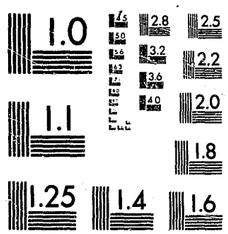
If you have issues viewing or accessing this file contact us at NCJRS.gov.

National Criminal Justice Reference Service

ncjrs

This microfiche was produced from documents received for inclusion in the NCJRS data base. Since NCJRS cannot exercise control over the physical condition of the documents submitted, the individual frame quality will vary. The resolution chart on this frame may be used to evaluate the document quality.



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

Microfilming procedures used to create this fiche comply with the standards set forth in 41CFR 101-11.504.

Points of view or opinions stated in this document are those of the author(s) and do not represent the official position or policies of the U. S. Department of Justice.

National Institute of Justice
United States Department of Justice
Washington, D.C. 20531

DATE FILMED

2/23/82

010 = Voice Recognition Using Color Encoded Voiceprints

030 = ~~Dr.~~ ~~Lester~~ A. Gerhardt

Associate Professor
Systems Engineering Division
040 = Rensselaer Polytechnic ~~Institute~~
Troy, New York 12181

041 = NILECJ

050 = NI 70-065-ME-9

060 = NCJRS 21

9892

U.S. Department of Justice
National Institute of Justice

This document has been reproduced exactly as received from the person or organization originating it. Points of view or opinions stated in this document are those of the authors and do not necessarily represent the official position or policies of the National Institute of Justice.

Permission to reproduce this copyrighted material has been granted by

Public Domain / LEAA

to the National Criminal Justice Reference Service (NCJRS).

Further reproduction outside of the NCJRS system requires permission of the copyright owner.

Table of Contents

Section	Title	Page
I.	Preface	3
II.	Summary	4
III.	Technical Discussion	5
	a. Introduction	5
	b. Background and Past Research	9
	c. Systems Description and Procedures	17
	d. Experimental Results	28
	e. A Real Time System Concept	45
	f. Conclusions	62
IV.	Professional Recognition and Publicity	65
V.	References	67

I Preface

This report describes a primarily experimental program concerning voice recognition using color encoded voiceprints. The work was performed at Rensselaer Polytechnic Institute, Troy, N.Y. by Dr. Lester A. Gerhardt, Associate Professor, Systems Division. The work was sponsored by the Department of Justice, of the U.S. Law Enforcement Assistance Administration, Washington, D.C., under its pilot grant program. A $10,000 grant, NI70-065-PG-9, was awarded for the period June 1971 to September 10, 1971. The financial support of the LEAA is gratefully acknowledged.

II. Summary

This report describes the results and conclusions of a primarily experimental study concerned with the generation, use, and evaluation of color encoded speech spectrograms or "voiceprints". Although the use of conventional "voiceprints" for suspect identification has been severely criticized recently by several individuals,[2,7] a belief that characteristic information or features are contained within the Time Varying Spectral Display (TVSD) of an individuals speech signal continues. The basic hypothesis of this study is that given the speech spectrogram as a means of identification, improvement over conventional spectrograms may be obtained using color encoding.

Following an introduction and a discussion of related research, the report first describes the electro optical system used to transform a conventional spectrogram to a color encoded display. Inputs to the system consisted of speech spectrograms produced primarily on a Kay sonograph. The output display appears in real time, on line with respect to the input on a TV type color display. Basically, the system transforms different density levels to different colors. The range and values of this transformation are operator controllable. A variety of digital encoding procedures are also possible.

III. Technical Discussion

a. Introduction

At the recent Seventh International Congress of Phonetic Sciences held this summer, Prof. Morris Halle of the Massachusetts Institute of Technology, a noted linguistic scholar, in referring to voice identification techniques stated "Such techniques should remain within the realm of scientific investigation".[7] It is in that context that this study was undertaken.

The overall goal of the study is to formulate and test methods for improving the display of conventional speech spectrograms (commonly called "voiceprints") with particular emphasis on color enhancement techniques. The basic hypothesis is that given a conventional speech spectrogram as a means of suspect identification, a color enhanced display will improve their readability. The work, therefore, is not initially concerned with proving the usefulness of the spectrogram as compared to other methods of identification.

Basic tests performed on hundreds of observers show that the spectrograms, or time varying spectral display (TVSD), is advantageous as a display in permiting an operator to extract certain periodic and aperiodic features. It is superior to a time domain representation, an instantaneous spectral display, or an acoustic signature in these respects. The advantages gained by using a color encoded TVSD are 1) the reduction of time needed to perform the feature extraction and classification and 2) the increased ease of readability. However, the color display does not lead to a significant reduction in the probability of error of classification.

The color display also serves to emphasize differences, for example, between spectrograms of the same speaker saying the same word, in fact, from the same recording as a function of the dispaly parameters. It also enhances differences between the same speaker saying the same word on different occasions etc. for the same display parameters. In this regard it again[2,7] leads to a questioning of the credibility of the basic conventional spectrogram. Therefore, in keeping with the idea of directing this study towards scientific investigation, the basic means of generating a TVSD was reviewed.

As a result, an improved real time on line system concept using a digital implementation and a Fast Fourier Transform and color encoded display was developed. With the increased flexibility, resolution, accuracy, and the ability to produce a speech spectrogram real time on line, this system offers the potential of a more realistic evaluation of improved speech spectrograms as a means of suspect identification. It is intended to continue this research but in the direction of developing a more efficient and useful system for generating speech spectrograms rather than with continued research using the existing approaches and data.

Thus this study, in addition to demonstrating the need for and capabilities of color encoded displays for reading speech spectrograms, has emphasized the need for research in this field towards more basic understanding of the speech signal and more effective means of displaying its essential characteristics. These characteristics must be present since both speaker and word identification are a regular everyday occurrence by individuals, and techniques for extracting and displaying them should be studied. It does not appear that a conventional spectrogram is the best way.

The results fall into three categories. The first is a comparative study of the effectiveness of different types of displays for detecting simple features. Using the simple two hypothesis problem of detecting a signal (in the acoustic range) in noise, the following displays were compared a) a time domain representation b) an instantaneous spectrum c) listening to the signal d) a TVSD and e) a color encoded TVSD. The effectiveness of the displays were measured using human observers and the results indicated that the most effective display was the TVSD and the effectiveness of the others appeared in the order as listed above. The results tended to substantiate the general use of a TVSD for data display.

The second major result was establishing the specific advantages of a color encoded display. A color encoded spectrogram was found to be easier and faster to read and interprete, but did not provide an extensive improvement in reducing errors of classification. As an example, for a S/N of -10 db the color encoded display results in an average of 25% less time required to arrive at a decision with the same level of confidence. Improvement in reducing classification errors was not as large as expected previously because the tests revealed the observer viewed the color encoding as a preprocessing of the information which thereby lowered the processing requirements on him, to which he accomodated.

Finally, hundreds of voiceprints were made and color encoded. The previously mentioned results were substantiated on actual data. The color encoded "voiceprints" were easier to read, analyzed in a shorter time, but were not classified with a substantially lower error rate (although a slight improvement was noted). Upon more detailed study, it became apparent that the techniques being used were limited by the relatively poor quality of the original data, specifically its limited dynamic range. Moreover, difficulties with conventional voiceprints reflecting their large variability and non-uniformity became more obvious with the color encoded approach. As such, since the techniques under study were being limited by the nature of the raw data, some effort was directed at considering new improved methods of generating speech spectrograms.

As a result, a systems concept was formulated for a digital real time, on line processor using a TV color encoded display and the feasibility was investigated. It is anticipated that development of this system would provide a means for studying and evaluating characteristics of voiceprints and lend to a better understanding of them for use in suspect identification. Further, this system can be implemented for field use.

In conclusion, it is felt that before the use of speech spectrograms is considered a dead issue, further research is required in the direction of developing a more effective means of generating speech spectrograms than is presently available to permit a more realistic evaluation. As a result of this study, the final system should definitely include a color enhanced display of the type described in this report.

b. Background and Past Research

Based on the relative ease with which individuals recognize and interpret speech both with respect to words and speakers, a large area of research has for some years been devoted to the extraction and display of essential characteristics of speech patterns for subsequent recognition by individuals or automatic means.[1,3,4,5,6] There has always been the underlying hope that speech characteristics are representative of an individual and that they can be used for identification purposes.[1,5]

Objection to this usage of interpreting voice characteristics became most severe[2] following the case of the People vs. Edward Lee King in California in which the State claimed to prove that the recorded voice of an unidentified youth who acknowledged setting fire to stores, was the same as that of the suspect, King, who was being held on another charge. A "voiceprint" was used for this purpose, and L. G. Kersta, sought to establish their similarities. Subsequent to this case, a great deal of conflicting views concerning the use of speech characteristics for suspect identification purposes have emerged.[5] To the author's knowledge, means are continually being sought to provide voice characteristics as admissable evidence in court for purposes of suspect identification at least in the states of Minnesota and New York. As such, the use of voice characteristics is still very much an issue in the field of Criminology and Law Enforcement.

However, it is not the original intent of this study to establish conclusively the use of voice characteristics as a means of suspect identification. In a scientific view, the use of color encoding of a display of voice characteristics was explored to determine more effective means of displaying for operator interpretation the available information. The type of display that was used is the speech spectrograms, or time varying spectral display TVSD, conventionally termed "voiceprint". It is this display of voice characteristics that was used in all referenced studies previously cited above. The remainder of the Section reviews the methods available to generate this information and some basic theory.

Whether listening to Sonar echoes, speech, or heart sounds, the hearing sense is able to delineate structural aspects (signal) in the presence of superfluous noise. The sense recognizes auditory patterns much the way the eye recognizes visual patterns. The time varying spectrum might be viewed as a transformation from an acoustical pattern to a visual pattern which retains and enchances the information bearing components of the sound. The justification for using the particular transformation that follows reflects back to the fundamental notions of a Fourier Series. Given a periodic signal, it is well known that a signal may be decomposed into a linear sum of sinusoids, where the coefficients are the information bearing elements containing the amplitude and phase information associated with each sinusoid. A similar transformation of information is obtained when the Fourier Transform is used for aperiodic signals. Unfortunately, the signals of concern here are generally non-stationary, and as such a single spectrum, usually defined as the magnitude of the Fourier Transform of the signal, [hereafter just designated f(t)] does not exist.

Thus, in accordance with the quasi steady state nature of such a signal f(t) the Fourier Transform is defined as a function of two variables ω and τ as

$$F(\omega,\tau) = \int_{\tau-\tau_m}^{\tau} f(t)\, w(t,\tau)\, e^{-j\omega(t-\tau)}\, dt$$

The integral is taken over the limits $\tau-\tau_m$ to τ rather than the standard infinite limits. These limits in conjunction with a weighting function, $w(t,\tau)$, exactly specify how much of the total signal should be used to generate the transform at the instant τ and how it should be weighted.

The role of a data window can be likened to the railroad passenger, seated by a window and facing to the rear. The window frame permits the viewer to see a panorama of present landmarks abreast of the train as well as some of those that have been passed.

The inferences or conclusions drawn by the observer will depend, in many cases, on his memory. The mental record would tend to discount, or forget, past events. Thus, the observer who must rely on the retention of data in his human memory would tend to weight his conclusions in favor of present or recent events. In an extreme case, where the observer can only see the present values and does not enjoy the benefit of a memory, his estimate of any attribute of the phenomena will be as variable as the data itself.

Data windows are therefore characterized by the amount of data observed as well as the "transmissivity". The length of data available for inspection is directly related to the memory capacity; the "transmissivity" depends on the particular discounting schedule associated with the memory system. On a clear day, the railroad passenger would observe the passing scene through a window that was essentially rectangular, while on a hazy day, the observation would be through a graduated window with decreasing transmissivity as events receded into the past.

If the weighting function is selected as exponential, for example, (implying that the signal past is of exponentially decreasing importance) this "short time" spectrum becomes

$$F(\omega,\tau) = \int_{\tau-\tau_m}^{\tau} f(t)\, e^{-\alpha(\tau-t)} e^{-j\omega(t-\tau)} dt = f(t) * e^{(-\alpha+j\omega)t}$$

An alternate view of the window or weighting function is that it reduces the error of a finite Fourier sum by modifying the coefficients of the series. This is done by multiplying the Fourier series coefficients by the window function. A multiplication in the time domain is a convolution in the frequency domain. The frequency content of a typical window function is primarily concentrated in the main lobe. It smooths out the sharp transitions of the magnitude function. The small sidelobes cause a greater stopband attenuation. Several existing window functions are presently used. Some of the more prominent ones are the Fejer, Hamming, Lanczos windows.

For any of the above, the short time power spectrum is defined as the magnitude squared of the quantity $F(\omega,\tau)$ and corresponds to the spectrum of a selected portion of the total signal weighted by the appropriate window function. The resulting spectrum is obviously a function of τ_m, which corresponds to the portion of the signal of interest. Thus, the idea of a short time spectrum leads directly to a three dimensional representation of the signal, where the power spectrum for a given τ_m and $w(t,\tau)$ is plotted as a function of frequency and τ. For any specified value of τ, the cross section is just the conventional spectrum defined for the last τ_m seconds of the signal.

The effective width τ_m of the window selected, for any given shape, is not arbitrary and does reflect in the frequency resolution obtainable in the transform domain. For the rectangular or gate window, for example, the frequency resolution can be improved by widening the window, i.e., using more of the signal history permits evaluating lower frequency components. In other words the Fourier Transform of the gate function, the sinc function, "widens" as the time domain window function narrows. Thus, this is nothing but a restatement of the uncertainty principle and the tradeoff between time and frequency.

The shape of the window function will influence the degree of smoothness between the spectrum at time τ_j and τ_{j+1}.

In a more practical light, the time varying spectrum defined $|F(\omega,\tau)|^2$ can be simply generated by selecting a proper segment of the total signal and filtering either using analog or digital filters to obtain the spectrum at a given instant. The window, the inverse transform of the filter function, selects the segment desired and determines the way in which the sample should be weighted. Selected methods for practically generating the time varying spectrum are now discussed.

The time varying spectrum, or sound spectrogram, may be obtained by repeatedly passing the prerecorded sample function to be analyzed through a narrow-band superheterodyne receiver as its local oscillator is swept slowly through the audio band. With the oscillator set at some frequency ω_e and a bandpass filter centered at ω_o, the component $\omega_o-\omega_e$ of the signal modulated by the local oscillator will fall in the filter bandwidth. As the original signal was prerecorded, any desired frequency band or subinterval of the total signal determined by the filter bandwidth may be analyzed. It may be easily shown that the output of such a bandpass filter is the desired function $F(\omega,\tau)$. The spectrum at any prescribed instant may be obtained by averaging the filter output in the neighborhood of the time τ, while the input function is continually presented and the oscillator scans slowly. The concept here is to move the input signal past a fixed filter set at ω_o, the center frequency, shifting the baseband spectrum by modulation. (An alternative approach is to keep the signal baseband fixed and move the filter center frequency.)

Several spectrum analyzers operating as above are available on a commercial basis (e.g. Kay Electronic Company and Signetecton Corporation). These devices permit a total stored signal length of up to about four seconds. Analysis can be performed using filter bandwidths of from 40 to 300 Hz. The analysis is performed as a continuous function of τ and the analysis time is about 1.3 minutes for a 2.4 second record, going up to about 16 minutes for a four second record on one device. Both the time varying spectrum is outputted as well as the instantaneous spectrum at an instant τ, as discussed above. A digital implementation is also available (Voiceprint Laboratories, Inc.).

Since the convolution of two functions in the time domain is equivalent to the multiplication of their respective spectra in the frequency domain, the operation of correlation of two signals $f_1(t)$ and $f_2(t)$ must be equal to the multiplication of the spectra of $f_1(t)$, $f_2(-t)$ or $F_1(\omega)$ filtered by $F_2(-\omega)$. Thus, filtering can be viewed as equivalent to a correlation type of processing. In this regard several references to the above type of filtering refer to it as heterodyning correlation. For the sake of completeness, it should be mentioned that there exists another realization of this correlation using a DELTIC correlator (Delay Time Compression). This approach is quite common in Sonar applications and can be realized both in analog and digital forms. The use of correlation should not be surprising since we are searching for periodicities of the signal which are retained and enhanced in the correlation function. This latter technique produces an output on line real time as compared to prerecording only a small segment as in the former approaches and processing non-real time off line.

In all the approaches mentioned above, the output is marked on an electrically sensitized paper -- whose marking density is approximately proportional to the absolute value squared of the output of the filter. Unfortunately, the dynamic range of the paper that is usually used is quite poor, having a dynamic range of at best 10 db. The information present at the recording mechanism does exceed 35 db. so that the method of display severely limits the use of this type of presentation. In an attempt to improve the resolution of the recording in the amplitude axis, modifications have been made so that equidensity contours of power are displayed in the uniform grey level. The circuitry, is now also available commercially as plug-in modules. The number of distinct levels that may be displayed are still limited to the dynamic range of the paper.

The color encoding methods used in this study extend the above systems in that color is used to display equidensity colors, and a TV display is used to provide for wider dynamic range. It will be seen that the limiting constraint on the obtainable performance is the quality of the raw data, the original speech spectrogram.

c. System Description and Procedures

This research is founded on the hypothesis that, given the voiceprint as a means of identification, color encoding and display of the existing information will result in an enhanced voiceprint which is easier and faster to read and interpret. Consequently, the problem reduces to one of image enhancement where the image is an original voiceprint.

Color enhancement techniques, as a possible means of increasing the amount of information that may be extracted from such recorded images is based upon the knowledge that the human eye can distinguish between colors much more easily than between shades of gray.[12] Developments in methods of producing color images in the infrared, microwave, and visible spectrums have already led to advances in other fields such as medicine, environmental protection, agriculture, and industrial production procedures.[8,9,10,11,12]

For example, infrared sensors have been used to scan the body to reveal color encoded variations in temperature that may be related to disturbances such as tumors, poor circulation, and inflamation caused by arthritis.[9] Aerial photographs taken using infrared color film, have been used to show the discharge and distribution according to color of pollutants in rivers, lakes, and other bodies of water, by showing the different levels of concentration.[11] In the field of agriculture, forests and crops may be checked for disease and healthiness by using color enhanced aerial infrared photographs. Electronic circuits may be non-destructively scanned and the color enhanced image used to detect "hot spots".

Airborne microwave systems are presently used to scan the earth and produce a color image from the apparent temperatures of the earth's surface that is scanned. This type of radiometer could become important in the collection of meteorological, glaciological, and oceanographic data.

In the visible spectrum, color enhancement techniques applied to black and white photographs of the sun, ice fields, clouds, and portions of the earths surface bring out various features in each. These techniques when applied to x-rays have been used to point up the contrasts between the different tissues.[10]

Methods of producing color enhanced images from black and white ones vary almost as much as the methods of obtaining them. Most techniques are costly and require relatively large amounts of time. One such technique starts with a black and white photograph and requires about a week and $500 to produce a color enhanced version. The process consists of making overlays by repeatedly rephotographing the original with high contrast black and white film using a variety of exposures to establish different light intensity levels. A selected color is then added to each overlay and the overlays are combined to form the final picture.[10]

One pyroelectric detector uses optical mechanical methods for color enhancement. In this system, the vertical scan requires 30 seconds and consists of 100 lines. The color is produced by placing a step color filter between the elements of the condensing lense. The filter is composed of six to ten sections bracketing the visible spectral range from blue to red. The color of light is then dependent upon the position of a galvanometer that controls the position of the filter. A camera and a wheel containing 8 filters are synchronized to the scanning rate of an infrared detector in another system, and are used to produce color thermograms by exposing one frame of color film to a number of different isotherm levels with a different color filter in front of the camera for each level.[8]

A system produced by Spatial Data Systems[13] provides more rapid color enhancement using a black and white television camera, a special digital video processor, and a standard 525 line color monitor. The digital video processor in this system analyzes the shades of gray, produces a digital code for each shade, and classifies them into a certain category. A color signal suitable for operation of the color monitor is then produced. The cost of this system is approximately $13,000.[13]

Recently, an image processing and/or enhancement system was developed at R.P.I. which among other features, also provides for real time color encoding of video signals. It transforms a range of densities or shades of gray, corresponding to variations in amplitude of the black and white video, into arbitrary colors.

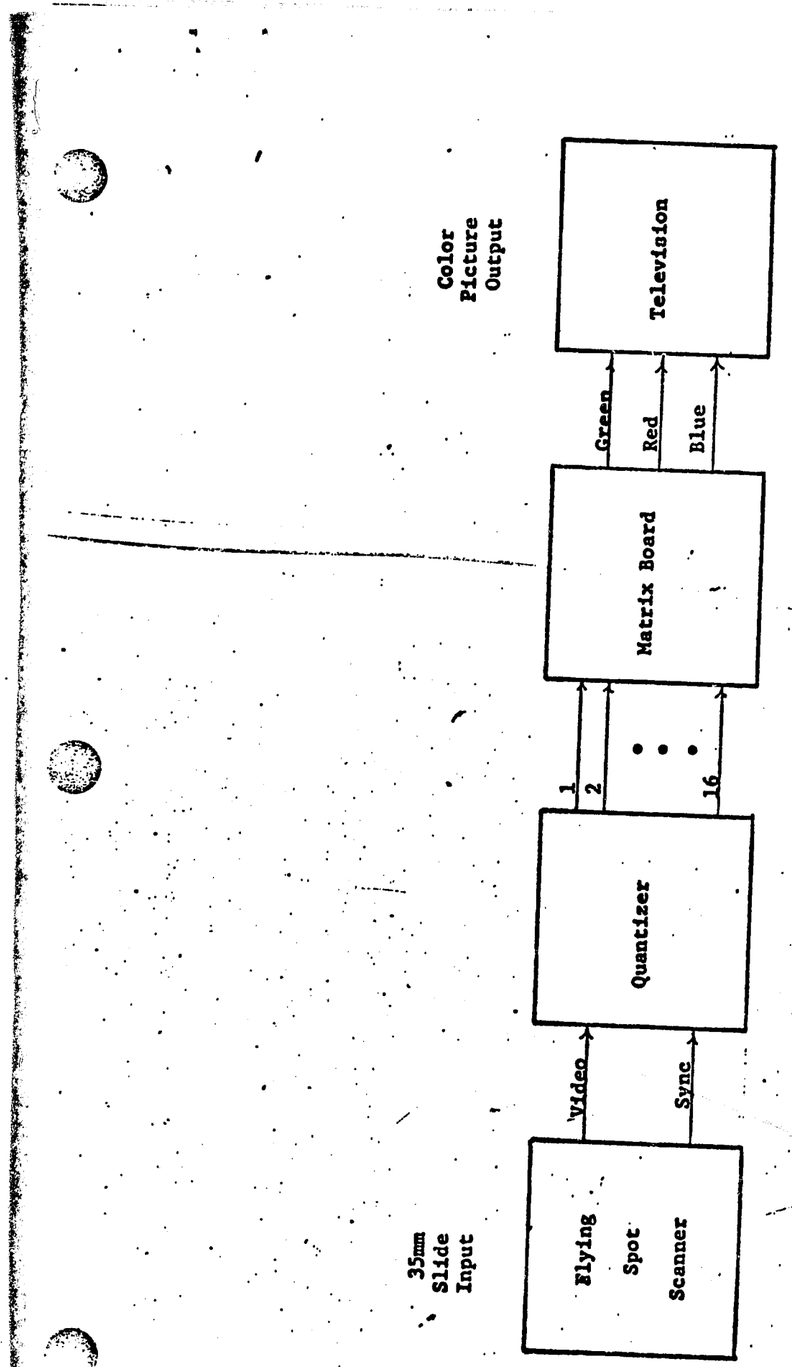
The color chosen to correspond to a particular density range is unique and completely independent of that density range. An additive background signal, consisting of the original black and white video, is also available, and makes it possible to display a portion of the voiceprint density color enhanced, while at the same time providing the unenhanced image in its original form.

The system operates in real time, and many of the parameters may be varied while in operation to provide greater flexibility and information. Image information such as the voiceprints, may be supplied in their original form via a closed circuit TV camera or as 35 mm slides.

The RPI system that has been developed consists of a quantizer, a console color television that contains a slide viewer and a TV camera. This total system costs about $4,000. The slide viewer is a flying spot scanner which scans the slides producing a beam of light that is modulated by the slide and detected by 3 photo multiplier tubes which produce an electrical signal. This signal is then converted to a normal video color signal. The flow of information is shown in Figure 1, where the system has been broken up into separate blocks for each function. The black and white image is first converted to an electrical signal in the first step by use of the flying spot scanner, or the camera depending on the source. The quantizer produces 16 outputs which are then used as inputs to a programmable matrix board. The matrix board then combines the 16 signals into three outputs which are returned to the color television to produce the colored image.

The television and slide viewer are mounted in the same console, but each has a separate chassis and power supply. The television is a Sylvania color television, chassis model D13-2.

The flying spot scanner slide viewer is completely transistorized except for the photo multipliers and the flying spot scanner tube. The circuits, Sylvania chassis H01-1, -2, for the slide viewer are broken up into three basic parts and each part has a different function. The three parts consist of the flying spot scanner and photo multiplier circuits, the preamplifiers for each of the three photo multipliers outputs, and a video processing board that contains circuits for DC restoration, automatic gain control, and matrixing the three color signals into one. It is in the slide viewer and camera output that the circuit modifications to enable the interconnection of the quantizer and matrix board were made.



Block Diagram Showing Signal Flow Through System

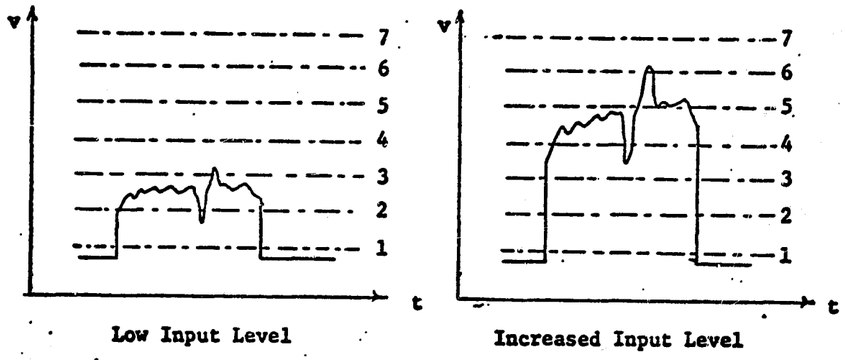
Figure 1

The quantizer is model 606, produced by Colorado Video, Incorporated, and produces 16 levels of quantization. Inputs to the quantizer include, video, horizontal drive, external, and a keyed video input for the feedback connections. The outputs consist of sliced and keyed video, and horizontal drive. Controls for the quantizer are, input level, analog level, bias level, function selector, and 16 threshold levels. The video input for the quantizer is from the camera or scanner.

The controls for the quantizer are all mounted on the front panel, and provide control over the various outputs and functions. The function selector switch has three positions, internal, external, and test. The internal position is the position in which the system is normally operated. In the internal mode, the signal is taken from the video input and sent to the individual quantizers. The output in this mode of operation is the quantized video signal. The test position is used when a video signal is being quantized, and there is either a separate horizontal drive signal or the information is contained on the video signal. This function allows the quantization levels to be adjusted, by generating a sawtooth wave which when quantized produces color bars on the television screen. The quantization thresholds can easily be varied by adjusting the threshold potentiometers while observing the width and position of the resulting transformation of density to color (color bars) on the television screen.

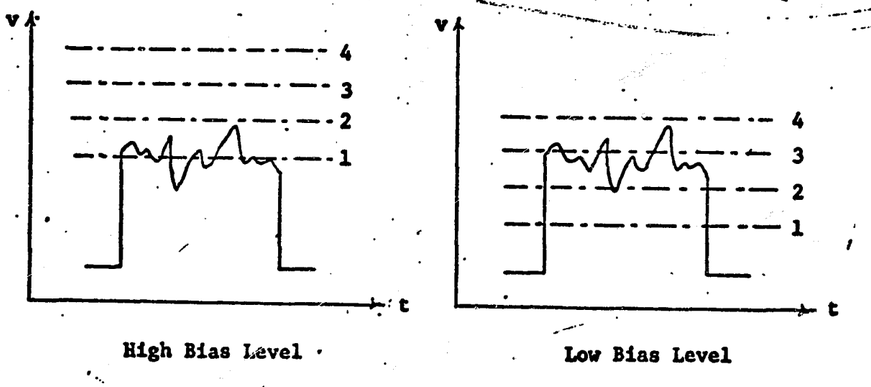
The input level control allows the amplitude of the input to the individual quantizers to be varied. This is useful in that it allows a linear change in the accuracy of quantization for fixed levels by changing the percentage of the signal between each slicing level as shown in Figure 2. The brighter portions of the signal may also be colored first by slowly turning up the input level until just one or two outputs turn on. This is equivalent to very coarse quantization as the whole signal is contained in a few levels. Combined with the bias control, this permits adjustment of the accuracy of quantization as well as the choice in the range of densities to be colored. The result of varying the bias level can be seen in Figure 3, and is one of the more important features of the quantizer. The following matrix board may also be used to select the desired color coding scheme, and ordering.

The analog level control is used to vary the amount of the black and white image that feeds through the quantizer to the television. The background level can be varied from a position that does not allow any of the signal to pass through to one that allows the normal black and white picture to be shown on the television.



Effect of Increase in Input Level

Figure 2



Effect of Change in Bias Level

Figure 3

Sixteen potentiometers that control the threshold voltage at which the corresponding quantizers turn on can be varied in any manner desired. If the keyed video feedback is being used, the threshold level of one quantizer may depend on the threshold level of another. An example of this could be if the output of quantizer number 2 is used to turn off quantizer number 1. In this case if the threshold of number one is set at a higher voltage than number 2, number 2 would turn on at the lower input voltage, and number one could never turn on. If the threshold levels are set in the normal fashion, the lower boundary of the slice is fixed by the threshold of the particular quantizer, and the upper boundary by the threshold of the next quantizer. The thresholds may be set to allow uniform or nonuniform quantization as well as varying the fineness.

The many variable parameters of the quantizer provide a large amount of flexibility, and proper use of the many controls can provide a very informative color enhanced image.

The quantizer outputs (16) are supplied to a programmable matrix board. The matrix board 16 x 3, has two layers of parallel contact strips positioned so that the upper layer of strips is perpendicular to the lower layer. Holes have been positioned at points where, if the upper layer were projected on the lower layer, the contact strips would intersect. Pins are inserted through both layers of the matrix board to connect the 16 inputs to the 3 outputs which are in turn fed to the color display. In this manner, ordering and selection of desired colors are obtained.

For the color enhancement of voiceprints, both the camera and flying spot scanner (FSS) inputs were used in conjunction with the color encoding system. In this way the original voiceprint as well as a 35 mm slide of the voiceprint could be accomodated.

The voiceprints were produced using a Kay Electric Sona-Graph 6061A. The Sona-Graph 6061A is an audio-frequency spectrum analyzer that produces permanent, graphic recordings of any type of complex wave in the range of 85 to 8000 hertz. Unlike conventional spectrum analyzers, the 6061A permits three different analyses to be displayed; the operator can select the display that most accurately shows the parameters he is studying. The No. 1 display gives an overall, three-dimensional picture of the signal being analyzed; frequency, amplitude and time are represented simultaneously on one display. The second type of analysis, a No. 2 display, permits the individual intensity of each frequency component to be displayed at any preselected point in time. This type of pattern is referred to as a Section. The third analysis that can be performed is similar to an oscilloscope display; it shows the average amplitude of all frequencies present, relative to time. With this pattern, the entire input signal can be examined for flatness, resonance peaks, or any amplitude study relative to time. Display No. 1 was used for all experimental work in this study.

In this mode, this unit will display any 2.4 second portion of audio in the 85 to 8000 hertz range. The input signal is first recorded on a continuous drum, and then played back at a higher speed during the analysis process. A frequency-heterdyne technique is used for the scanning system, and there are two plug-in filters available for increased flexibility. Constrained by the uncertainty principle, the narrow filter emphasizes frequency resolution, and the wide filter emphasizes time resolution. A build-in calibration-tone generator can provide frequency markers every 500 hertz along the frequency scale of the pattern, simply by depressing a switch.

An adjustable AGC control is present and can be used to extend the dynamic range of the pattern; also the darkness of the pattern can be adjusted to obtain the best contrast. For monitoring purposes, a VU meter and a loud speaker can be used simultaneously either when recording the input signal and when performing the analysis.

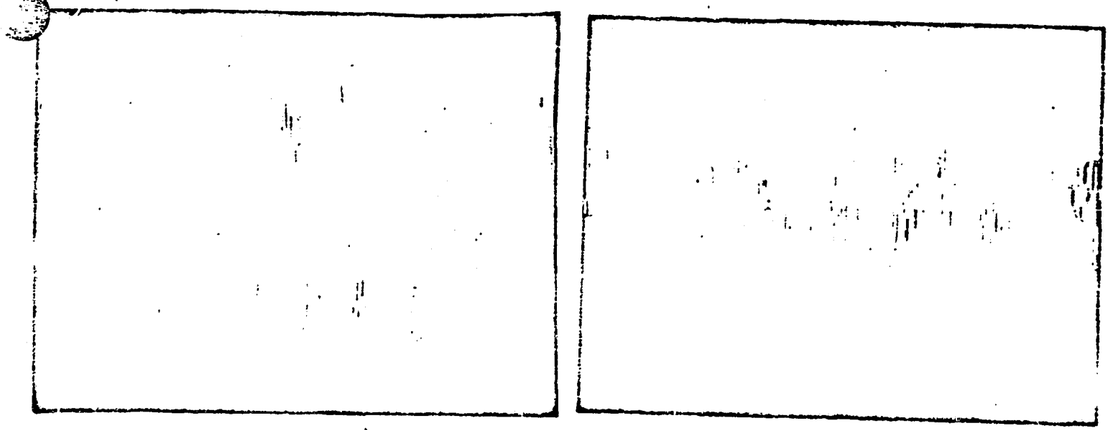
After deciding on the experiment to be performed (explained in the next Section), an appropriate voiceprint was made using the Kay Sona-Graph. The original was placed before the camera and enhanced in real time. When the desired result as viewed live by the operator was achieved, the color TV monitor display was photographed using a Nikon 35 mm camera at a speed of 1/8 second at f 2.8 nominally. As an alternative a black and white slide was made from the voiceprint and the scanner input used to enhance it in real time. Note that in both cases the voiceprint, once obtained, is enhanced and displayed in real time; on line.

The majority of experiments were performed in this manner. Methods for obtaining other selected results are described in the next Section as required.

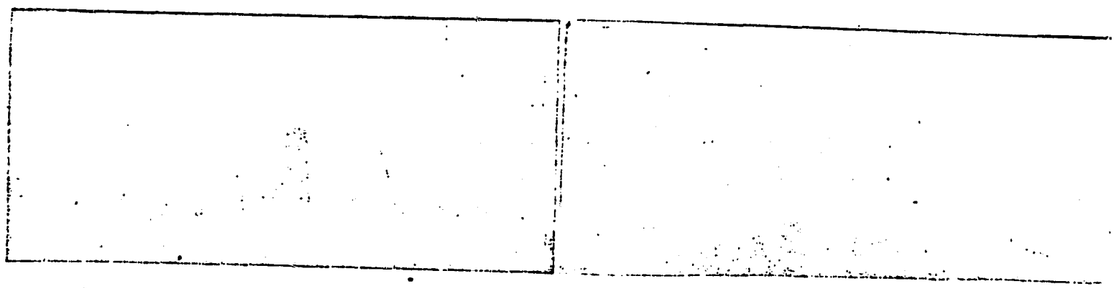
d. Experimental Results

As preliminary to evaluating the effectiveness of color encoded voiceprints vs. conventional voiceprints, tests were first conducted using human subjects to determine the contribution of different types of data presentations including a time varying spectral display (TVSD) of which the voiceprint is a prime example. The data consisted of different amplitude sinusoids imbedded in additive Gaussian white noise of zero mean and constant variance. The types of displays included a) the signal and noise displayed as a function of time on an oscilloscope trace, b) the instantaneous spectrum of the signal and noise displayed on an oscillograph record, c) an audio display of the signal and noise, d) the time varying spectrum of the signal and noise, and e) the time varying spectrum of the signal and noise color enhanced. Two examples of presentations a) and b) are shown in Figure 4a and 4b respectively.. The time display shows the signal in noise for a signal to noise ratio (S/N) of 5 db and -10 db while the frequency display is for a S/N of -6 db and -25 db respectively. The signal in noise displayed on a TVSD appears as a line (the signal spectrum) parallel to the time axis against a background of "pepper and salt" white noise.

Several tests were run for each type of display on up to 20 individual subjects for S/N ratios varying from 12 db to -30 db. As a result, the ability of the individuals to detect the presence of the signal in this simple two hypothesis problem (signal or no signal) as a function of the display type is summarized below.



a) Time Domain Signal in Noise



b) Instantaneous Spectrum

Figure 4 Display Comparison Test Formats

Display in Order of Increasing Effectiveness

1. Oscilloscope time domain display
2. Instantaneous spectral display
3. Listening to the acoustical signal
4. Time varying spectral display

The least effective display was the direct time domain representation which proved useful for S/N ratios to about -6 db. Improvement was obtained as more cycles were displayed to the subject, but a limit in performance for a specified S/N was reached when approximately 10 cycles were shown. The TVSD permitted detection of the signal at S/N ratios as low as -25 db. Note that special high resolution intensity modulated display was used for these cases, since the dynamic range of the Kay equipment used is limited to, at best,, 8 levels.

The contribution of the TVS color encoded dispaly in this series of tests proved to be primarily in the reduction of time needed by the observer to detect the signal with the same level of confidence. (This result was substantiated in similar tests on actual voiceprints.) The attainable detectability as a function of S/N ratio was approximately the same as for the Black and White presentation provided the dynamic range was sufficient. The observation time necessary to arrive at a decision is demonstrated in Figure 5. Here the percentage of permitted maximum observation time (10 seconds) actually required (on average) by the subjects to detect the signal is plotted as a function of S/N ratio for the conventional display and color display. The advantage of color is greatest at the poorest S/N ratio and reduces the time required by over 25% at -10 db.

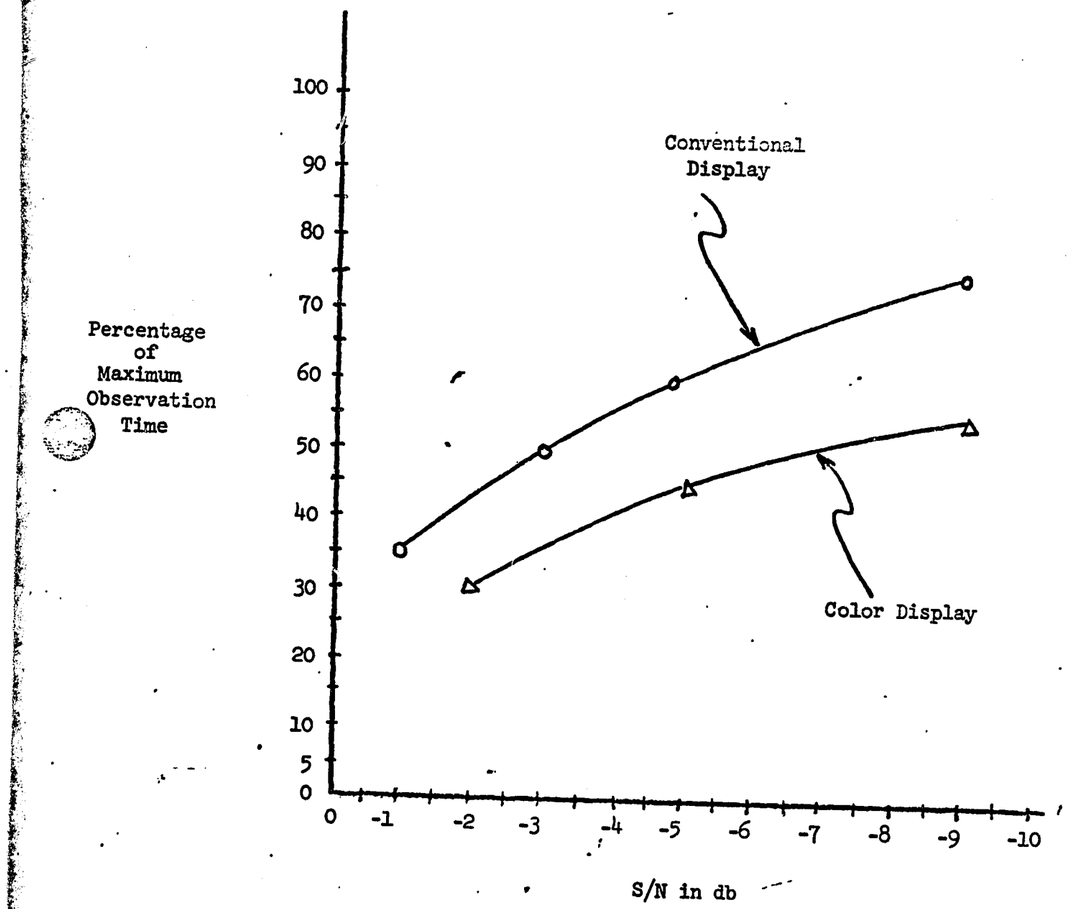


Figure 5 Reduction in Observation Time Using Color

The fact that significant differences in detection capability did not result using the color display (although a few db improvement was obtained) reenforces the idea that color encoding can only dispaly information that already is present but does so in a more comprehensible form.

The specific colors that were used did not, in general, effect the subject's performance (provided reasonable sensitivity to the eye was maintained to the hues selected).

However, the number of colors used did effect the performance. Because of the limited dynamic range of the conventional sonagram, only three colors or less were required. Most of the tests were, in fact, run with two colors, the threshold between them being set at the theoretical level to minimize the probability of error of detection assuming equal costs of errors and apriori specified probabilities. In the cases where a CRT display was used and more dynamic range was available, (thereby not limiting the number of colors) an increase of performance still did not result by increasing the number of colors. More than about five colors tended in fact to confuse the subject and resulted in poorer performance. As a result, the subsequent tests on actual voiceprints use less than five colors.

The remaining results of this Section deal with actual voiceprints. At this point, the basic hypothesis of this research should be reemphasized - given the conventional voiceprint as a means of suspect identification, color encoding and display of the existing information will result in an enhanced voiceprint which is easier and faster to read and interpret. In all tests conducted in association with this project, the hypothesis was borne out.

This does not, however, in any way substantiate the use of the conventional voiceprint. In fact, in several instances, the color enhanced voiceprints served to point out the shortcomings of the original data. Moreover, the results should not be interpreted as suggesting the use of the present color enhancement process for regularly converting conventional voiceprints to be used for suspect identification. What is shown is that color display of information "of this type" is more effective in permitting a comparison to be made, but should be utilized with only the proper input information e.g. an improved voiceprint.

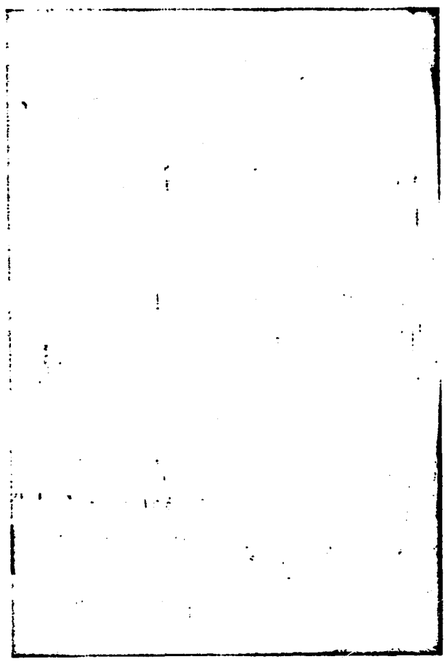
As general results on actual voiceprints, all subjects tested were able to classify the color enhanced voiceprints more easily and more rapidly than the conventionally displayed information. The errors made did not vary significantly whether the color encoded prints were used or not, regardless of the quality of the original information. However, it must be noted that in no case was the original voiceprint information deemed sufficient to permit extraction, by color encoding, of the maximum amount of features usually found in similar applications. Any lack of anticipated improvements is attributed to the relatively poor quality and nature of the classical voiceprint. It is for this reason that a subsequent Section is solely concerned with a new, more efficient, real time, high resolution, adaptive method for producing voiceprints. Implicit in this approach is a color encoded display.

Based on the literally hundreds of images considered, some basic observations were made. These are exemplified by the pictures shown in Figures 6 through 12.

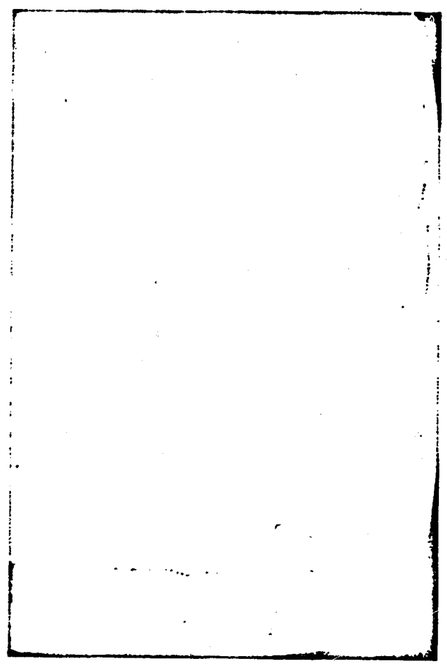
Figure 6 is an attempt to demonstrate the effects of alteration of the parameters available in producing a voiceprint from the same acoustical recording as recorded in a color encoded format. The author's name spoken in the sentence "This is Lester Gerhardt calling" served as this particular record for the voiceprint. Figure 6 a, b, and c are all voiceprints or more appropriately TVSD's of the same recording by the same speaker. The differences lie only in the Sonagraph parameters used to make the TVSD. Specifically, the first used a wide filter setting, the second an H-S filter setting, and the third a narrow filter setting. It is not the intent of these results to say unambiguously that 6a, b, and c are "obviously" made by the same speaker. They are provided for the reader to make his own decision of similarity based on features that he may find. Figure 6d is the author's name said in the same context but by a different speaker. Conclusions at this point are left to the reader.

Figure 7 shows the color encoded quantization levels used to produce Figure 6. The arrow below the figure shows the range of the video signal and bias position used (as explained in a previous Section). Only two colors were used here as a result of the prior tests already cited.

Figure 8 is an alternate display of Figure 6a with a different set of color encoded parameters. The associated color encoded quantization levels should be self explanatory. The display is obviously different and was adjusted such that selected contours were enhanced in yellow. Note, however, the retention of basic characteristics.

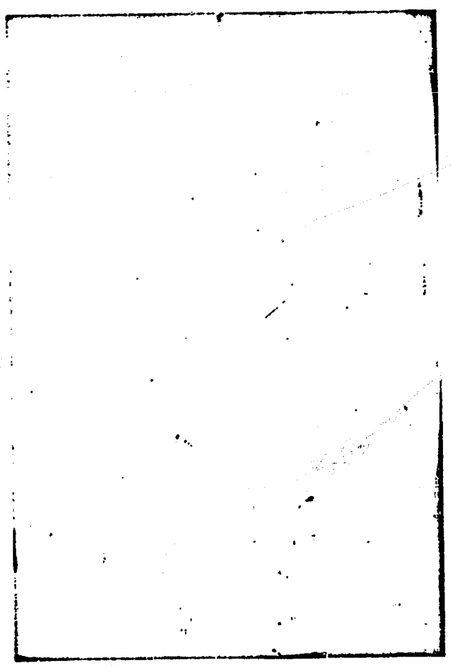


b) Speaker A - Condition 2

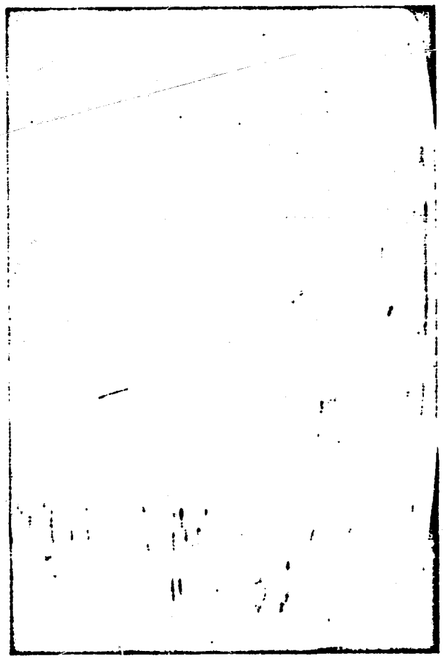


d) Speaker B

Figure 9 Color Encoded Spectrograms - Two Colors



a) Speaker A - Condition 1



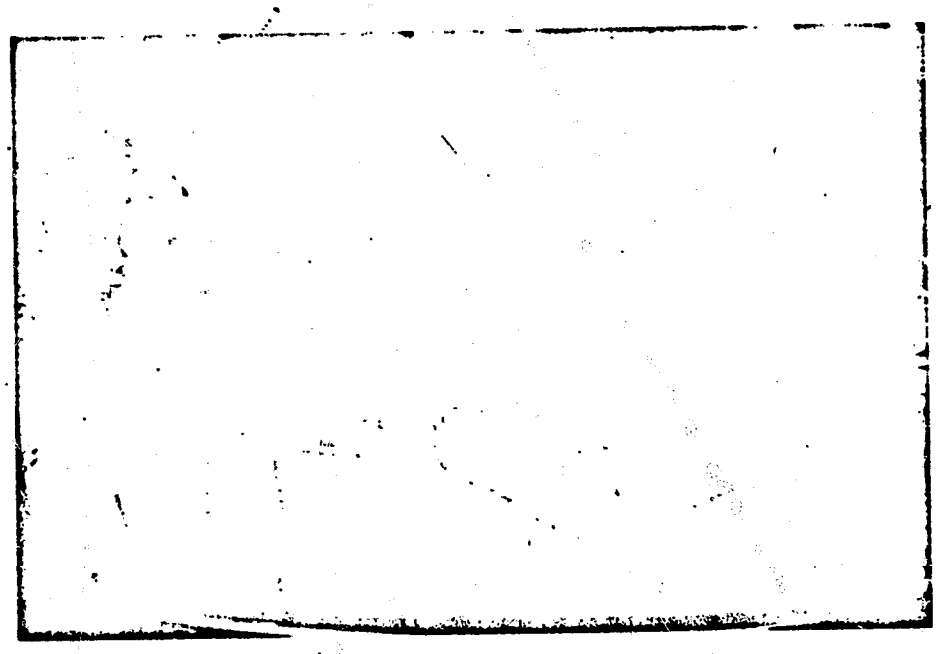
c) Speaker A - Condition 3

Figure 9 shows additional versions of the same recording made into a TVS color encoded display. Here the "slices" are comparatively narrow, and show the nonuniformities in the background as what might be considered as additional features, but which of course are not. The point of this Figure is to demonstrate that additional false "information" can be found by an operator of such a device if he is not selective and knowledgeable of the problem.

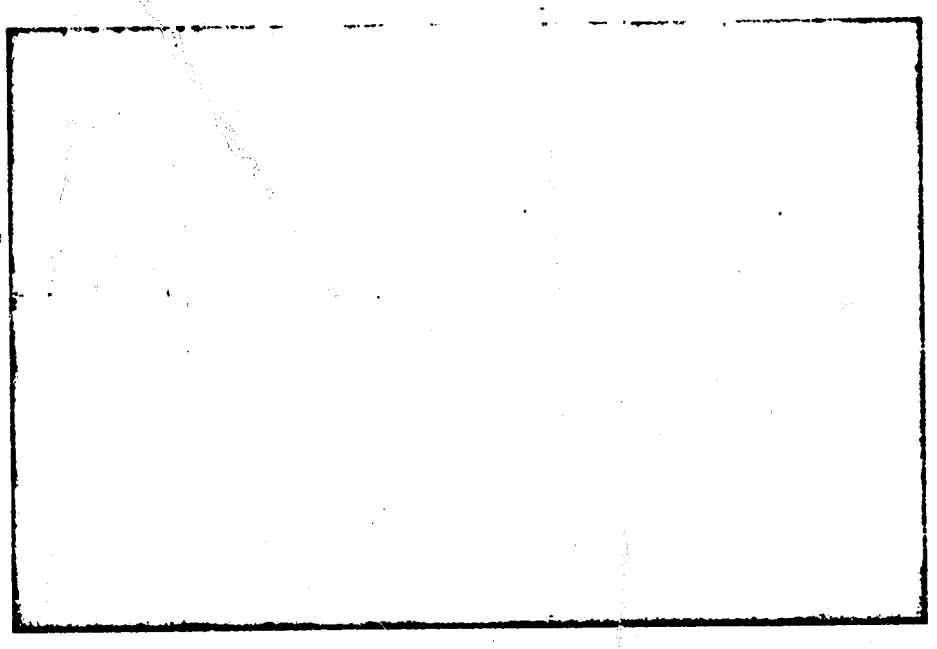
Figure 10 shows the TV's color encoded displays made from conventional spectrograms for three different speakers. The phrase "Who is this?" served as the recording. It is to be emphasized that no adjustments of the encoding parameters were made between the four cases. Once again, in all cases, subjects were able to classify color encoded sonagrams easier and quicker than conventional formats to achieve the same classification error. Whether the differences between speakers A, B and C are greater than between A disguising his voice is left to the interpretation of the reader.

Figure 11 is a comparison of the same information as in the previous Figure but using different levels and more colors. Both more real and artificial information is seen to be available than in the previous representation.

Figure 12 is a final alternate display with its associated color bars. It is to be emphasized that all the same recordings were used for Figures 10, 11, and 12.



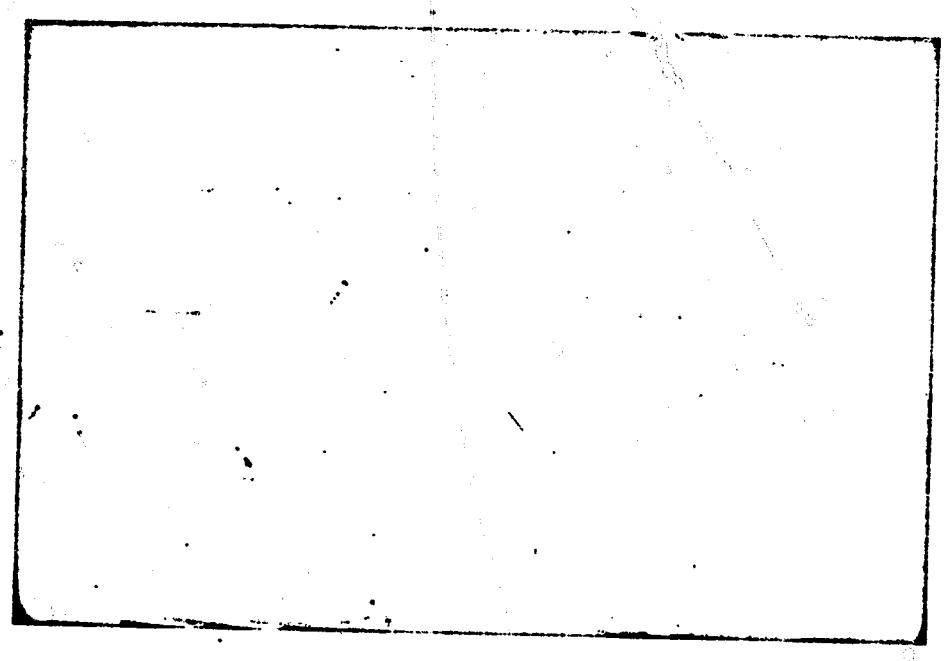
a) Version of Fig. 6a



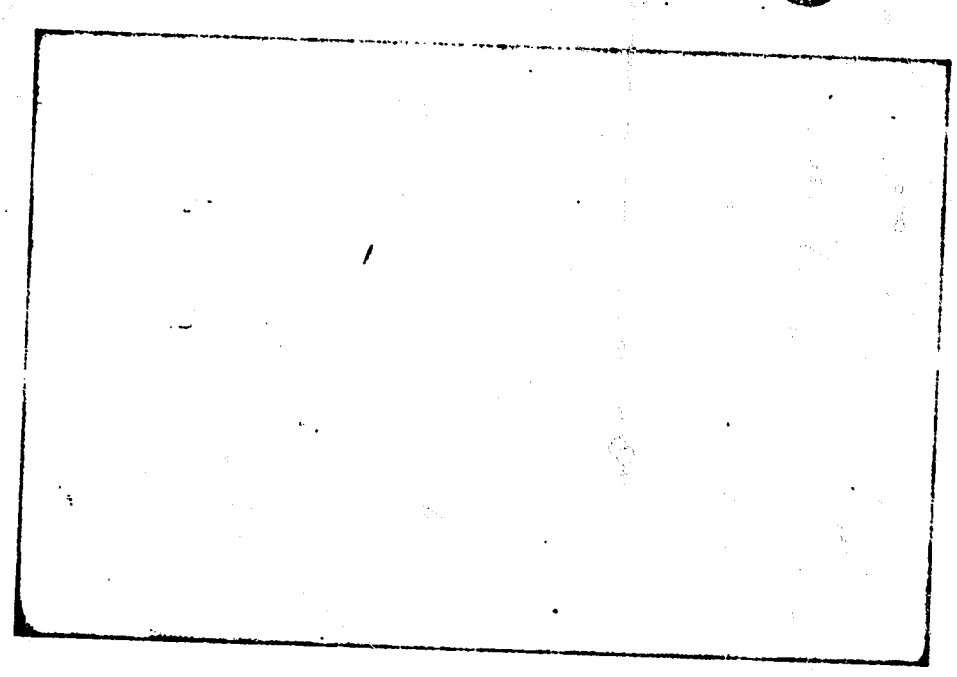
b) Version of 6d

Figure 9 Alternate Color Encoded Voiceprints - Several Colors

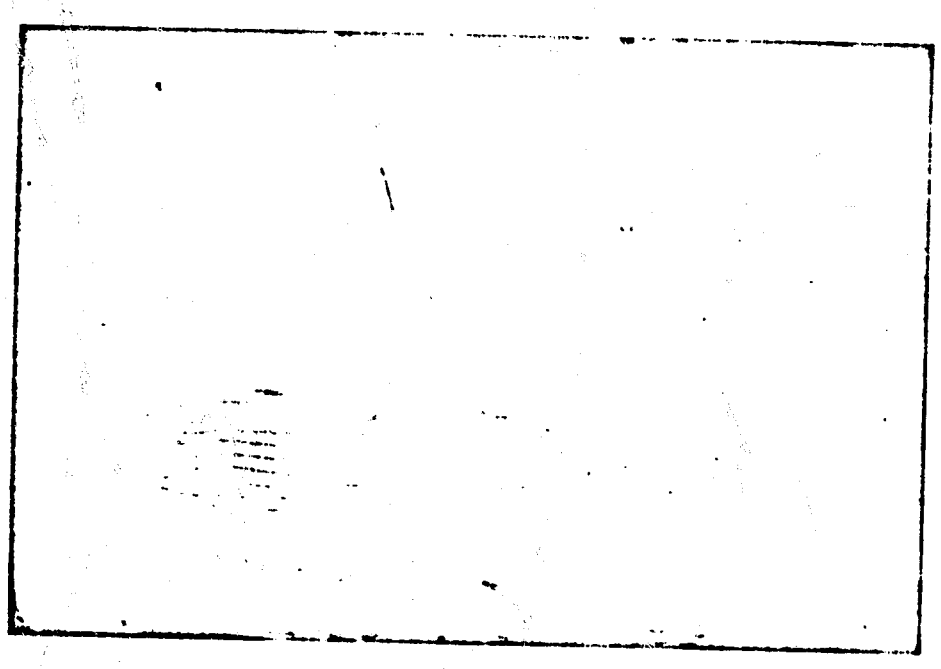
Figure 10 Color Encoded Voiceprints - Different Speakers



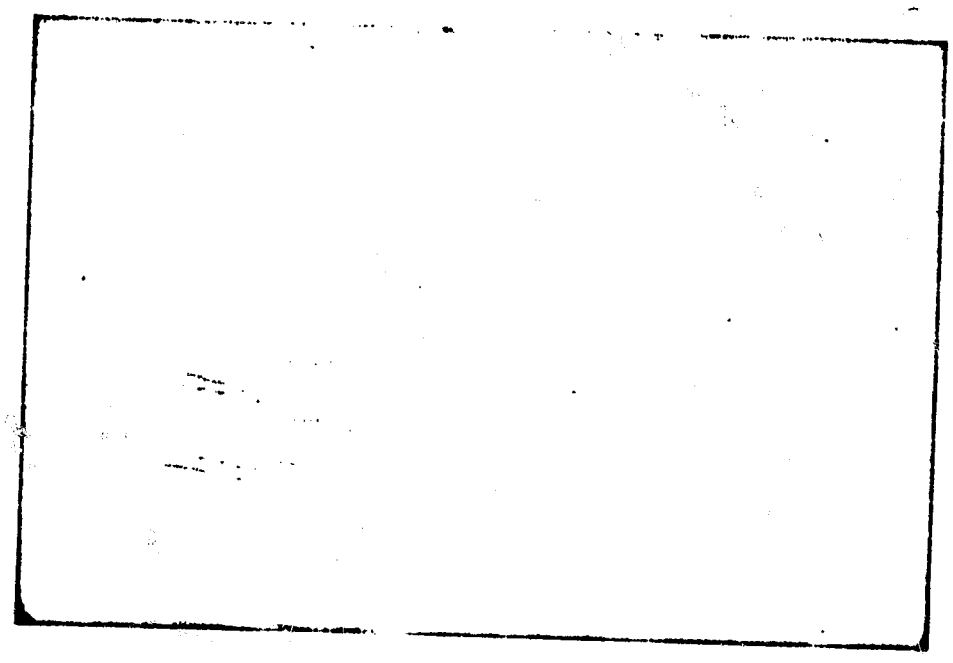
a) Speaker A - Normal Voice



b) Speaker A - Disguised Voice

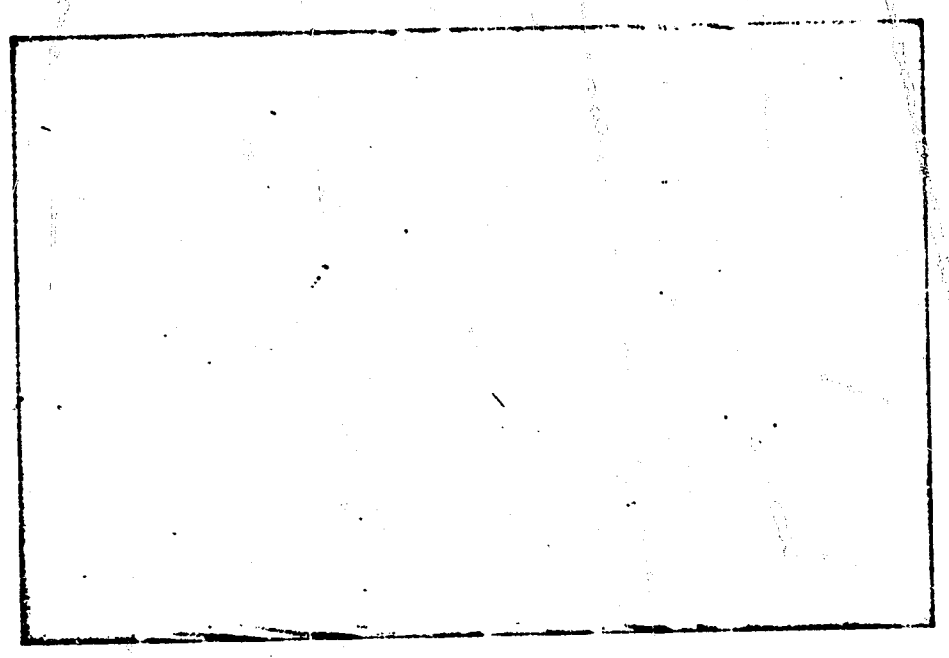


c) Speaker B

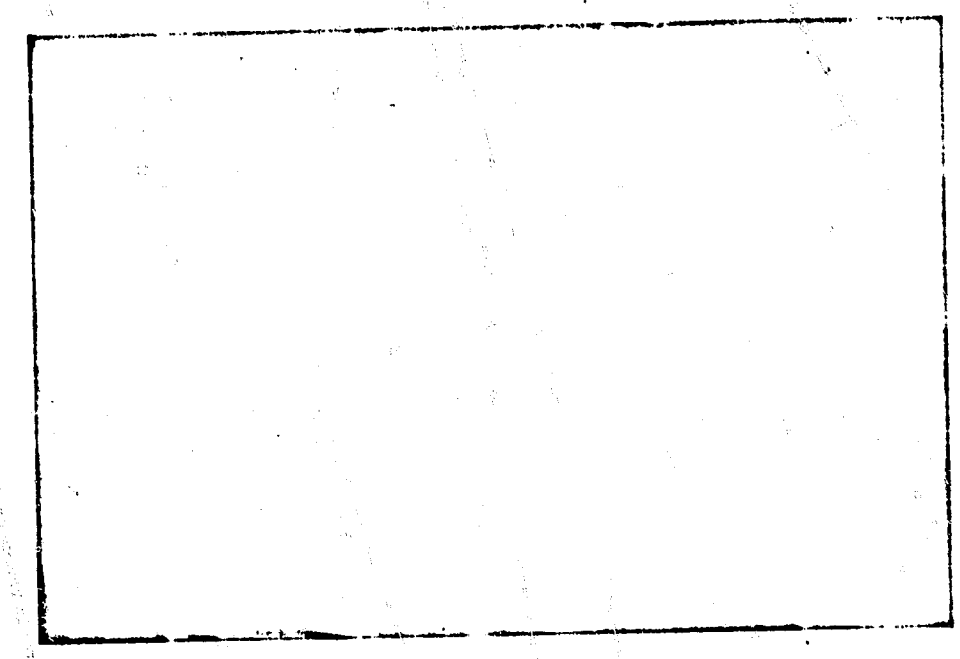


d) Speaker C

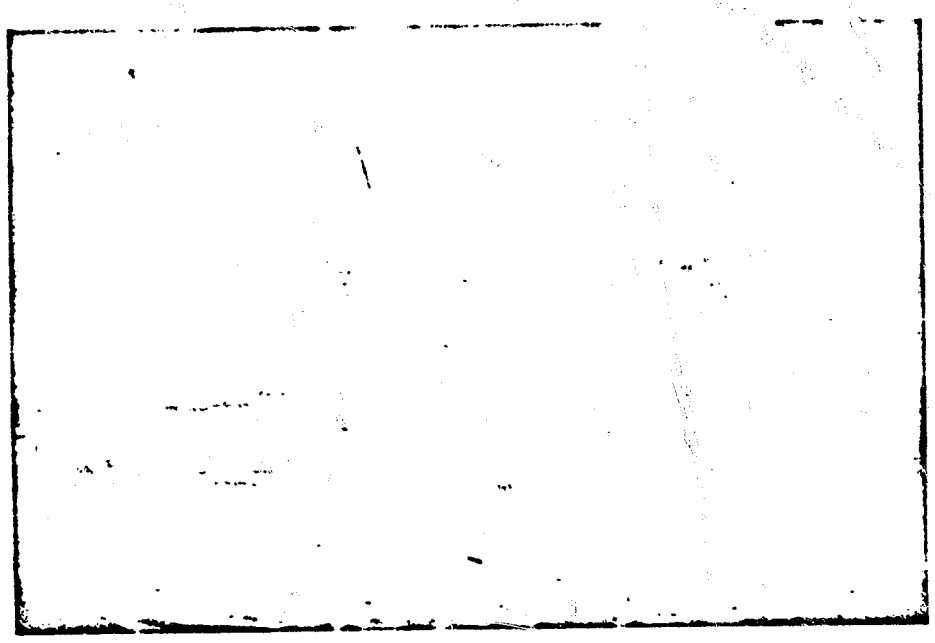
Figure 11 Alternate Display of Figure 10



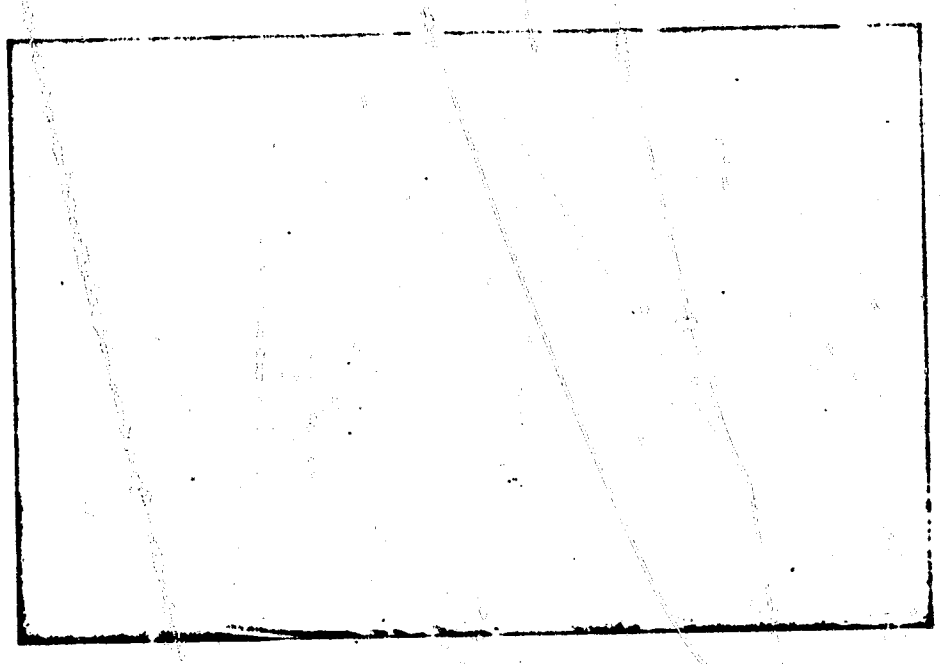
a) Speaker A - Normal Voice



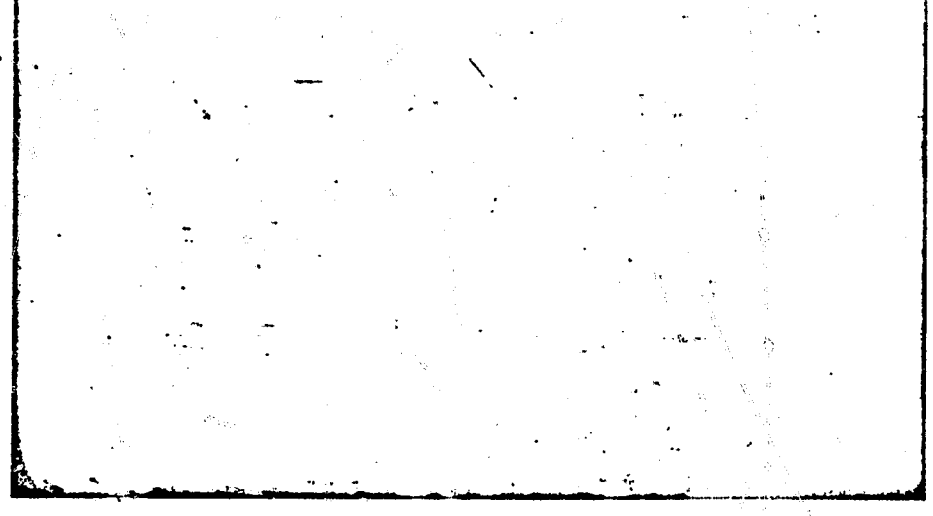
b) Speaker A - Disguised Voice



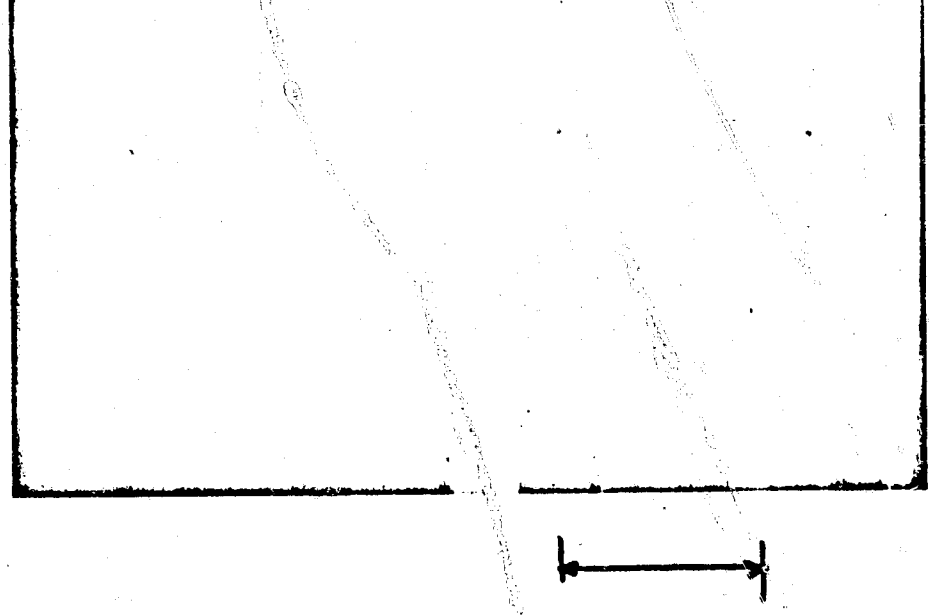
c) Speaker B



d) Speaker C



a) Redisplay of Fig. 11d



b) Color Pattern

Figure 12 Multiple Color Display and Associated Color Pattern

As already pointed out, a severe limitation of the conventional display is its limited dynamic range. In an attempt to see if an improved dispaly would yield improved color encoded capabilities, sonagrams were contoured plotted using the Kay plug-in module made for this purpose. Unfortunately as was expected, although contour plots do provide more sharp gradations in gray scale between adjacent levels, they are still recorded on the same paper with its limited dynamic range. As such greatly improved classification capabilities are not achievable. Improvement in performance is only marginal compared to the high resolution CRT display used in the S/N experiments and the associated improvement in time responses observed there. There was no available means to generate a high resolution (in all three axes) Time Varying Spectral Display to effectively test the improvements gained by color encoding this type of display of similar information. Continuation of the program toward the development of a real time high resolution digital system would permit such an evaluation to be conducted.

One could conclude from these results, that with additional parameters such as those used in the color encoding process, the sonagrams can be made to show almost whatever the observer desires. While this is not quite the case, sufficient variation and improper adjustment can lead to severe mis-interpretation. It must be noted however, that "information" cannot be and is not created in such a process. Whatever is displayed is present in some form in the original data. The fact that the reader finds he cannot convincing state that all enhancements are obviously made from the same recording (as they are noted) forces him to question the validity of the very nature of the original data, the sonagram or "voiceprint".

It is felt, therefore, that continued improvement in techniques of suspect identification using acoustical patterns can only be gained by considering new and improved methods of generating a Time Varying Spectral Display of speech signals. The results of this color encoding has served to bear this out, and the next Section describes a real time, on line concept for just such a system utilizing digital techniques. Based on the generally favorable results regarding ease and simplicity of analysis of valid data using color, the system will implement a color encoded display.

e. A Real Time System Concept

It is clear, by previous research and the results of this program, that recognition is optimized when all three variables of a voiceprint, frequency, time and particularly intensity are present with higher resolution than previously utilized. The use of higher resolution in the intensity axis was initially provided by the Kay Electric Sona Graph plug in Contour Unit and later by more sophisticated Voiceprint Laboratories equipment. The main disadvantage is that these techniques are off line and non-real time as previously pointed out. The enhancement techniques discussed here permit easier extraction of the previously cited contours and are real time on line with respect to the original voiceprint (which is itself produced off line). The ultimate system is one in which the speaker's voiceprint would be continuously generated on a real time color encoded time varying spectral display and is relatively inexpensive. It is with these objectives in mind that the proposed system of this section was conceived.

The basic system uses digital signal processing techniques to produce a continuously varying time varying spectral dispaly of approximately the last three seconds of voice information with a color encoded format. It is anticipated that more optimum processing, filtering and encoding (including adaptive techniques) can be easily instrumented in this system such that a much improved "voiceprint" or signature can be obtained which contains more meaningful clues or identifying characteristics, then enhancement of conventional voiceprints have been able to demonstrate. Because of the considerable use of digital techniques in this proposed system, some background in this area is first given followed by a description of the system itself.

Recently, great strides have been made in the field of computing with strong emphasis on real time on line processing techniques. The Discrete Fourier Transform can be obtained rapidly using digital filtering techniques using a Fast Fourier Transform algorithm. The discrete approaches are different in the sense that they deal with a sampled signal and operate on a finite number of time domain samples. Regardless of the approach, given only a finite number N of time samples, a discrete spectrum results which itself can possess only a finite number N of discrete frequencies. Once defining a Discrete Fourier Transform, procedures for generating the time varying spectrum are basically similar to those explained previously.

To elaborate a little further, a digital filter is the algorithm by which an input sequence of data is transformed into an output sequence. The process is assumed to be linear and is defined as the convolution.

$$y(nT) = \sum_{m=0}^{n} h(mT)\ x(nT - mT)$$

where $y(nT)$ = output sequence of numbers;

$x(nT)$ = input sequence of numbers;

T = the sampling interval of the continuous signal $x(t)$; and

$h(mT)$ = weighting sequence defining the digital filter.

The convolution is transformed into a multiplication in the frequency domain by the z-transform. The z-transform of an infinite sequence of number $x(nT)$ is defined as:

$$X(z) = \sum_{n=0}^{\infty} x(nT)\ z^{-n}$$

where $z = e^{sT}$ = unit delay operator

$s = j\omega$ for real frequencies.

The convolution may be written as:

$$Y(z) = H(z)\ X(z)$$

$H(z)$ is the frequency domain representation of the digital filter. The most general form of $H(z)$ is the general recursive digital filter function:

$$H(z) = \frac{a_o + a_1 z^{-1} + a_2 z^{-2} + \ldots + a_n z^{-n}}{1 + b_1 z^{-1} + b_2 z^{-2} + \ldots + b_m z^{-m}}$$

A nonrecursive filter is defined with only the numerator coefficients a_k nonzero. A purely recursive filter is defined with only the denominator coefficients b_k nonzero and at least one numerator coefficient nonzero. As such, the z transform in the discrete case in analogous to the Laplace transform in the continuous case. Therefore, filtering and general signal processing can be performed in the discrete domain quite simply.

As an example of the analogy, assume the input is a pure sinusoid at frequency as

$$x(t) = e^{jwt}$$

Let $t=mT$ where m is an interger and T is the sampling instant. This leads to the DFT

$$\bar{H}(jw) = \sum_{m} h_m\ e^{-jmwT}$$

where h_m corresponds to a set of smoothing weights. Note that $\bar{H}(jw)$ is periodic with period $2\pi/T$. If it were desired to average a particular sample with its two nearest neighbors so that

$$y(nT) = \frac{1}{3}\ (x(h-1)T + x(hT) + x(h+1)T)$$

it is clear that

$$h_{-1} = h_0 = h_1 = \frac{1}{3}$$

and

$$\bar{H}(j\omega) = \frac{1}{3}(1+2\cos \omega T)$$

which acts as a digital low pass filter (which is also periodic).

Therefore, even by these simple examples, it should be apparent that the desired processing of the audio signal can be accomplished in the discrete domain, which offers an inexpensive, reliable and efficient approach. The one distinct tool that makes it possible to accomplish this processing on line and in real time is the Fast Fourier Transform. Because of its importance and contribution to the proposed program, it is discussed next.

This section presents the development of the FFT from the basic DFT for the case of N samples where $N=2^M$. The material here is the foundation for the digital processing used in the proposed system.

The DFT (Discrete Fourier Transform) of N points, X_p for $0 \leq p < N$ is defined as

$$A_r = \sum_{p=0}^{n-1} X_p \exp(-2\pi jrp/N) \qquad 0 \leq r < N$$

The A_r are the N complex frequency components derived from the X_p and have the following properties:

(a) If the X_p are sampled over a time interval T at a rate (N/T) from a real signal, the A_r are separated in frequency by (1/T)Hz.

(b) If the X_p are real values, the A_r, are symmetric about the (N/2T) value and therefore have only N/2 independent values. ($A_i = A_{n-i}$ for $0 \leq i \leq N/2$).

The computational reduction obtained by the FFT is basically derived from decomposing the indices r and p into their prime factors and then taking advantage of the circular symmetry of the resultant exponential terms in the summation.

For example, for any term $0 \leq p < 2^m$ the binary representation $p=p_{m1}2^{m-1}+p_{m2}2^{m-2}+\ldots+p_1 2+p$, where $p_i = 0$ or 1. In compact notation we can write $p=(p_{m-1}p_{m-2}\cdots p_i p_o)$ as a "conventional" binary number.

A similar procedure can be done for r. After a reasonable amount of manipulation*, and using compact binary notation can be rewritten

$$A_r = \sum_{p_k=o}^{1} \exp(-2\pi j p_k \frac{(r_k r_{k+1}\cdots r_{m-1})}{2^{m-k}})$$

An illustration for the case of N=8, M=3 will now be considered.

$$A_{(r_o r_1 r_2)} = \sum_{p_o=o}^{1} \exp\left((\tfrac{2\pi j}{8})p_o(r_o r_1 r_2)\right) \sum_{p_1=o}^{1} \exp\left((\tfrac{2\pi j}{4})p_1(r_1 r_2)\right)$$

$$\sum_{p_2=o}^{1} \exp\left((\tfrac{2\pi j}{2})\, p_2 r_2\right)\; X_{(p_2 p_1 p_o)}$$

Using the notation $\exp(2\pi j \frac{d}{e}) = W_e^d$ the terms reduce to

$$A_{(r_o r_1 r_2)} = \sum_{p_o=o}^{1} W_8^{p_o(r_o r_1 r_2)} \sum_{p_1=o}^{1} W_4^{p_1(r_1 r_2)} \sum_{p_2=o}^{1} W_2^{p_2 r_2} X_{(p_2 p_1 p_0)}$$

The value of $A_{(r_o r_1 r_2)}$ is iteratively computed according to this equation in M=3 stages. At each state only two values of X are required for each intermediate A value. For each intermediate level it is possible to compute a pair of current results from a pair of previous results directly requiring no additional storage.

The general iterative step in the computation of the FFT involves a pairwise operation in which two new intermediate results are computed from a pair of results from the previous step.

*K. W. Drake, Interoffice Memo, Bell Aerospace Company, June 1971.

The system concept presented below uses digital techniques to provide compatible generation and display of real time frequency spectra from audio range signals. First, the system configuration, performance parameters, and basic logic and timing features are delineated. Signal resolution, display features and the pertinent operating parameters are given next. Detailed logic and timing is then presented for the digital processor incorporated in the system.

The system contains four basic elements, a console, a color display, operating controls, and electronics, and can be easily packaged in a desk type unit. The color display can be a 525 line TV industrial type display (300 line resolution) of approximately 12-inch size as determined best for the anticipated viewing distance. A turntable mounting can allow orienting the viewing screen as required. All operating controls for display parameters and input levels adjustment and monitoring are placed in a separate control panel to allow manipulation while monitoring the display. Both the analog and digital electronics along with their associated power supplies will be mounted in separate pull-out electronics drawers. These drawers should incorporate quick removal connectors and cables. All circuitry should be mounted on individual functional boards to enhance the ease of maintenance.

The system will sample an analog acoustic signal at a rate of 8000 samples per second and convert each sample to a 6 bit (2's compliment coded) digital word.

The output rate will be consistent with the standard TV rate of 30 frames per second. Each TV frame will display approximately the most recent 2.7 seconds of real time data as shown in Figure 13. The scan line-to-scan line time difference in display time is 10/8000 seconds (.00125).

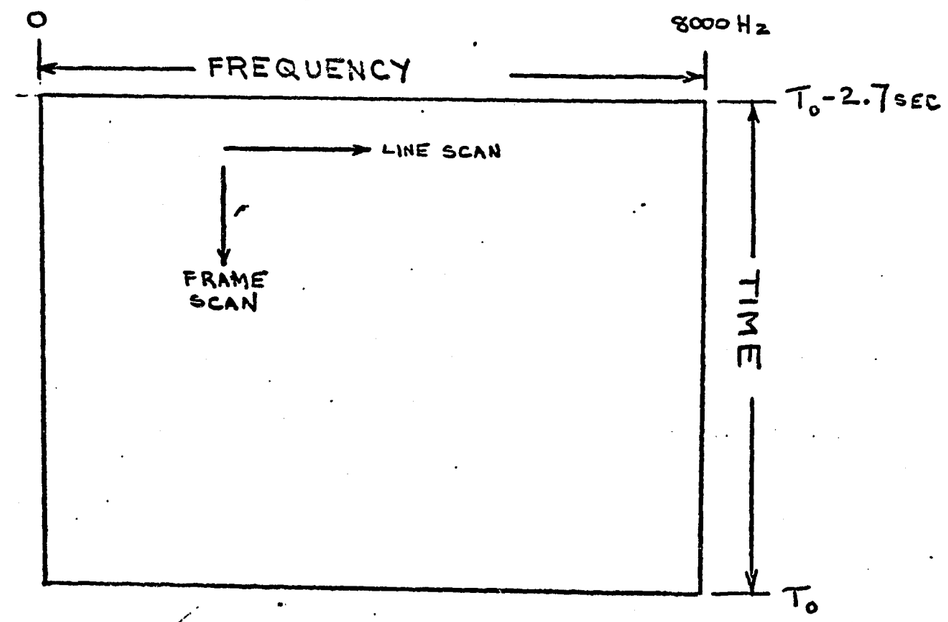


Figure 13 TV Dispaly Format

The real time data signal is sampled at an 8000 Hz rate based upon the assumption that the signal is frequency limited to below 4000 Hz. The FFT of the data is assumed made upon segments of 256 samples yielding a frequency resolution of 8000/256 (31 Hz) between 0 Hz and 4000 Hz. These numbers imply that the FFT will operate on sequences of 256 samples for each scan line of the display. Each successive scan line will have ten samples replaced by new values. For each new TV frame there will be 60 new samples requiring the computation of 6 new transforms for 6 new lines of display.

The system operation is best described in reference to the TV frame rate of 1/30 second. During this basic time period the following must occur (not necessarily in the order listed).

(a) Sixty new data samples are taken and placed in the converter memory.

(b) The sixty samples from the last cycle are shifted into the transform memory and the sixty oldest words are dumped.

(c) The FFT takes 256 word records from the transform memory and produces 128 word frequency spectra. There is a total of six records displaced by ten words (256 + 50) in the 306 word memory.

(d) The display memory is shifted six records (i.e., scan lines) with the oldest six records dumped and the newest six loaded from the FFT output.

(e) The display memory is scanned once producing a 500 line by 128 words per line by three bits per word TV frame.

The functional blocks, Figure 14, are now described in more detail. The signal conditioning circuits consist of preamplifiers and level control amplifiers for the raw acoustic signals as well as the monitor amplifiers. An A/D converter is also contained in this section and operates under control of the master logic. Conversion back from D/A is accomplished in the color encoding circuits (previously described) which also level select the signal and provide the color drive to the electronic monitor. The monitor is a standard 525 line three color TV display which has all color and synchronization signals digitally generated from the master logic.

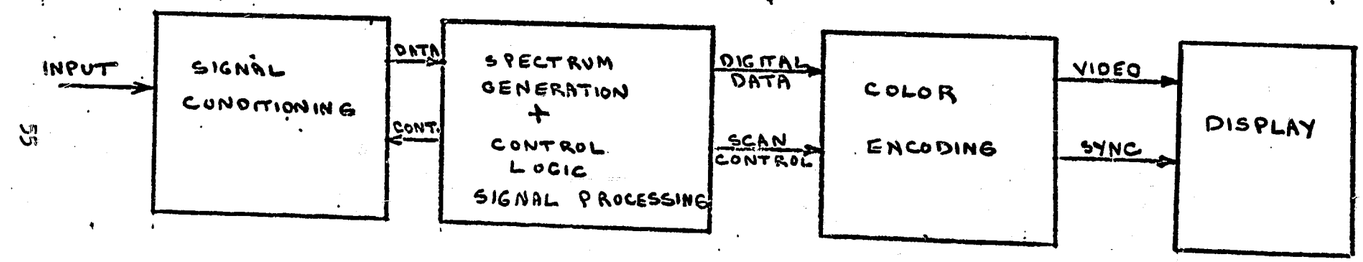


Figure 14 Functional Block Diagram

The Signal Processor Unit utilizes digital filtering techniques to produce a time varying spectrum for display on a television raster. In particular, the signal processing implemented is the Fast Fourier Transform (FFT) so that a real time display is produced. In general, this block accepts inputs from the Analog-to-Digital (A/D) converter on the left. The digital information is processed through this system from left to right in the block diagram. Each line of information for the TV display is outputted through a D/A converter into the horizontal and vertical scanning circuitry. In addition to the displayed information, television control signals are provided to sync the television to the internal clock in the Signal Processor Unit. This unit is composed of only digital hardware which is run synchronously from an internal frequency standard. All the computations and shifting operations, as well as the Analog-to-Digital (A/D) and Digital-to-Analog (D/A) operations are controlled by an internal master clock, divided down to produce the correct control frequency.

Under control of the Signal Processor Unit (SPU), the A/D converter is commanded by the Start Convert command to begin an A/D conversion. When this data is ready, the End of Convert signal is returned to the SPU. Upon receipt of this signal one shift pulse is gated through shifting the digital input word. This A/D conversion cycle is done at a rate such that 60 data words are in the line computation cycle of 31.808 m sec. During the acquisition time of the present 60 data words the previous data words are in the computation registers.

The computation iteration begins when the 60 input data samples are transferred from the input registers into the next set of registers. Fifty of the data input samples are stored in one set of registers while the first ten samples are stored in still another set of registers. By the use of this shifting technique the 256 words of data are updated with ten new samples, shifting the oldest ten samples from the register. These updated 256 data words are now serially shifted.

At the completion of the required 256 shift pulses to load the computation registers, the computation cycle is initiated. The computation cycle is composed of four clock periods per data word or 1024 clock periods for the complete 256 data word computation. The computation process consists of many multiplications and additions in order to produce each FFT constant. In particular, the functions A+WB and A-WB must be generated for each data word pair. The W is a constant supplied from a Read Only Memory (ROM) during the computation cycle.

Because the FFT computation is a process requiring the use of complex arithmetic, each data word has both a real and an imaginary part and thus two sets of computation registers are provided. Since in general, both W and B are complex quantities it is required that four multiplicative operations take place. That is $B_R W_R$, $B_R W_R$ and $B_I W_I$ must be formed. (R = real, I = imaginary.) A ROM is used to store and access W_I and W_R for higher speed multiplication to form these terms. Each of these products is stored in turn into four word Scratchpad memory. At the completion of this multiplication process, the real and imaginary terms are added along with the A_R and A_I terms to produce the terms

$$A_R - (B_R W_R + B_I W_I) \quad \text{and} \quad A_I - (B_R W_I + B_I W_R)$$

and

$$A_R + (B_R W_R + B_I W_I) \quad \text{and} \quad A_I + (B_R W_I + B_I W_R)$$

The computation and shift cycle is repeated for eight iterations.

At the completion of this cycle, the 256 FFT terms are contained in real and imaginary registers. Then, the (magnitude)^2 term is produced by applying the A_R terms to both parts of the multiplier to produce A_R^2. Similarly, the A_I terms are also applied to the multiplier to produce A_I^2. Then, the sum $A_R^2 + A_I^2$ is produced and placed in a register. After the required shifting operation to return the data to the correct position, the square root operation is performed. It is noted that the FFT process implemented here results in 128 FFT constants while the remaining 128 FFT terms are redundant. Thus, the operations to perform the (magnitude)^2 and magnitude operation are only on the first 128 numbers in the shift registers.

The basic operational devices in the Signal Processor Unit are the arithmetic units. The operations that must be performed are multiplication, addition, subtraction, and square root. The addition and subtraction operations should be instrumented with dual carry-save adders in conjunction with look ahead carry logic elements to produce the sum and carry with minimal gate delays. For the multiplication, it is possible to reduce the total delay through two logic levels to approximately 65 nseconds. The square root operation may be performed by using the multiplier in conjunction with a binary up counter and comparator. The counter state is applied to both multiplier inputs producing the (counter)^2 term. The comparator is used to indicate when the counter state is equal to or greater than the (magnitude)^2 term. That is, the counter is incremented until the (counter)$^2 \geq$ (magnitude)^2. At this time the counter is read and the square root of (magnitude)^2 is present in the counter.

The three most significant bits of the 128 FFT terms are now shifted into other registers and circulated for reordering of the FFT constants. At the completion of 256 circular shifts, the numbers appear in correct order. This 128 bits of information represents one line of data on the TV display. This data is temporarily stored to synchronize transfer into the TV line registers.

At the same time data is being shifted into the proper location after the magnitude operations, ten new samples are entered. This 8 iteration computation cycle is repeated. In 31.808 msec, 6 128 FFT constants are produced. Since the television frame rate is 1/30 sec., or 33.3 msec, 1.5 msec is used as a wait period before the next computation cycles are initiated.

In actuality, however, the TV system is an interlaced system. That is, the odd lines are scanned first in 1/60 second and then the even lines are scanned in the second 1/60 second period.

Each of the lines is clocked out is sequence (odd then even) by a high frequency clock. The output multiplexer in conjunction with the clock control places the selected data output line into the D/A converter. As each 3 bits of data appears from the shift registers, a start convert command is generated by the Signal Processor Unit.

At the completion of the 500 lines of data to the D/A a complete TV display has been generated. As each line of data is read into the TV system it is restored in the same location in the register. At the completion of the frame, while the vertical trace is returning to the start position, the shift registers are rotated six places. Enough shift pulses are applied to properly rotate the data. The new data enters into a register while the oldest data is lost.

In addition to the above shifting control, the SPU provides the required control signals for the TV display. In particular, a vertical and horizontal sync signal are provided, for timing, as well as a horizontal and vertical blanking signal to control the picture tube during retrace time.

It is anticipated that the type of system can be manufactured in limited production quantities (about 100 units) for a price of less than $10,000 per unit. The justification of this price is not based solely on a change from prototype research and development to production techniques. Also to be considered is the virtually assured continued decrease in price of digital components including integrated circuitry, which is of significance since the bulk of the circuitry used in the signal processing is digital in nature.

The entire display matrix may be stored in 128 bit static shift registers, as an example. These registers represent a considerable portion of the total material costs. Within the next several years it is expected that quad and even hex shift registers can be packaged in the same physical area as conventional shift registers at approximately the same cost. As a comparison, if hex units were used, the quantity of shift registers would be reduced from 307 to 52 and would result in a direct hardware savings of approximately $2,000.

As an other example, the multiplier instrumentation presently considered consists of 18 fast carry-save adders and 12 dual input nand gates. However, development is underway to competitively produce a multiplier chip. The next few years, it is expected that a dual multiplier will be available at a price comparable to the price of a dual adder today.

The two ROM's needed for the system require a $1,000 mask charge per ROM. In future units this mask charge is not present, reducing the total cost by still another $2,000.

In addition, with the development of product techniques and the many competitors in this area, it is quite realistic to expect approximately 10 percent reduction in the cost of IC's over the next three years. Thus, this would again reduce the material costs.

If higher quantities of each item were purchased as for a production buy, the prices could be reduced by at least another 10-15 percent.

Finally, the system design concept is based upon the use of modular construction and modern integrated circuit devices. As a result of this, a minimum of labor costs are required in component assembly. Consequently, it is, at this point, based on such factors as the above quite reasonable to estimate a total system cost of $10,000 for a limited unit production.

f. Conclusions

The effectiveness of color encoded sonagrams, speech spectrograms, or "voiceprints" has been studied. It has been clearly established that the use of color greatly improves the ease and speed with which such recordings may be read. Using conventionally generated data, however, the use of color encoding does not (except for a few db) reduce the errors of classification as may have been anticipated by the reader. This may be explained by considering that the observer who is classifying the original TVSD's is already performing a great deal of processing and feature extraction. The color enhancement of the data acts as a means of preprocessing the information, and thereby reduces the amount of processing required by the observer. As such the advantages of the technique are easier and faster interpretations but not an overall lower classification error. The observer apparently just automatically reduces the level and amount of processing he performs to maintain the same overall level of processing between the color preprocessing and his own as he alone provided before.

In a larger scope, the study was severely restrictive in its use of conventional "voiceprints" as raw data. The success using a color encoded format cited in this study is not representative of what is potentially achievable due primarily to the lack of dynamic range in the original data. Other areas of application to which color encoded has been applied has yielded greater improvements because of the improved nature of the raw data used. Moreover, the color enhancement techniques served to point out the variability of the conventionally produced "voiceprint" and the susceptability of it to misclassification. It is for these reasons that the last part of the study was directed at investigating an overall new method of generating a TVSD or "voiceprint" with high resolution in all three axes.

As a result a real time, on line system concept has been developed. Using a digital implementation and a FFT it will be possible to process voice signals so as to produce a TV type display of the TVS or "voiceprint" of the last three seconds of voice information in a color encoded format. (It is now possible to perform a 1024 point transform for 20 KHz signals in real time whereas the proposed system only requires a 256 point transform for 8 KHz signals.) The development of such a system would permit the recording not only of the voice signal but the color encoded "voiceprint" associated with it immediately using inexpensive video recorders. A conventional "voiceprint" can easily be reproduced on the same system if desired. Given a voice recording, the system may be used as a testbed for evaluating certain parameters of the speaker, track formants, etc. It is intended to serve as a laboratory research tool to more critically produce and evaluate "voiceprints" and to measure quantitatively their effectiveness in suspect identification, as well as for field use once the feasibility has been established.

In summary, this study has served to establish the effectiveness of color encoded displays of sound spectrograms, and at the same time has not encouraged the use of conventional "voiceprints" for suspect identification. However, useful and reasonable features or characteristics of a speaker are felt by the author to be detectable and classifiable provided a more effective means of generating and displaying this information is available.

It is therefore, proposed that a program be sponsored by LEAA to provide for the development and realization of a digital real time TV format system using a color encoded display for presenting more meaningful sound spectrograms similar to the one outlined in this report. The results of this study should be viewed positive in the respect that a) they substantiate the basic use of color encoded sound spectrograms displays and b) they provide a direction for developing more effective and realiable means of suspect identification based on voice characteristics.

IV. Professional Recognition and Publicity

The work conducted under this grant has been publicized and/or presented in part in technical papers as indicated in the quarterly progress reports. For convenience, a complete listing of these items is given in this section.

1. A newspaper article "Technological Watson Color Codes Voices" by Grace O'Connor, appeared in the Albany Times Union Sunday, October 18, 1970 which described the intent of the LEAA sponsored research.

2. Two radio interviews with Dr. Gerhardt were conducted on WRPI, Troy, N.Y. on November 29, 1970 and December 6, 1970 discussing his work on voiceprint classification sponsored by LEAA.

3. An invited paper "Processing and Display of Time Varying Spectral Information with Application to Voice, Sonar, and Medical Signals" was presented by Dr. Gerhardt at the XXIst Air Force Avionics Panel meeting in Rome, Italy, May, 1971. In part, it described voiceprint research and credited LEAA with partial support of this work.

4. Dr. Gerhardt was invited to participate and present a paper "Voice Processing Research at RPI", at the upcoming Rome Air Development Center Workshop on Recongition Problems in Speech, September 22-23, 1971, Rome, N.Y.

5. Dr. Gerhardt was invited to organize and chair a session on "Signal and Image Processing for Societal Problems" for the 1972 IEEE International Convention, March 1972. One of the prime areas of application is criminalogy and suspect identification.

In addition, several discussions were held with the NYS Police Laboratories and the NYS Attorney General's Office and Dr. Gerhardt on the subject of voiceprint recognition based on the work being conducted under the LEAA grant.

In all of these, LEAA was always cited as the sponsoring agency. It is felt that in all cases this resulted in favorable publicity for the LEAA pilot grant program.

V. References

1. "Signal Theory in Speech Transmission", E.E. David, IRE Transactions on Circuit Theory, December 1956.
2. "The Voiceprint Mystique", P. Ladefoged and R. Vanderslice, Working Papers in Phonetics, UCLA, November 1967.
3. "Application of Spatial Filtering to Speech Recognition", D.L. Thorne thesis, Air Force Institute of Technology, June 1970.
4. "Synthetic Voices for Computers" J.L. Flanagan et al, IEEE Spectrum, October 1970.
5. "An Experiment on Voice Identification by Visual Inspection of Spectrograms", O. Tasi et al, Michigan State University. Presented at the Acoustical Society of America 80th meeting, Houston, Texas, November 1970.
6. "Automatic Word Recognition" G.L. Clapper, IEEE Spectrum, August 1971.
7. "Voiceprints Under Fire", D. Vineberg article in Montreal Star, August 28, 1971.
8. "Thermal Imaging with Real Time Picture Presentation", Borg, Sven-Bertil, Applied Optics, Vol. 7, No. 9, p. 1699, September 1968.
9. "Diagnostic Thermography", Barnes, R. Bowling, Applied Optics, Vol. 7, No. 9, p. 1673, September 1968.
10. "Color This Brain Visible", Life, vol. 66, no. 9, p. 28-30, March 7, 1969.
11. 'Big eye' in the sky: Sensors to monitor earth, deAtley, Elizabeth, Electronic Design, 7, p. 25, April 1969.
12. "Thermography: Coloring with Heat, Time, Science, p. 46, August 17, 1970.
13. Datacolor Systems, Spatial Data Systems Product Information

END