the multidimensional vector sets into a single 'proximity index', and finally, all of the above techniques have been systematically and conveniently integrated into a solid reliable computerized working system.

This research project is considered to be unique in a sense that it has been conducted at the very site of the law enforcement environment where the product is most needed. In fact, during the four year long project, the research staff have been constantly exposed to a variety of real criminal voice cases, which have been handled by the conventional method, aural-spectrographic method, of voice identification. Under such circumstances, mainly because of the requirements by C.A.V.I.S., and partly because of the need for enhancement of some rudimental processing aspects in the conventional aural-spectrographic voice identification method, it was only a natural course that useful

by-products emerged. By-products of C.A.V.I.S., such as digital audio processing techniques for editing, filtering, searching for words, and real time frequency analysis, have been assimilated into the conventional process of voice identification. It has been confirmed that the same by-products can be applied effectively toward the analysis of general types of acoustic events generated and recorded during the course of real criminal actions. These events include gun shots, explosives, and sound generated by a tossed piece of galvanized pipe allegedly used at a murder scene, and so forth.

In relation to the existing methodology, we are urged to make a following note. What has been accomplished by the sophisticated statistical computation is not going to be a replacement of the conventional method of voice identification when text-dependent samples are available, but, rather to be a reinforcement working in a complimentary fashion with the existing methodology and technology of voice identification.

Human speech production is a complex and dynamic phenomenon requiring even more complex mechanisms of processing by the human brain. It is fair to say, for the system as it stands now, that C.A.V.I.S. is limited in its capability to capture the speaker dependent information embedded within the <u>semantic and linguistic</u> aspects of the voice. That type of discrimination still calls for the intervention by the experienced examiner through the critical listening and extraction of such information.

We reported that C.A.V.I.S. yielded 98% correct identification, and 2% incorrect identification (false

identification) when performed thoroughly algorithmically with a minimum amount of operator intervention. At this performance rate we feel very strongly that our system is ready to provide services upon request and contribute in the investigative process in which the identity of recorded voices bear evidential relevancy. In order to reduce this 2% error of false identification, the system needs to be provided in the future with more analytical capability and information processing strategy that parallels with the brain of the experienced human voice expert. For this provision, there is a realistic optimism exemplified in the survey report by the Federal Bureau of Investigation (Koenig, 1986). Koenig reported that the 2000 voice identification comparisons by the spectrographic technique conducted by the FBI examiners yielded a 0.31% false identification error rate and a 0.53% false elimination error rate.

Until a fully developed automatic computer system of voice identification is established for forensic use, this 'man-machine' interactive system appears to be the best direction to pursue to fight against the crime. The tool developed in this project is ready to aid the voice examiner in analyzing the increasing numbers of voice identification cases.

8 REFERENCES

Atal, B. S. 'Automatic recognition of speakers from their voices' in Automatic Speech & Speaker Recognition, N. Rex Dixon and Thomas B. Martin (ed.). IEEE Press, New York, 1978, pp. 349-364

Atal, B. S. 'Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification', J. Acoust. Soc. Amer., 1974, Vol. 55, pp. 1304-1312.

Bowie, G. 'Application of Extreme Value Statistics For Project C.A.V.I.S.', Memorandum submitted to the Los Angeles County Sheriff's Department, 31 August, 1986.

Atal. B. S. 'Automatic speaker recognition based on pitch contour', J. Acoust. Soc. Amer., 1972, Vol. 52, pp. 1687-1697.

Bunge, E. 'Automatic speaker recognition system AUROS for security systems and forensic voices identification' in Automatic Speech & Speaker Recognition, N. Rex Dixon and Thomas B. Martin (ed.). IEEE Press, New York, 1978, pp. 4124-420.

Digital Audio Corporation, "PDF 2048 user's manual", 6512 Six Forks road Ste. 203B, Raleigh, NC 27609-2946, March, 1986.

Doddington, G. R. 'A method of speaker verification', Paper presented at The Eightieth Meeting of the Acoust. Soc. Amer., 1970, Nov. 3-8, Houston, Texas.

Doddington, G. R. 'Speaker verification - Final report', Rome Air Development Center, Griffiss AFB, N.Y., Tech. Rep. April, 1974, RADC 74-179.

Furui, S. 'Cepstrum analysis technique for automatic speaker verification'. IEEE Trans. Acoust., Speech, and Signal Processing, April, 1981a, Vol. ASSP-2, No. 2, pp. 254-272.

Furui, S. 'Comparison of speaker recognition methods using statistical features and dynamic features', IEEE Trans. Acoust., Speech, and Signal Processing, 1981b, Vol. ASSP-29, No. 3, pp 342-350.

Furui, S., Itakura, F., and Saito, S. 'Talker recognition by longtime averaged speech spectrum', Electronics and Communications in Japan, 1972, 55-A, pp. 54-61.

Gumbel, E.J., Statistics of Extremes, Columbia University Press, N.Y., N.Y., 1958.

He, Q., and Dubes, R. 'An experiment in Chinese speaker identification', presented at 1982 IEEE Int'l Conf. Trans. Acoust., Speech, and Signal Processing.

Hunt, M. J., Yates, J. W., and Briddle, J. S. 'Automatic speaker recognition for use over communication channels', IEEE Int'l Conf. Record on Acoust., Speech, and Signal Processing. May 9-11, 1977, pp. 764-767.

Jain, A. K. and Dubes R. 'Feature definition in pattern recognition with small sample size', Pattern Recognition, 1978, Vol. 10, pp. 85-97.

Koenig, B.E. 'Spectrographic voice identification: A forensic survey', J. Acoust. Soc. Amer., 1986, Vol. 79(6), pp. 2088-2090.

Kinnison, R.R. 'Applied Extreme Value Statistics', MaCmillan Publishing company, New York, 1985.

Luck, J. E. 'Automatic speaker verification using cepstral measurements', J. Acoust. Soc. Amer., 1969, Vol. 46, pp. 102-1032.

Majewski, A. W., and Hollien, H. 'Cross correlation of long-term speech spectra as a speaker identification technique', Acustica, 1975, Vol, 34, pp. 20-24.

Markel, J. D., and Davis, S. B. 'Text-independent speaker recognition from a large linguistically unconstrained time-spaced data base', IEEE Trans. Acoust., Speech, and Signal Processing, February, 1977, Vol. ASSP-27, No. 1, pp. 74-82.

Markel, J. D., Oshika, B. T., and Gray, A. H. 'Long-term feature averaging for speaker recognition', IEEE Trans. Acoust., Speech, and Signal Processing, 1977, Vol. ASSP-25, pp. 330-337.

Naik, J.M. and Doddington, G.R. "Evaluation of a high performance speaker verification system for access control", Proc. IEEE Intl. Conf. Acoust., Speech, Sig. Processing, pp. 2392-2395, April, 1987, Dallas, Texas, USA.

Nakasone, H. and Melvin, C.A. "Computer Assisted Voice Identification System", Proc. IEEE Intl. Conf. Acoust., Speech, Sig. Processing, pp. 587-590, April, 1988, N.Y., N.Y., USA.

Nakasone, H. 'Computer voice identification method by using intensity deviation spectra and fundamental frequency contour', 1984, Unpublished Ph.D. dissertation, Michigan State University.

Noll, A. M. 'Cepstrum pitch determination', J. Acoust. Soc. Amer., 1967, Vol. 41, pp. 2932-309.

Paul, J. E., Rabinowitz, A. S., Riganati, J. P., Richardson, J. M. 'Development of analytical methods for a semi-automatic speaker identification system', 1975 Carnahan Conf. on Crime Countermeasures, 1975, pp. 52-64.

Tosi, O. Voice Identification: Theory and Legal Applications. University Press, Baltimore, 1979.

Tosi, O., Pisani, R., Dubes, R., and Jain A. 'An objective method of voice identification', Current Issues in the Phonetic Sciences, Harry & Patricia Hollien (ed.). In series of Current Issues in Linguistic Theory Vol. 9 in Amsterdam Studies in the Theory and Hearing of Linguistic Science IV., Amsterdam-John Benjamins B.V., 1979, pp. 851-861.

APPENDIX A

Mathematics for Computing the Area (Probability of Match of Two Speakers) Formed By Two Probability Density Functions Derived From the C.A.V.I.S. Time Domain Parameters

Let

$P_{(x)}$ = probability

$p_{(x)}$ = distribution function

$$P_{(x)} = \exp\left(-((x-a)/b)^c\right)$$

Then, $$\frac{dP_{(x)}}{dx} = \frac{-c}{b}\left(\frac{x-a}{b}\right)^{c-1} P_{(x)}$$

The distribution is

$$p_{(x)} = \frac{c}{b}\left(\frac{x-a}{b}\right)^{c-1} P_{(x)}$$

Integrate $P_{(x)}$ from $x = a$ to $x = \infty$

$$\int_a^{\infty} P_{(x)}\,dx = \int_a^{\infty} \frac{c}{b}\left(\frac{x-a}{b}\right)^{c-1} \exp\left(-\left(\frac{x-a}{b}\right)^c\right)dx$$

$$= \left[-\exp\left[-\left(\frac{x-a}{b}\right)^c\right]\right]_a^{\infty}$$

$$= 0 + \exp\left(-\left(\frac{a-a}{b}\right)^c\right) = 1$$

Integrate $P_{(x)}$ from $x = a$ to $x = x\_cross$

$$\int_a^{x\_cross} p_{(x)}\,dx = \left[-\exp\left(-\left(\frac{x-a}{b}\right)^c\right)\right]_a^{x\_cross}$$

$$= 1 - \exp\left(-\left(\frac{x\_cross - a}{a}\right)^c\right)$$

The above algorithm exemplifies an example when there is only one cross over (intersect, or probability of match) made by two probability density functions of one of our time domain parameters (one for a known and one for an unknown speaker). In the equations, a, b, and c represent three Weibull parameters, and throughout in this final report, they are denoted as e0 (threshold parameter), v0(characteristics value), and k0(shape parameter), respectively. (By courtesy of Dr. Glenn Bowie, 1989)

APPENDIX B

List of Major Program Names Developed For C.A.V.I.S.[1]

Program Names	Descriptions
ADA10240.EXE	A 12-bit Analog to Digital Conversion program with a 10240 Hz sampling rate. Initiated by automatic (external) trigger.
ADM10240.EXE	A 12-bit Analog to Digital Conversion program with a 10240 Hz sampling rate. Initiated by manual (internal) trigger.
ADA20480.EXE	A 12-bit Analog to Digital Conversion with a 20480 Hz sampling rate. Initiated by automatic(external) trigger.
ADM20480.EXE	A 12-bit Analog to Digital Conversion with a 20480 Hz sampling rate. Initiated by manual (internal) trigger.
AVGWAVE.EXE	To compute the averaged smoothed and the sorted wavelets from each voice sample.
CAV3D.EXE	To plot speakers dynamically onto three dimensional spaces based on the speaker specific parameter set.
CAVEXP2P.EXE	To perform voice identification and voice verificaton experiments.
CAVIS.EXE	The main driver program which integrates the C.A.V.I.S. major programs used in analog to digital, digital to analog conversions, filtering, editing, parameter extractions, and other signal processings.

[1] All the main programs listed above are coded in Microsoft "C", versions. 3.1 and 5.1. Function modules called by these main programs are not listed. A few functions are written in the Assembly language where the intensive computation is required, or where the speed of data transportation is critical during the ADC or DAC processes.

CHARTIT.EXE	To produce a hardcopy of sound file.
CKLEVEL.EXE	To calibrate the optimum input level of the analog signal before the ADC process.
COMPLEXT.EXE	To synthesize complex tones used during the debugging stage of the system software development.
IDSPROT.EXE	To compute the IDS spectra.
DELPAUSE.EXE	To remove pauses by the automatic method.
DISP-AVG.EXE	To plot the long-term averaged spectrum generated from the individual speaker sample.
DISP2P10.EXE	To plot simultaneously all the frequency domain parameters of the known and the unknown speakers.
DISPAFT.EXE	To display in a water fall mode a set of 1024 point FFT frames taken from a sample of a speaker.
DISPJOIN.EXE	To display the entire speech parameters both from the time and the frequency domain in the final graphic output format.
DISPLET.EXE	To display graphically the each individual wavelets segmented.
DISPPROT.EXE	To display graphically the IDS spectra.
DO375.EXE	To perform the numerical conversions from SD 375 data format to the DOS binary format.
DSPFFT.EXE	Perform a series of 1024 point FFT's by the use of a TMS 320 based Digital Signal Processor Board installed in the system computer
EDIT1024.EXE	To perform editing of the sound file. Uses the convenient graphic display combined with the instant DAC feature.
FFT1024.EXE	Perform a series of 1024 point FFT's by the "C" language written software.
FRVID.EXE	To compute the F-ratio statistics of the speech parameters.
GENAVGFD.EXE	To compute the averaged IDS shape from

the entire speaker set.

GEN-PDF.EXE	To generate the 512-tap convolution coefficient set that is used to set the PDF 2048 into the arbitrary shape such as a low or high pass for the evaluation of the system performance.
GET-AVG.EXE	To retrieve the 1024 data from the SD-375 for later use of computing the convolution coefficients.
IDSPROT.EXE	To plot the IDS spectra as many as 10 at a time, each IDS being displayed in its own color.
LOOPIT.EXE	To playback (DAC) the sound file for aural evaluation in a continuous mode.
MATCH.EXE	This program takes a table of speech files of the length between one second to any length that is limited by the maximum memory capacity of the storage medium in use. It conveniently facilitates the short-term memory aural analysis of the speech samples through a 12-bit DAC with a 10240 Hz sampling rate. It is not included as the required element of the C.A.V.I.S. voice identification and verification experiments, but has been used as a daily laboratory tool for the analysis of the actual voice cases.
MATRIX3W.EXE	To generate an N x N matrix of the probabilities of match computed between every possible combinations of the speaker samples along a given time domain parameter.
PACKIT.EXE	To concatenate the signals that are automatically segmented by the program delpause into a sound file.
PICKIT.EXE	To detect interactively (operator and software) the wavelets from the unsmoothed and unrectified sound file, and to store the addresses of the detected wavelets.
PICKSRT.EXE	To detect interactively (operator and software) the wavelets from the smoothed and rectified sound file, and to store the addresses of the detected wavelets.

PLAYIT.EXE	To playback a sound file.
POPEVK.EXE	To estimate the population values of the three Weibull parameters.
SCOPEIT.EXE	To display in a real time mode the input analog signal on the system computer monitor for the purpose of calibration.
SD-PDF.EXE	To generate the 512-tap convolution coefficient set that is used to set the PDF 2048 into the individualized shape for each speaker's sample.
SMOPROT.EXE	To smooth and normalize an IDS.
STEREO.EXE	To display graphically two sound files simultaneously, to playback the designated portion of either file, and to perform a 512 point FFT and display the results.
W1.EXE	To compute the three parameters of the Weibull function from the "wavelet intensity", (or the 3w1) C.A.V.I.S. speech parameter by the method of manual removal (with an optical mouse) of bad data from both the low and the high extreme ends.
W1AUTO.EXE	To compute the three parameters of the Weibull function from the "wavelet intensity", (or the 3w1) C.A.V.I.S. speech parameter by the method of automatic iterative removal of bad data from both the low and the high extreme ends.
W2.EXE	To compute the three parameters of the Weibull function from the "pitch", (or the 3w2) C.A.V.I.S. speech parameter by the method of manual removal (with an optical mouse) of bad data from both the low and the high extreme ends.
W2AUTO.EXE	To compute the three parameters of the Weibull function from the "pitch", (or the 3w2) C.A.V.I.S. speech parameter by the method of automatic iterative removal of bad data from both the low and the high extreme ends.
W4.EXE	To compute the three parameters of the Weibull function from the "auto corre-

	lation of successive wavelets", (or,3w4) C.A.V.I.S. speech parameter by the method of manual removal (with an optical mouse) of bad data from both the low and the high extreme ends.
W4AUTO.EXE	To compute the three parameters of the Weibull function from the "auto correlation of successive wavelets", (or, 3w4) C.A.V.I.S. speech parameter by the method of automatic iterative removal of bad data from both the low and the high extreme ends.
W5.EXE	To compute the three parameters of the Weibull function from the "variation of total wavelet intensity", (or 3w5) C.A.V.I.S. speech parameter by the method of manual removal (with an optical mouse) of bad data from both the low and the high extreme ends.
W5AUTO.EXE	To compute the three parameters of the Weibull function from the "variation total wavelet intensity", (or, 3w5) C.A.V.I.S. speech parameter by the method of automatic iterative removal of bad data from both the low and the high extreme ends.
W8.EXE	To compute the three parameters of the Weibull funaction from the "average energy distribution curve", (or 3w8) C.A.V.I.S. speech parameter by the method of manual removal (with an optical mouse) of bad data from both the low and the high extreme ends.
W8AUTO.EXE	To compute the three parameters of the Weibull function from the "average energy distribution curve", (or, 3w8) C.A.V.I.S. speech parameter by the method of automatic iterative removal of bad data from both the low and the high extreme ends.
WAVELETN.EXE	To compute the intermediate data sets for 3w1, 3w2, 3w4, 3w5, and 3w8 parameters concurrently while graphic displays are in progress. The program generates the output data files in ASCII format.
WEIBPOP.EXE	To compute and display the estimated population Weibull density functions.

WEIBPROB.EXE	To compute and display the intersect (probability of match) of two samples (one taken from an unknown, and the other, from a known speaker) represented by the pair of Weibull density functions of one of the time domain parameters.
WPROB10.EXE	To compute and display the intersect (probability of match) between all the possible combinations of Weibull density functions of 5 voice samples of each of the two given speakers.

APPENDIX C

Samples of five time domain parameters

FILE 1 = ts311x 3w1 Parameter = Normalized_Glotal_Intensities

	01	02	03	04	05
E0	22657.1719	20488.5059	20895.1172	22713.9785	22186.0039
V0	29688.6680	29829.9219	29828.4961	29573.2813	29484.0586
K0	5.5000	7.6875	7.7188	4.7813	5.8125
Weibull					
MEAN	29148.6602	29268.5234	29293.3086	28995.6738	28944.6016
VAR	1856818.0000	1828104.0000	1659948.8750	2246330.0000	1818408.25
S.D.	1362.6511	1352.0740	1288.3900	1498.7761	1348.4836
CVAR	0.0467	0.0462	0.0440	0.0517	0.0466
SKEW	-0.3181	-0.5137	-0.5158	-0.2225	-0.3534
KURT	2.9559	3.2905	3.2949	2.8470	3.0050
Descriptive					
MEAN	29149.4004	29270.0137	29295.3848	28995.6055	28945.6426
VAR	1771618.3750	1741265.1250	1601123.8750	2144848.7500	1724990.00
S.D.	1331.0215	1319.5701	1265.3552	1464.5302	1313.3888
CVAR	0.0457	0.0451	0.0432	0.0505	0.0454
SKEW	-0.3700	-0.5290	-0.5309	-0.2895	-0.4350
KURT	2.0501	2.2110	2.5441	1.9662	2.1043

FILE 2 = ts311y 3w1 Parameter = Normalized_Glotal_Intensities

	01	02	03	04	05
E0	24249.5430	24511.9570	21270.5117	24108.7266	21826.0215
V0	29249.9922	29137.9941	30035.4238	29429.1621	30146.3496
K0	3.5234	3.1797	7.1250	3.8906	6.2500
Weibull					
MEAN	28750.2793	28654.0098	29477.0566	28923.6211	29562.2891
VAR	2004393.3750	2041696.0000	1839833.1250	1918412.7500	2084222.12
S.D.	1415.7660	1428.8793	1356.4044	1385.0677	1443.6835
CVAR	0.0492	0.0499	0.0460	0.0479	0.0488
SKEW	0.0192	0.1123	-0.4731	-0.0647	-0.3980
KURT	2.7133	2.7151	3.2086	2.7378	3.0741
Descriptive					
MEAN	28748.2500	28652.1152	29479.5137	28923.0879	29565.7402
VAR	1941193.0000	1974012.8750	1769301.0000	1848970.6250	1947324.12
S.D.	1393.2670	1404.9957	1330.1508	1359.7686	1395.4656
CVAR	0.0485	0.0490	0.0451	0.0470	0.0472
SKEW	-0.0671	0.0440	-0.4619	-0.1161	-0.4021
KURT	2.1977	2.0387	2.4498	2.0835	2.0509

APPENDIX C (Continued)

Program E:\PRNW10.EXE Friday, 29 Sep 1989 - 10:58:43.78

FILE 1 = ts311x 3w2 Parameter = Fundamental_Frequency

	01	02	03	04	05
E0	79.5969	63.6851	94.9284	83.7398	86.2383
V0	159.8024	163.2805	138.3546	139.8305	134.3863
K0	3.8906	2.7734	1.9883	1.6602	2.3203
Weibull					
MEAN	152.1814	152.3382	133.4182	133.8727	128.8979
VAR	435.9698	1195.0221	409.1013	962.0613	380.8488
S.D.	20.8799	34.5691	20.2263	31.0171	19.5153
CVAR	0.1372	0.2269	0.1516	0.2317	0.1514
SKEW	-0.0647	0.2471	0.6390	0.9022	0.4444
KURT	2.7378	2.7677	3.2601	3.8730	2.9527
Descriptive					
MEAN	151.0349	150.9860	132.3643	132.5053	128.0031
VAR	419.3026	1119.2240	386.3436	845.9623	363.8753
S.D.	20.4769	33.4548	19.6556	29.0854	19.0755
CVAR	0.1356	0.2216	0.1485	0.2195	0.1490
SKEW	-0.0203	0.1513	0.4988	0.6179	0.3964
KURT	3.1325	2.2197	2.9759	2.1504	2.7210

FILE 2 = ts311y 3w2 Parameter = Fundamental_Frequency

	01	02	03	04	05
E0	84.9856	94.5377	92.0400	95.6651	98.7871
V0	148.2320	133.1453	145.3337	137.3328	137.6367
K0	3.5703	2.1250	2.3086	1.5039	1.6387
Weibull					
MEAN	141.9519	128.7301	139.2557	133.2687	133.5462
VAR	313.5562	286.4622	470.7950	648.7505	473.6034
S.D.	17.7075	16.9252	21.6978	25.4706	21.7624
CVAR	0.1247	0.1315	0.1558	0.1911	0.1630
SKEW	0.0077	0.5521	0.4504	1.0674	0.9230
KURT	2.7152	3.1076	2.9603	4.3752	3.9313
Descriptive					
MEAN	140.9141	127.8341	137.9062	131.9510	132.4043
VAR	298.2512	273.4631	420.8198	567.5229	433.7365
S.D.	17.2700	16.5367	20.5139	23.8227	20.8263
CVAR	0.1226	0.1294	0.1488	0.1805	0.1573
SKEW	0.0541	0.5805	0.3014	0.9240	0.7537
KURT	2.9785	3.5108	1.9641	3.5009	2.9760

APPENDIX C (Continued)

Program E:\PRNW10.EXE Friday, 29 Sep 1989 - 10:58:48.39

FILE 1 = ts311x 3w4 Parameter = Corr_Succ_Wavelets

	01	02	03	04	05
E0	995.8730	998.0584	999.4634	1000.0118	992.9894
V0	1054.3058	1069.1749	1062.3230	1095.9496	1091.0413
K0	1.2031	1.3594	1.2578	1.0391	1.1914
Weibull					
MEAN	1050.8015	1063.1921	1057.9275	1094.4802	1085.3953
VAR	2102.7480	2347.6687	2188.9890	8269.3271	6063.4858
S.D.	45.8557	48.4527	46.7866	90.9358	77.8684
CVAR	0.0436	0.0456	0.0442	0.0831	0.0717
SKEW	1.5151	1.2555	1.4160	1.8884	1.5377
KURT	6.2065	5.0597	5.7405	8.2797	6.3177
Descriptive					
MEAN	1050.4858	1062.9525	1057.5776	1093.6901	1084.7631
VAR	1980.1344	2249.2109	2066.2898	7685.3330	5644.1934
S.D.	44.4987	47.4258	45.4565	87.6660	75.1278
CVAR	0.0424	0.0446	0.0430	0.0802	0.0693
SKEW	1.2643	1.0921	1.1949	1.5569	1.2648
KURT	3.9924	3.5610	3.8785	5.1141	3.8528

FILE 2 = ts311y 3w4 Parameter = Corr_Succ_Wavelets

	01	02	03	04	05
E0	993.2094	989.7971	1007.8661	1000.4515	996.4571
V0	1066.7833	1140.9352	1088.2687	1072.0464	1074.1483
K0	1.2578	1.1367	1.0234	1.2070	1.3281
Weibull					
MEAN	1061.6387	1134.1312	1087.5076	1067.6971	1067.9124
VAR	2998.8169	16193.1445	6056.5181	3132.0676	2951.2512
S.D.	54.7615	127.2523	77.8236	55.9649	54.3254
CVAR	0.0516	0.1122	0.0716	0.0524	0.0509
SKEW	1.4160	1.6506	1.9317	1.5077	1.3020
KURT	5.7405	6.9000	8.5538	6.1704	5.2476
Descriptive					
MEAN	1060.9895	1132.6504	1086.6530	1066.8909	1067.1355
VAR	2743.9182	14742.5215	5508.9697	2834.8247	2615.0359
S.D.	52.3824	121.4188	74.2224	53.2431	51.1374
CVAR	0.0494	0.1072	0.0683	0.0499	0.0479
SKEW	1.1063	1.3349	1.6432	1.1823	0.9899
KURT	3.4460	4.0949	5.5306	3.7464	2.8148

APPENDIX C (Continued)

Program E:\PRNW10.EXE Friday, 29 Sep 1989 - 10:58:53.28

FILE 1 = ts311x 3w5 Parameter = Normalized_Glotal_Shimmer

	01	02	03	04	05
E0	188.1446	173.0270	200.2955	155.4339	172.0871
V0	795.1268	755.9966	775.0175	827.5623	758.7400
K0	1.4199	1.2773	1.2988	1.3750	1.4551
Weibull					
MEAN	740.1666	713.4104	731.1919	769.8192	703.7613
VAR	155484.4375	181640.2969	169889.1250	204474.4844	137840.812
S.D.	394.3152	426.1928	412.1761	452.1886	371.2692
CVAR	0.5327	0.5974	0.5637	0.5874	0.5275
SKEW	1.1717	1.3829	1.3478	1.2331	1.1265
KURT	4.7399	5.5927	5.4402	4.9719	4.5773
Descriptive					
MEAN	737.9327	710.8639	728.3773	766.8989	701.3884
VAR	148999.0938	173552.7031	162551.9063	195759.8906	132423.156
S.D.	386.0040	416.5966	403.1773	442.4476	363.8999
CVAR	0.5231	0.5860	0.5535	0.5769	0.5188
SKEW	1.0170	1.1874	1.1564	1.0381	0.9684
KURT	3.2492	3.7495	3.9382	3.3973	3.2808

FILE 2 = ts311y 3w5 Parameter = Normalized_Glotal_Shimmer

	01	02	03	04	05
E0	198.7635	258.3114	144.5535	168.7787	280.4189
V0	873.7714	1066.9147	849.1837	838.5959	771.9771
K0	1.3809	1.0762	1.7188	1.3906	1.2363
Weibull					
MEAN	815.3454	1044.3638	772.7999	779.9177	739.4171
VAR	204312.5156	534299.3750	141838.5625	198102.7969	139406.890
S.D.	452.0094	730.9579	376.6146	445.0874	373.3723
CVAR	0.5544	0.6999	0.4873	0.5707	0.5050
SKEW	1.2249	1.7918	0.8481	1.2113	1.4537
KURT	4.9400	7.6939	3.7281	4.8879	5.9139
Descriptive					
MEAN	811.7903	1038.3081	770.4039	776.1975	734.9787
VAR	194357.3594	502868.0938	136839.1875	187000.9063	129614.492
S.D.	440.8598	709.1320	369.9178	432.4360	360.0201
CVAR	0.5431	0.6830	0.4802	0.5571	0.4898
SKEW	1.0378	1.6002	0.7111	1.0171	1.2394
KURT	3.6058	6.0751	2.7310	3.2966	4.2145

APPENDIX C (Continued)

Program E:\PRNW10.EXE Friday, 29 Sep 1989 - 10:58:58.17

FILE 1 = ts311x 3w8 Parameter = Energy_Distribution

	01	02	03	04	05
EO	324.0961	361.5950	251.9547	352.6517	312.7665
VO	1116.5532	1155.4116	1102.8757	1123.3718	1117.1232
KO	1.6641	1.7617	1.8125	1.7344	1.7539
Weibull					
MEAN	1032.2529	1068.3148	1008.4199	1039.4347	1029.0492
VAR	191141.2969	171615.4688	186757.0000	166737.0938	177718.609
S.D.	437.1971	414.2650	432.1539	408.3345	421.5669
CVAR	0.4235	0.3878	0.4285	0.3928	0.4097
SKEW	0.8985	0.8107	0.7686	0.8343	0.8174
KURT	3.8627	3.6336	3.5330	3.6927	3.6501
Descriptive					
MEAN	1029.0122	1065.4376	1006.2689	1036.5476	1026.1066
VAR	180086.5000	163098.6094	179005.4688	158756.3125	169110.718
S.D.	424.3660	403.8547	423.0904	398.4424	411.2307
CVAR	0.4124	0.3791	0.4205	0.3844	0.4008
SKEW	0.7202	0.6528	0.6434	0.6750	0.6534
KURT	2.5568	2.5726	2.5619	2.6738	2.5886

FILE 2 = ts311y 3w8 Parameter = Energy_Distribution

	01	02	03	04	05
EO	264.3383	308.5601	342.9292	269.4112	351.4659
VO	1120.0905	1109.8516	1117.4316	1145.6213	1095.3193
KO	1.7734	1.7422	1.6797	1.8672	1.6387
Weibull					
MEAN	1025.9232	1022.3918	1034.5601	1047.3958	1016.9997
VAR	196914.1563	178665.3750	179245.7344	187195.8438	173626.828
S.D.	443.7501	422.6883	423.3742	432.6613	416.6855
CVAR	0.4325	0.4134	0.4092	0.4131	0.4097
SKEW	0.8008	0.8275	0.8838	0.7258	0.9230
KURT	3.6094	3.6754	3.8225	3.4367	3.9313
Descriptive					
MEAN	1022.8646	1019.4186	1031.4552	1044.5837	1013.8563
VAR	186499.4688	169834.1250	169467.0313	178119.1250	164381.062
S.D.	431.8558	412.1094	411.6637	422.0416	405.4394
CVAR	0.4222	0.4043	0.3991	0.4040	0.3999
SKEW	0.6176	0.6518	0.7055	0.5504	0.7438
KURT	2.3919	2.5511	2.5702	2.3355	2.7131