A SYSTEM FOR SPEAKER IDENTIFICATION

Investigators

Harry Hollien, Ph.D.
Mark Yang, Ph.D.
Donald G. Childers, Ph.D.
Ruth Huntley, M.A.

FINAL REPORT

Grant: 84-IJ-CX-0014

Report Prepared By:

Harry Hollien, Ph.D.

:ute for Advanced Study of the Communication Processes
University of Florida
Gainesville, Florida 32611

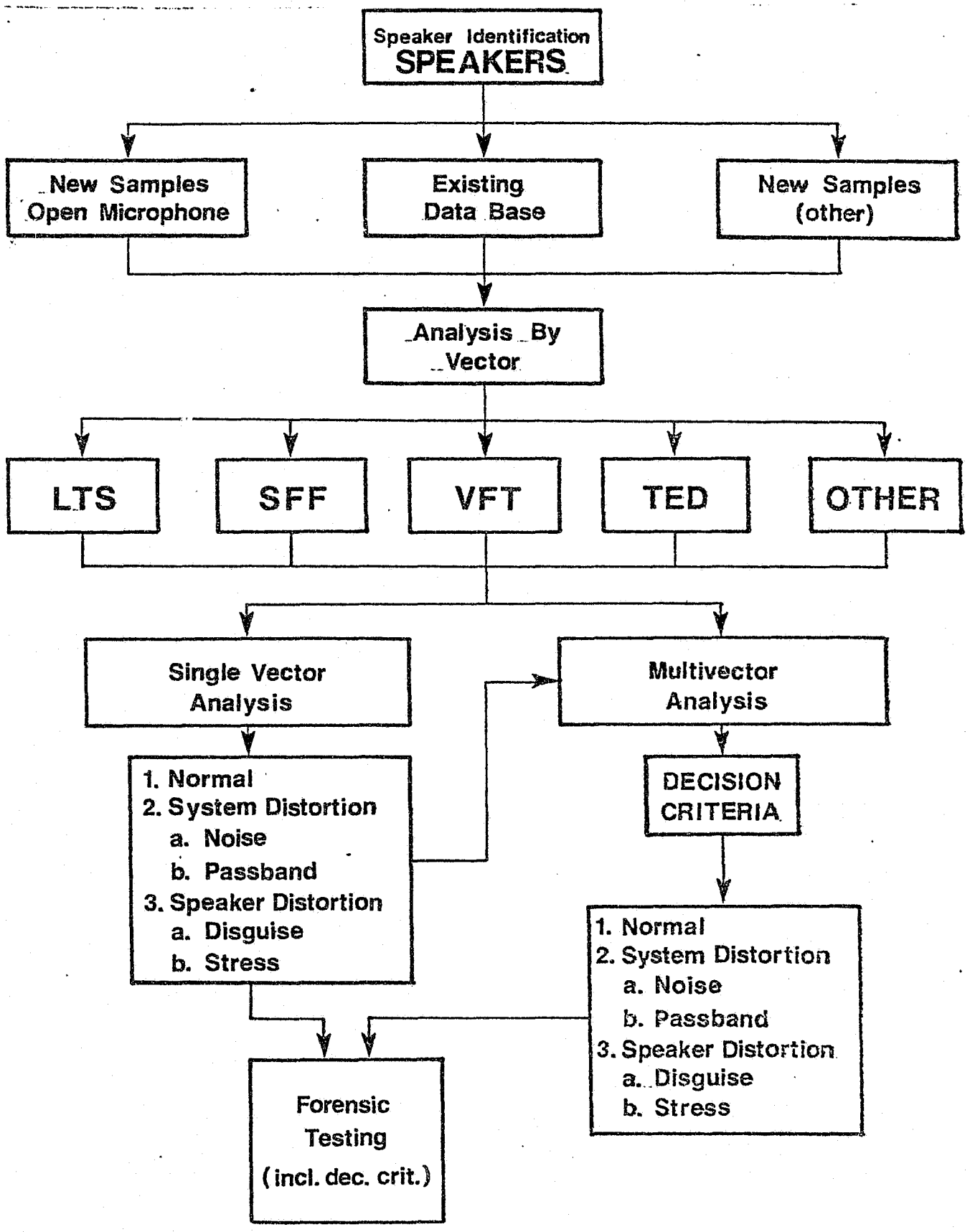
December 31, 1985

ABSTRACT

The basis of this research is the theoretical construct that the speech signal contains features which are sufficiently unique and consistent within a given individual -- and reasonably different among individuals -- to permit successful speaker recognition. That is, both data and logic permit the assumption that certain elements within a talker's speech are relatively idiosyncratic and discriminative as a result of his or her basic glottal and supraglottal structure and the habituation of patterns used in speaking. Social, economic, geographic and educational factors as well as maturation level, psychological/physical states, sex and intelligence all affect speech patterns in specific ways and can combine with those idiosyncratic attributes of the talker's anatomy and physiology to create recognizable features in that person's speech and voice -- ones that can be used in the recognition process. Second, it also is our position that, while there may be no single attribute within a person's speech of sufficient magnitude to permit that individual to be differentiated from all other talkers in all situations, the use of groups of features will permit the recognition process to occur. Third, we postulate that the simple analysis of captured signals (i.e., signal analysis by conventional means) will not permit successful speaker recognition -- especially if channel or speaker distortions are present. Rather, to be successful, a speaker identification system must focus on the natural features within a talker's speech -- that is, if the resulting procedures are to prove robust for criminal justice and law enforcement purposes. Finally, if a speaker identification system is to be employed in the field it must be based on results from systematic research at two levels -- i.e., those that employ 1) the basic research model (investigation of the basic relationships among the parameters and vectors applied to the task) and 2) the forensic or field model (evaluation of the procedures/techniques developed for controlled experiments of this type parallel to, or occuring in, the forensic milieu). Our generalized approach to the problem may be best understood by consideration of Figure 1. As can be seen, human speech samples (drawn as needed from a variety of controlled sources) are analyzed by a variety of methods (i.e., vectors or multidimensional sets of parameters) either singly or in combination, and under a variety of distorting conditions. These procedures are followed for either the basic research or forensic models. Finally, the ultimate goal of this project has been, and is, the development of a valid/reliable (semiautomatic or computer assisted) speaker identification system appropriate for use by criminal justice and law enforcement personnel.

As stated, the approach utilized for the cited purposes incorporated both basic and applied experiments. Subject populations included large (N=25) homogeneous groups of males (i.e., similar dialects, socio-economic backgrounds, education, size, and so on) as well as smaller sets of subjects (known and unknown talkers plus 6-10 foils) both randomly drawn from a large male population and chosen for confusability from prior experiments (i.e., multiple "sound alikes"). As may be seen from the report to follow, yet other groups of male speakers were chosen where warranted and some preliminary research was carried out on female talkers also. Second, the vectors utilized were based on recognizable features in the human voice (also

FIGURE 1. Flow chart of the research approach utilized to develop a speaker identification system for use in the criminal justice and law enforcement milieu. The procedures employed permit appropriate research to be carried out as a function of either the basic research or forensic models.



upon deductive logic and research data); they consisted of 25-40 parameter multi-dimentional -- but related -- sets of speech characteristics. These experimental vectors included 1) general voice quality (long-term spectra or LTS), 2) speaking fundamental frequency level/variability (SFF), 3) vocal intensity level/variability (INT; deferred for improved vector construction), 4) articulation -- both vowels and nasal consonants (vowel formant tracking or VFT), 5) prosodic or time features of speech (rate measures plus time-energy distribution or TED) and 6) vocal roughness (jitter or JIT -- not yet satisfactorily developed). Of the six vectors, the four that have been proven most robust for the speaker identification task are LTS (voice quality), SFF (speaking fundamental frequency), VFT (vowel/syllable features) and TED (speech time/rate). Third, a large number of basic experiments have been carried out as have a number investigations related to the forensic model (either real or simulated "cases") -- plus some highly controlled experiments where field conditions have been rotated/evaluated. Finally, a substantial number of distance measures (simple and multidimensional space), statistical techniques and decision criteria have been studied.

The report to follow provides information about some of the more relevant experiments carried out under the aegis of this grant. While not highly technical, it is somewhat long and detailed. Accordingly, the results of this project -- plus some conclusions and recommendations -- will be briefly summarized in this abstract (please note, however, that necessary clarifications can be found in the body of the report). First, the research to be reported is divided into two sections -- experiments based on the forensic model and, later, studies representative of those carried out in the basic research area. The initial summary is of seven actual civil or criminal cases; all were of field quality and, in several, there was a serious degrading of signal integrity/quality. Positive identifications were made in five cases, eliminations in two. Data from extensive aural-perceptual experiments (large populations; 2-3 groups) proved to be in close agreement with the computer-based machine data in <u>all</u> instances -- and subsequent case disposition appeared to validate the decisions made on the basis of this information. Second, four simulated field experiments were carried out under conditions of severe speech degradation. Correct identifications were realized in all four instances -- and at rather high probability levels. Third, nine seperate experiments were carried out (again under simulated field conditions) and correct identifications were made in all instances. Also reported are 10 experiments where all or nearly all of the foils were individuals who sounded very much like the known talker or had been confused with him in many of the basic experiments. In this case, a second set of variables also was evaluated; i.e., four distance measures. The Steinhaus and absolute distance procedures were found superior to the other two; indeed, 60% correct identification was realized even under these negative conditions. Thus, it can be concluded that the semiautomatic speaker identification procedures resulting from this project display exceptionally good performance in the field even though the basic research experiments (and procedural refinements) are <u>not yet</u> complete!

The results of the several hundred basic experiments funded by the 84-IJ-CX-0014 grant are much more difficult to organize for brief

summary. In this case, protocols were employed (initially anyway) which degraded the entire procedure in such a manner that large response ranges were possible -- thereby permitting the subtle differences among the parameters and the vector strengths and weakness to be efficiently evaluated. Identification of those vectors of demonstrable robustness, modifications for vector improvement and determination of optimum use under specific field conditions also were determined by this approach. Lastly, vector strength under moderate experimental conditions (non-contemporary speech, text independent samples, homogeneous groups of subjects and so on) was carried out.

Briefly, the results demonstrated that the LTS vector is nearly 100% accurate for many conditions and SFF is nearing these levels. Secondly, even though all of the basic research on VFT was carried out on single vowels and nasal syllables, we found this vector to be (potentially) a very powerful one. Indeed, scores here are already so high that it appears that it soon will rank with SFF and perhaps even LTS. On the other hand, while the upgrading of TED has progressed at a somewhat slower rate, the more recent modifications are encouraging as identification scores based on this vector have improved markedly. Moreover, TED is the vector that shows the greatest resistance to bandpass effects and, especially, speaker disguise. To be explicit, the basic experimentational data may be summarized as follows. First, LTS demonstrates 100% correct speaker identification under many conditions; it is resistent to noise, stress and related degradations; the LTS weakness includes decreased effectiveness for telephone bandpass and speaker disguise. Efforts now will be made to improve this vector's performance in these two areas without degrading its strength in the others. Second, SFF is now demonstrating robustness in many areas; however, it is most vulnerable to register/voice changes in disguise -- especially in different psychological environments -- and limited bandpass occasionally results in somewhat unstable SFF data. As with LTS, research focused on these deficiencies is underway (but awaiting funding). Third, the VFT vector appears to be a potentially powerful one but occasionally is difficult to extract from noisy signals. Future thrusts here will be designed to evaluate its strength when multiple stimuli are employed and to see if the vowel formant ratios will permit the high identification levels expected. Also to be researched are several digitization techniques and especially which of them will overcome the occasional instability found in VFT. Fourth, it is apparent that the TED vector should be expanded to include more prosodic parameters. We are especially desirous of improving the general performance of this vector (especially for speaker stress) without degrading its robustness as an identification cue in the areas of disguise and telephone bandpass. Finally, the issue of combining the vectors is a critical one. As it turns out, only modest improvement in identification currently is realized when vectors are combined. However, we are beginning to discover that this situation arises because the strength of a vector varies in different environments and one or two of them prove dominant in most configurations. Thus, we now believe that it is important to conceptualize the SAUSI (semiautomatic speaker identification) profile in <u>two</u> ways. First, it will be necessary to develop profile configurations that fit each of a number of rather specific forensic situations and second, it also will be important to improve the mathematical process of combining the vectors within multi-

multidimensiona space.

A final comment about the basic research thrust carried out under the aegis of this grant seems warranted. As may be seen, a substantial number of studies focused on the four distance measures and on several statistical procedures which have been carried out. This research also should be continued as several new distances must be evaluated and a number of the novel statistical procedures (developed by Dr. Yang) should be evaluated. In any case, currently it appears that the Steinhaus distance measures are superior to the three others. Nevertheless, it is necessary to determine which of the distance measures works best and under which conditions. Once available, these data/procedures can be combined with the vectors in the SAUSI profile and applied differentially to the cited forensic situations. In any case, data outlining many of these relationships can be found detailed in the report to follow.

A few of the more relevant conclusions should be included in this abstract. First, it is now obvious that the natural speech feature approach is one of merit -- and perhaps even the only viable approach currently available to speaker identification (and speaker verification) in the field. Second, the systematic research approaches applied here appear necessary if a good speaker recognition method is to accrue; moreover, the two level (basic/forensic) approach has been found to be quite effective. Third, it now is quite apparent that a profile approach will be necessary if speaker identification is to be carried out successfully in the field. Fourth, the vectors we have chosen appear adequate to the task (they seem especially effective in field situations). Fifth and perhaps most important, the statements made above can be confirmed in a great measure by the exceptionally good (and somewhat unexpected) performance of SAUSI in the field. Finally, even though our method already appears adequate for field use, it has not yet been refined -- and hence additional research is necessary if premature application is to be avoided.

INTRODUCTION

There are three independent yet related areas within the Communication Sciences that are of substantial importance to criminal justice and law enforcement; they are speech recognition, speaker verification and speaker identification. Even though rather substantial progress has been made with respect to the development of on-line methods related to the first of these problems -- and modest progress re: the second -- the fact remains that there are no independent systems currently in existence that permit speech/speaker recognition tasks to be carried out. The problem is complicated by the fact that nearly all of the research being carried out in these three areas is concentrated in the first two. The reasons for this situation are clear. First, the task is a formidable one. Speaker identification -- unlike speaker verification -- always involves an "open" set of suspects (i.e., the criminal may or may not be among subjects in the set), yet one of the subject/suspects is likely to be selected as the individual most similar to the unknown anyway. Second, the signal usually is degraded by system or channel distortions such as noise; limited bandpass and so on and/or by speaker distortions such as disguise, stress and so on. Moreover, the forensic model is one where the process may not be text-independent; it certainly involves non-contemporary matches. Third, there are social implications that sometimes tend to discourage relevant agencies from supporting research in this area.

It is the fourth problem associated with speaker identification that is of greatest importance. Specifically, human speech has been thought to be so variable that there may be no characteristics within the (resultant) acoustic wave which would permit reasonable levels of identification. Yet, it can be observed that, from time-to-time, every normally hearing individual is able to recognize known talkers from the perception of their speech alone; thus, the logic that speaker identification is possible by signal processing would appear irrefutable. Moreover, the results of some completed experiments would suggest the argument that analysis of combinations or groups of speech features will operate powerfully (where single parameter analysis would fail) and permit speaker identification to be carried out on a scale far greater than previously considered possible. Questions remain, however, concerning the potential universality of speaker recognition, the identity of those features that can be most successfully used as cues, the best parameter combinations for each environment and the practical application of those procedures or systems which result from research in this area. In short, two basic questions must be addressed: 1) Is interspeaker variability always greater than intraspeaker variability (i.e., are there certain speech features so idiosyncratic to an individual that identification always is possible) and 2) can these features -- or a profile based on them -- actually be applied to the forensic model. In response, it appeared that both basic and applied research should be pursued in this area -- and such was our approach re: the 84-IJ-CX-0014 project.

Progress has been substantial and, while we concede that there are still some basic (and applied) questions to be answered, it now should be possible to use our procedures in the field -- at least in

a limited way. Moreover, if our procedures are properly applied, there should be a reasonable probability of useful results. We base our position in this regard, in part on the available data and in part on our successful use of the forensic model. That is, while we concede that there may not be a single speech vector which will permit very high levels of correct identification, there appear to be a group of vectors (i.e., a profile) which will permit successful application of the process. Please note, however, that we also concede that additional research will have to be carried out to further validate our procedures, refine them and develop an easily used set of field techniques.

AN APPROACH TO SPEAKER IDENTIFICATION

To date, three general approaches to speaker identification have been utilized; they are: 1) aural/perceptual recognition, 2) spectrogram matching and 3) machine recognition. Perhaps most is known about the aural/perceptual approach. While we concede that definitive data have not yet been reported -- even for small scale sorting and matching experiments of restricted voice samples, we have found it possible to base our efforts on a tentative aural/perceptual model. Indeed, we have found that, in aggregate, most of the strategies used by humans in the identification task have been defined and, thus, coupled to our models, can provide the substructure to our approach. That is, we were able to determine that the following parameters are used by listeners in the speaker identification process: 1) speaking fundamental frequency (f0) level and variability (vocal pitch), 2) speech spectra (voice quality), 3) nasal resonance (articulation), 4) vowel placement (articulation), 5) vocal tract turbulence (quality of speech) and 6) prosody (speech timing/rate). As will be seen (below), the research to follow is organized in response to these observations. That is, we argue that it is only the natural features of speech that will resist channel and speaker distortions. Our results appear to demonstrate the strength of this contention.

It should not be necessary to comment on the second approach to speaker identification -- i.e., the spectrum matching or "voiceprint/gram" approach. As a method, it has been described/defended by a very few authors and studied and/or attacked by numerous scientists. Moreover, when the data from all reported experiments are combined and synthesized, the inescapable conclusion is that the "voiceprint/gram" method of speaker identification is not a valid one. However, data accumulated about vowel/consonant formants suggest that they contain useful identification cues and this relationship is of some importance in our model.

We selected the third approach to speaker identification as the only practical one possible -- and for several reasons. First, we noted that nearly all research being carried out in the area of speaker verification is machine oriented. Second, we believe that there is no practical way that human auditors can be utilized objectively to carry out speaker identification tasks and that modern technology has eliminated the need to do so. In any case, we would argue that appropriate quantitative (machine) processing of human

speech (in order to identify those natural speech features idiosyncratic to the individual) will lead to the most robust method of speaker identification. Indeed, the research to be reported will serve to verify this contention.

To summarize; we now can assume that speech is so unique to the individual that a selected cluster of features can be used as stable, valid identity cues and that natural speech features are among the most powerful signal-contained elements for this purpose. In response to these assumptions, we developed and tested a number of vectors for this purpose.

OBJECTIVES

The basic purposes of this project were to obtain data on the process of human speaker identification and to test the usefulness of a speaker identification system. Also considered important were efforts to test several models of speaker identification, distance evaluation schemes -- and, especially, our postulate that a natural speech feature approach is potentially robust enough to be effective in the field.

To be specific, our objectives were to develop a systematic and sensitive approach to speaker identification by:

1) Identifying, structuring and analyzing those speech features (vectors) that could serve as predictors of a speaker's identity,
2) Evaluating the profile or multiple vector approach to the identification task and determine which configurations were resistant to distortions of various classes and types,
3) Adapting and analyzing various data storage/retrieval and statistical procedures for specific use in decision criteria,
4) Structuring and testing the forensic model as an assessment of the robustness of our techniques for field use.

In short, we believe that our systematic approach to the problem, plus the data generated, have resulted in development of a number of useful constructs and provided sharply upgraded knowledge about the problem. Further, this effort has permitted testing of several speaker identification models and the development of a field system that promises (after additional refinement) to be a reasonably efficient one.

METHOD

As stated, the basis of this project has been the theoretical construct that the speech signal contains features which are sufficiently unique to a given individual -- and enough different among individuals -- so as to permit effective speaker identification to be realized. That is, both data and logic permit the assumption that certain elements within a talker's speech are relatively idiosyncratic -- that they result from habituation of speaking patterns which, in turn, are based on social, econcomic, geographic

and educational factors as well as maturation level, psychological/physical states, sex and intelligence. In any case, we maintain that these factors combine with behaviors related to the talker's anatomy and physiology so as to create recognizable features within a particular person's speech and voice. Further, this position is based on two postulates that: 1) natural speech characteristics provide the most robust identification cues and 2) while no single speech attribute may be powerful enough to permit differentiation from all other talkers, the use of feature groups will permit the recognition process to occur -- even under unfavorable conditions. A brief review of the general methods employed will proceed specification of the results.

The Data-Base

One of the features of this project was that we already had assembled a large portion of the required data-base. In all, recordings were available of 250 men and 136 women (N=386) who produced speech samples of several types (N=4669). This data-base was established so that the effects of speaker and system/channel distortions upon our vectors could be evaluated and identification cues studied in a variety of environments. The primary system distortions included: A) limited passband (including a "telephone data-base") and B) noise, with talker distortions including: 1) several types of stress, 2) disguise (nine types), 3) dialect (two dialects), 4) sex, 5) age and 6) speech materials (four types). Field conditions were simulated by combining system and speaker distortions or by using tape recordings from actual cases. Finally, these 4,669 samples were supplemented as needed (see protocols listed below).

The Experimental Vectors

Even though it was obvious that analysis of certain speech features would lead to the successful identification of talkers, the specific features that would prove most useful had not been identified. Hence, we evaluated available information on fundamental frequency, glottal volume velocity, vowel formant frequencies/bandwidths, turbulent phonemes, nasal consonants, prosodic/timing features and the similar elements; those vectors to follow were selected for intensive evaluation.

Long-Term Speech Spectra (LTS). The use of power spectra as an identification cue has had a relatively long history in this area. We postulated LTS to be an index of general voice quality and that, as such, it is a good cue to a speaker's identity. As will be seen, we have discovered that LTS can correctly predict the identity of speakers at very high levels; at least when data are normalized and for laboratory type research. We also have been able to demonstrate that LTS is relatively resistant to the effects of speaker stress and a varied of noise conditions. However, it does not function well as a predictor of identity when talkers disguise their voices. The LTS data extraction system includes a Princeton 4512, FFT spectrum analyzer coupled to our PDP-11/23 computer; we hope to use ILS (also) in the future. The vector utilizes 40 parameters to generate a power spectrum curve covering a frequency range of 60-10,240 Hz. A number of mathematical "distances" provide the desired statistical

comparisons.

Speaking Fundamental Frequency (SFF). While perception of f0 has been shown to be a reasonable cue for speaker recognition, feature analysis has been only mildly encouraging as researchers have reported varying degrees of success with it. Accordingly we developed a 30 parameter SFF vector and results have been substantially good. The parameters making up this vector include SFF mean, SFF standard deviation, the number of semitone (ST) intervals containing energy plus the number of waves in each of the ST interval "bins." Fundamental frequency data are obtained automatically by means of FFI-8 -- the output of which was fed directly to the PDP-11/23 computer. FFI-8, a digital readout f0 tracking device, consists of a series of successive low-pass filters, with cutoffs at half-octave intervals, coupled with high-speed switching circuits which are controlled by a logic system. FFI measures each wave (it does not "sample") by producing a string of pulses -- each pulse marking a boundary of a fundamental period from complex speech waves -- which are delivered to the computer. An electronic clock marks the time from pulse-to-pulse and these values are processed digitally to yield (among other data) the geometric mean frequency level and standard deviation of the frequency distribution. In order to provide for the new multiple-parameter SFF vector, a subroutine was written to permit the semitone interval analysis; further semitone histograms and frequency or period information were used to compute the actual features such as mean value, modal value, modal frequency, variance and histogram entropy.

Vowel Formant Tracking (VFT). Much use has been made of vowel formant center frequencies, bandwidths and transitions by individuals using time-frequency-amplitude spectrographic techniques in speaker identification. Admittedly, that approach is not a very sophisticated one; however, the research conducted by these individuals has suggested that vowel formants and their nature definitely are important to the speaker identification task. We have upgraded the procedure and are using a Princeton 4512 FFT coupled to our PDP-11/23 computer for this purpose (again we wish to use ILS in the future). In any case, research in both the aural/perceptual area and relative to study of the issue via machine approaches can be used to argue the importance of elements within the formants as speaker identity cues. When selecting the vowel characteristics for this vector, we focused our efforts on several features: 1) the center frequencies of the first three formants (F1, F2, F3) and 2) ratios among these three formants (F1/F2 and F2/F3). As will be seen, it appears that F1, F2 and (perhaps) F3 center frequencies are robust indicators of speaker identity -- (especially when several vowels such as /i,a,u/ plus the syllable /na/ are used as stimuli). Initially, we used spectral analysis protocols (Princeton 4512) with the frequency/intensity of those three energy peaks corresponding to the first three vowel formants used as recognition cues.

As may be seen from the report to follow we have not yet been able to assess the robustness of vowel formant ratios (in process). Nevertheless, we would suggest that they are important for the following reasons. Formant frequencies are generally dependent on

the size and shape of the vocal tract. Therefore, they are based both on anatomy and vocal tract articulatory movements. However, since the range of formant frequency shifts and the F1/F2 and F2/F3 ratios probably cannot be significantly altered at will, they should convey idiosyncratic information about a specific talker. Further, the relative "position" of a speaker's vowels (as determined by both the formant frequencies and ratios) also should be idiosyncratic of an individual's overall speech pattern. Finally, it would appear that formant frequency analysis/ratios appear resistant to many kinds of system distortions and, therefore, should be useful for speaker identification purposes even though those types of interference are present. As stated, however, we have only been able to extract the necessary ratios (by the end of the current grant) and have not analyzed them. Hence, they cannot be used to contribute materially to this report.

Temporal Vector (TED). Very little research on speaker identification has focused on any of the temporal parameters that can be found within the speech wave; and there are but few exceptions here. Nevertheless, there is strong logic that there are prosodic speech elements that can be extracted and used for recognition purposes. For example, given the hypothesis that talkers differ with respect to the durational use of acoustic energy in speech, it is possible that the amount of time a speaker is or is not producing an acoustic signal during a specific amount of connected discourse may be useful in the identification task. Moreover, certain individuals appear to employ a greater number and/or longer silent intervals in producing a linguistic message than do others. In any case, a rather substantial number of temporal speech features appeared amenable to study; we selected the following: a) Total Speech Time (TST) -- defined as the period (in ms) it takes to produce an utterance of a set number of syllables, b) Speaking Time Ratio (S/T) -- defined as a measure of the total time for which acoustic energy is present during a set utterance, c) Silent Interval (SI) -- a reciprocal of speaking time (ST), d) Speech Rate (SR) -- a measure of the speech material completed during a fixed time period (based on syllable rate not word rate) and e) Consonant/Vowel Duration ratios (C/V) -- the (time) ratio between a particular consonant and a vowel in a specified CV utterance. An additional temporal vector also was evaluated; specifically, a time-energy distribution (TED) vector which reflects the total time a talker's speech bursts remain at specific energy levels (relative to his peak amplitude). This also provided an indication of the speaker's speech pattern with respect to speech bursts and pause periods. The TED vector is totally software based; all digitization is obtained utilizing our PDP-11/23 computer with an A/D converter.

Other Vectors: The above cited four vectors provided basis for the primary thrust of this research. However, a number of other vectors also were researched but only on a preliminary basis; they included: 1) vocal intensity (deferred), 2) jitter, 3) shimmer and 4) phonemic unit. However, research here has been preliminary in nature and, to date, no robust vector has been identified from among this group and only a small amount of data will be reported. Actually, a great deal of research was carried out on the vocal intensity vector

(VI). However, it did not prove to be very robust and, hence, research here has been deferred until a more promising VI approach is identified. On the other hand, the jitter vector appears promising -- - at least as a subvector to SFF. In any case, some data for each are reported in the results section to follow. In summary, the natural speech vector approach to speaker identification has proved to be effective -- and at least four vectors have been identified as useful.

Experimental Approaches Utilized:

The experimental approach we have employed is considered somewhat unique; hence it should be emphasized. Fundamentally, research was carried out on two levels; these levels were based on: 1) a basic research model and 2) a forensic model.

The Basic Research Model: The goals of this approach were to cycle the selected vectors through a variety of structured laboratory evaluations in order to study the effects of a large variety of system (channel) and speaker distortions -- and their effect on both the vector and the identification process. Since it was considered desirable to observe differences that could be rather subtle, it was necessary to spread the identification responses over as large a continuum as possible. To be specific, a correct identification range of from 20% to 80% will permit a 60 percentile point dynamic range. Of course, it is conceded that as the vector under study becomes better understood, it also is desirable to assess its maximum potential for correct identification. Nevertheless, the best evaluation of the strengths and weaknesses of a vector can be determined by the process of degrading it in a variety of ways and studying it under these adverse conditions. The maximum identification levels are studied at a later stage (in this case only for LTS so far -- and to some extent SFF).

As would be expected, if an extensive spread of responses is required, the research must be designed appropriately. In this case, a group of 25-30 healthy young males was chosen as the primary experimental population. These males were by-and-large quite similar in education, dialect, size, age, health and so on. Hence, a large, quite homogeneous population was identified and employed in much of the basic research conducted. It should be noted especially that at least five pairs of subjects were relatively close "sound-alikes". A population such as this one will tend to make attempts at speaker identification quite difficult -- especially if channel effects, text, environment and talker status are applied as single or combined environmental effects. It should be stressed again that this approach lends itself equally well to single or multiple vector analysis and evaluation of these vectors under conditions of distortion (i.e., band pass, noise, speaker disguise, etc). In any case, the approach has proved to be quite successful as it has led to: 1) the modification (improvement) of vectors, 2) an understanding of which vectors are most robust under specific speaker/system conditions, 3) the elimination (or deferral) of one vector, 4) the discovery that "wrap-around" (during data

processing) will be a major problem for any computer based approach to the problem and 5) an enhancement of the theory that natural speech features exhibit a robustness not enjoyed by other signal processing approaches.

The Forensic Model: While it is useful to employ a basic research approach to study the conditions under which the experimental vectors will be most powerful and how they can be upgraded, the large/homogeneous population approach cited above is one that will be but very rarely encountered in the criminal justice/law enforcement milieu. Rather, the typical paradigm is one where there is an unknown speaker and a group of 1-N (sometimes 6-10) suspects. An alternate scenario is where there is more than one unknown speaker and perhaps but a single suspect. In this case, it is necessary to provide additional suspects/subjects. In any case, the forensic model is one of match/no match and the most powerful approach is not one-on-one (i.e., not one known vs one suspect) but rather where the unknown voice is compared to the one or more suspects (i.e., the knowns) and the comparisons are made in the milieu of up to 10 foils (or innocent talkers). We included a number of "simulated" studies of this kind among our protocols and, more important yet, have applied the vectors and procedures to several actual cases. In all of these forensic cases, high levels of secondary reliability were included. That is, the actual "unknown" speaker actually was known in the simulated experiments. The results in the actual cases were "validated" by: 1) verification of findings by two or more blind listening panels and/or 2) ultimate knowledge of the actual match between the unknown voice and the suspect. The high level of success in these actual and simulated cases is encouraging. Accordingly, the results based on the forensic model will be considered first.

RESULTS

Forensic Model:

The data generated by the experiments based on the forensic model are presented first. This approach is employed because these results demonstrate that the procedures utilized have merit even though they have not as yet been finalized. Indeed, it still is necessary to 1) identify a final set of vectors for the SAUSI profile, 2) determine the robustness of each under all environmental conditions and combinations (upgrading any that require improvement), 3) evaluate additional distance measures, 4) improve decision criteria and 5) convert all processing to software. Nevertheless, substantial success with this approach appears to have been realized; data from key experiments follow.

Table 1 summarizes the decisions projected for seven actual criminal/civil cases; the tapes for these cases were of field quality but were only poor (noisy) for cases No. 2 and 6. The names have not been revealed for obvious reasons but are available upon request.

Table 1. Summary table of the results of seven actual cases evaluated on the basis of the Forensic Model (details can be provided upon request). Profiles of 3-5 vectors were utilized with LTS, SFF and TED always included. Each decision based on this profile was contrasted with scores obtained from 2-3 listener groups making (blind) ABX judgements. The unknown was compared to all knowns and foils; sometimes there was more than one known suspect. Since information about the suspect (and case disposition) ultimately became available, it is included in the last column.

Case No.	Speakers K-1	K-2	F1	F2	F3	F4	F5	F6	F7	F8	F9	A-P Level/Data Above	Below	%	Apparent* Guilt Yes	No
1	11	NA	-	-	-	-	-	84%	-	NA	NA	X(for F6)		2		X
2	76	-	18	-	-	-	-	-	-	-	NA		X	5	X	
3	88	-	-	-	-	-	-	-	-	-	-	X		3	X	
4	69	61	-	-	-	-	-	-	-	-	NA		X	10	X	
5	-	NA	-	-	57	11	-	43	-	-	-	NA**	NA	NA		X
6	89	82	-	-	-	35	-	-	-	-	NA		X	5	X	
7	86	NA	-	-	-	-	-	-	11	NA	NA		X	7	X	

NA = Not Applicable
\- = Chance levels
A-P = Aural-Perceptual level
* = Apparent guilt is based on case disposition.
** = 70% of all listeners indicated that K and U were different.

The results were accepted by all parties (including the court in question) in four instances, although no trial was held in two. Two of the other cases were internal to a large company and, finally, a court rejected the approach as "premature" in one instance. In five of the cases (2,3,4,6,7), the known and unknown talkers apparently were the same individual (on the basis of external evidence) and in two cases (1,5) they were thought to be two different people. The agreement with the aural-perceptual data was extremely good except for the case No. 5 where listener responses were quite variable. Table 2 provides similar data for simulated field cases; the data are presented in a somewhat different form. Moreover, the known talker was tested against himself in two cases and samples of the so-called "unknown" talker's speech entered twice in two others. By-and-large, these data show a clear cut identification in the proper direction (known with known; unknown with known) except for Experiment No. 2 where one of the foils (F3) was sometimes identified as being the same person as the known talker. It should be of interest to note that this research was completed before work on the intensity vector was deferred and, while data from this parameter set contributed to the decisions made, the contribution was but minimal.

In some ways, the data from Tables 3 and 4 reflect the basic research approach as the experiments are carried out on the 25 (homogeneous) male subjects utilized in that series; i.e., these nine subpopulations are drawn at random from the larger group of men used in the basic experiments. Nevertheless, the forensic model can be seen to be controlling in both sets of studies. Please note also that the data are for a single vector (LTS or power spectra) and good recordings of normal speech were used. In the first of the two sets of investigations (Table 3), speakers were drawn at random with one of them selected as the known. Thus a sample of the "known" or test subjects voice was compared to another sample of his voice plus those of seven other subjects. As can be seen, the known and known were matched in all cases (100% identification). In the second experiment (see again Table 4), an attempt was made to "stress both the model and the vector. In this case, selection for each of the subjects was made as a consequence of evaluation of the basic research data. In all cases the "known" subject was a person whose speech had been seriously confused with at least two other individuals (and sometimes as many as five) and whose identity had been confounded (at least occasionally) with all of the other foils utilized. Thus, the task was to see if the experimental (known) talker could be identified even though other talkers who sounded very much like him were included in the group. Please note also that four distances were compared for robustness as decision approaches. In any case, and as may be seen from examination of Table 4, the absolute and Steinhaus distances resulted in 60% correct identifications (plus several "near misses"); the Euclidean distance approach correctly identified half of the talkers. Admittedly, the speech samples used were of good quality and reasonably contemporary. It is a striking finding, nevertheless, that individuals who sound similar to each other -- and whose identity is often confused -- could be deferentiated in so many cases and on the basis of so little data.

Table 2. Summary table of four field experiments using a four-way profile, i.e., the combined SFF, LTS, INT and TED vectors. In each case, the known talker (K) was matched to each of the suspects (U's) and foils (F's). In Experiment #1 and #3, the known talker was, in addition, matched to himself. Profiles are based on 10 nearest neighbor approach.

Experiment	Subject										
	K	U1	U2	F1	F2	F3	F4	F5	F6	F7	F8
Experiment #1											
Rank	1	3	2	5	4	6	10	9	6	8	NA
4-V Score	1.0	3.8	3.1	6.0	4.5	7.0	8.0	7.4	7.0	7.2	NA
Percent	100	82	86	-	51	-	-	-	-	-	NA
Experiment #2											
Rank	NA	1	2	-	-	3	-	-	-	-	-
Percent	NA	87	63	-	-	53	-	-	-	-	-
Percent profile	NA	91	82	-	-	44	-	-	-	-	-
Experiment #3											
Rank	1	2	NA	-	-	-	3	-	4	NA	NA
4-V Score	1.0	2.1	NA	-	-	-	3.8	-	4.7	NA	NA
Percent	98	88	NA	-	-	-	40	-	36	NA	NA
Experiment #4											
Rank	NA	1	NA	3	-	4	-	-	-	2	NA
Percent	NA	81	NA	36	-	35	-	-	-	44	NA
Percent profile	NA	81	NA	37	-	37	-	-	-	47	NA

K = Known Talker
U = "Suspect" or Unknown
F = Foil

NA = Not Applicable
- = Chance Level

Table 3. Summary table of nine experiments -- utilizing the LTS vector (only) in a typical forensic situation. Subjects were drawn randomly from a homogeneous (age, size, health education) group of young males. A three nearest neighbor approach was used. Note the 100% correct identification.

Test Subject	Experimental Subjects							
1	1 1*	F6 -	F9 -	F11 -	F14 3	F18 2	F21 -	F23 -
5	F2 -	5 1*	F9 -	F14 3	F18 2	F20 -	F21 -	F25 -
9	F4 -	F8 -	9 1*	F10 3	F16 -	F18 -	F22 -	F24 2
12	F1 -	F5 -	F9 -	12 1*	F15 -	F20 2	F24 -	F26 3
15	F3 -	F7 -	F12 -	15 1*	F17 3	F20 -	F23 -	F26 2
17	F2 -	F4 3	F8 -	F15 -	17 1*	F18 -	F21 -	F25 2
23	F1 -	F6 -	F13 2	F14 3	F19 -	F21 -	23 1*	F24 -
24	F2 -	F7 -	F12 -	F15 3	F18 -	F23 -	24 1*	F26 2
25	F3 -	F6 -	F11 -	F14 2	F17 3	F19 -	F22 -	25 1*

1* = correct identification

Table 4. Summary table of LTS (only) matches in a typical forensic situation. Foils were either subjects that sounded quite similar to the test subject or had been systematically confused with him in a number of prior experiments.

Distances	Test Subject	Experimental Subjects						
	2	2	F4	F6	F9	F13	F15	F20	F26
EUC		1*	2	-	-	-	-	-	3
ABS		1*	3	-	-	-	-	-	2
STH		1*	3	-	-	-	-	-	2
MAX		1*	3	2	-	-	-	-	-
	3	F1	3	F6	F9	F10	F14	F19	F21
EUC		-	-	-	-	2	1	-	3
ABS		-	3	-	-	2	1	-	-
STH		-	3	-	-	2	1	-	-
MAX		-	3	-	-	2	1	-	-
	4	F1	4	F5	F7	F8	F13	F19	F25
EUC		-	2	-	-	3	-	-	1
ABS		-	1*	-	-	3	-	-	2
STH		-	1*	-	-	3	-	-	2
MAX		3	-	-	-	2	-	-	1
	6	F2	F3	6	F7	F10	F12	F18	F23
EUC		-	3	1*	-	2	-	-	-
ABS		-	3	1*	-	2	-	-	-
STH		-	3	1*	-	2	-	-	-
MAX		3	-	1*	-	2	-	-	-
	8	F1	F5	F7	8	F16	F19	F24	F26
EUC		2	-	-	-	3	-	-	1
ABS		-	-	-	-	2	3	-	1
STH		-	-	-	-	2	3	-	1
MAX		1	-	-	2	-	-	-	3
	10	F1	F6	F8	10	F13	F21	F24	F25
EUC		-	-	2	1*	-	-	3	-
ABS		-	-	-	1*	2	-	3	-
STH		-	-	-	1*	2	-	3	-
MAX		-	-	3	-	-	-	1	2

Table 4 (Continued)

Distances	Test Subject	Experimental Subjects							
	11	F4	F5	F10	11	F17	F19	F22	F24
EUC		-	-	1	3	-	2	-	-
ABS		-	1	2	-	-	3	-	-
STH		-	1	2	-	-	3	-	-
MAX		-	-	1	2	-	3	-	-
	14	F2	F3	F7	F10	F11	14	F19	F20
EUC		2	-	-	3	-	1*	-	-
ABS		2	-	-	3	-	1*	-	-
STH		2	-	-	3	-	1*	-	-
MAX		1	2	3	-	-	-	-	-
	16	F3	F6	F13	F15	16	F19	F25	F26
EUC		-	-	1	-	2	-	3	-
ABS		-	-	1	-	2	3	-	-
STH		-	-	1	-	2	3	-	-
MAX		-	-	1	-	2	-	3	-
	21	F3	F6	F8	F10	F17	F18	21	F24
EUC		-	-	-	3	-	2	1*	-
ABS		-	-	-	3	-	2	1*	-
STH		-	-	-	3	-	2	1*	-
MAX		3	-	-	2	-	-	1*	-

EUC = Euclidean Distance
ABS = Absolute Distance
STH = Steinhaus Distance
MAX = Maximum Distance

1* = Correct Identification

The final experiment to be reported in this section involves comparisons of the robustness of the same two vectors as identification cues (and generally the same population) for both the basic and forensic research models. Examination of Table 5 will reveal several relationships that should be of interest. First, as expected, LTS was the better predictor of speaker identity (than was SFF at that time anyway). Second, it can be seen how this approach permits various of the experimental distances to be evaluated for predictive strength. For example, the Steinhaus distance appears to be the overall most powerful (at least for these conditions). These data can be used for other evaluations also. However, the main contrast in this case is the one between the two research approaches (basic/forensic). While several comparisons can be made, there are one or two that are rather significant. That is, it can be seen that if the forensic model is utilized with the Steinhaus distance, 96% correct identification can be expected for LTS (this vector reflects general voice quality) and 72% for SFF (a quantitative measure of f0 which, in turn, reflects pitch level). Also demonstrated is the cited fact that, while the basic research approach is appropriate for the evaluation of vector strength, the forensic model permits a far superior prediction of speaker identity.

Basic Research Approach:

So many experiments have been carried out on the basic nature and development of our speaker identification vectors/approaches that a serial listing would be counterproductive and, perhaps, even confusing. Accordingly, we have elected to include only a number of sample experiments -- primarily to demonstrate 1) the investigational approaches utilized, 2) the depth and breadth of this project, 3) how the information derived from the basic research can lead to improved profiles/vectors and 4), of course, a summary of the key results.

LTS: A Sample Vector:

Since more research has been carried out on the long term spectral vector (a vector that is thought to reflect the general voice quality of a speaker), the data here can best be used to describe the cited elements of this research program.

As can be seen from Table 6, a vector can be used to study sample size; in this case, however, the vector was too robust (i.e., the scores were too high for proper evaluation); hence this research currently is being replicated on degraded speech and with other vectors. Table 7 provides data on two variables: 1) the use of different mathematical (distance) approaches in the decision process and (again) the effect of the presence of individuals who sound very much like the individual to be identified. Note that this experiment reflects combination of both basic and forensic protocols, and the control of speech contemporariness/subject/sample quality. In any case, certain relationships are observable (again the scores are a

Table 5. Summary table of means from several parallel experiments carried out on the same two vectors (LTS and SFF) twice: once utilizing the basic research approach (N=25) and the other, the forensic model (N=K + F1 to F7). Speech samples were of good quality but were text independent and non-contemporary. Values are percent correct identifications.

		Basic			
	Distance	1	2	3	Forensic
SFF					
1)	Euclidean	19	39	50	56
2)	Absolute	27	42	58	60
3)	Steinhaus	27	42	58	72
4)	Maximum	19	39	46	56
LTS					
1)	Euclidean	62	69	73	88
2)	Absolute	27	35	46	88
3)	Steinhaus	62	69	73	96
4)	Maximum	54	65	69	88

Basic results are derived from the three-nearest-neighbor approach.

Table 6. Percent correct classification for various length (10-40 sec.) contemporary speech samples. The LTS vector served as basis for the comparisons. All values are in percent; N=25.

A. Long Contemporary Samples

Test Sets	Reference Sets First	Second
First	-	100
Second	96	-

B. Short Contemporary Samples

Test Sets	Reference Sets First	Second	Third
First	-	92	96
Second	96	-	96
Third	92	100	-

little too high but are reasonably useful). First, the Steinhaus distance appears to be somewhat more useful as an indicator of speaker identity -- at least from the multidimensional set of parameters that make up this procedure. Second and as expected, randomly selected subjects were identified somewhat more accurately than those where doubles (sound-alikes) were introduced into the foil group. What was a little startling was the robustness of Steinhaus for this second procedure. If this relationship holds up, it very well may lead to improvement in the predictive ability of the entire system.

The data reported in Table 8 contrast different reference sets and different sample combinations as a function of vocal disguise. As can be seen, the LTS vector (in its present form) is not yet a good predictor of speaker identity when speech disguise is employed. Also apparent is the fact that different types of disguises will affect the identification process by different magnitudes. For example, the "pencil in mouth" and "pinched nose" procedure have the least effect on the data while register shifts (falsetto) appear to be the most detrimental to this vector as an identification cue. Finally (re: this section anyway), Table 9 provides data demonstrating that the approach being utilized in this case is completely text independent -- a result that was totally unexpected. Note also that the four distance measures of interest are again contrasted but in this case, the unexpected high identification scores may have obscured the actual differences among them.

Other Vectors:

As stated, work on the INT vector (INT is thought to reflect the level and variation of vocal intensity and/or perceived loudness) has been deferred until an improved set of parameters can be developed. Thus, the systematic and long-term research that has been carried out was concentrated on SFF or speaking fundamental frequency (pitch of voice), TED or Time vs Energy Distributions (plus speaking rate) and VFT or vowel formant tracking; a JIT or vocal jitter vector has been introduced but the research in this area is just being initiated.

In some cases, it is useful to contrast a developing vector against a more established one. Such contrast can be found in Table 10; here it can be seen that the identification value of SFF can vary between 44% and 78% under various conditions while LTS will vary from 84 to 100% in the same environment. The problem with this set of experiments lay in combining the two vectors of interest (LTS/SFF). In this case, correct identification by LTS was so high that no data relevant to the effects of adding the SFF vector could be obtained. We are now replicating this segment of the research but with the speech samples sufficiently degraded that the cited effects may become apparent. SFF also has been tested for a group of nine disguises in much the same manner as LTS (see Table 11 and see again

Table 7. Comparison of four distance criteria -- Euclidean, Absolute, Steinhaus and Maximum -- based on 16 forensic type experiments (i.e., the unknown matched to the known plus 7 foils). The LTS vector was utilized, as were good samples of contemporary (same day) speech. Eight of the experiments utilized subjects randomly drawn from a homogeneous population of young adult males; only "sound-alikes" and talkers confused with the unknown were utilized in the other eight. All values are in percent.

Distance	Randomly Selected Subjects/Foils	Foils Confused with Unknown	Mean
Euclidean	62	50	(56)
Absolute	88	75	(82)
Steinhaus	88	100	(94)
Maximum	88	75	(82)
All Distances	(82)	(75)	

Table 8. Percent correct classifications for disguise (LTS is the vector of choice). Reference sets consisted of four normal readings, each matched individually to the Test set, and matched in combination; it consisted of approximately 40 seconds of the "Grandfather Passage," whereas the Test sets contained 20 sec.

	Reference Sets					
	1	2	3	4	Combined No.1	Combined No.2
Test Sets						
Normal	52	72	64	64	64	82
Pencil in Mouth	28	40	32	32	40	40
Muffled with Hand	16	24	24	36	28	36
Pinched Nose	44	36	40	40	40	48
Free Disguise	8	20	8	24	16	24
Falsetto	4	12	4	4	8	16

Table 9: Summary table of research focused on text dependency utilizing the LTS vector. All values are correct identifications (in percent) for 25 subjects recorded under good quality conditions; samples were contemporary. Four distances were tested as was text dependency (R-R) and sample size (R=40 sec.; T=20 sec.). The reference level data are obtained in the same manner as studies reported in the early literature.

Distance	Reference Level	R1-R2	R1-R3	R2-R3	T-R1	T-R2	T-R3	Overall
Euclidean	100	88	96	92	88	85	92	90
Absolute	100	92	96	92	85	92	100	93
Steinhaus	100	92	92	81	85	88	100	90
Maximum	81	69	88	62	65	68	85	73

Table 10. Summary table of reliability runs of the verification procedure. Subjects were 25 young adult males; speaking condition was normal.

Analaysis	Percent Correct Verification			
	SFF-1	SFF-2	LTS	LTS/SFF-2
Posterior probability	71	78	100	100
Jackknife test	53	52	96	95
Identification test				
First Run	60	44	96	84
Second Run	60	52	100	100
Third Run	60	68	100	100
Fourth Run	50	48	92	92
Fifth Run	53	52	92	100
Mean	57	53	96	95

Table 11. Summary table of the effects of speaker disguise on the identification ability of the SFF vector. Subjects were 20 young adult males; comparisons were made for nine types of disguise to a normal reference set. Only the SFF vector was utilized.

Condition	Disguise to Normal Comparison percent correct
1. Pencil in mouth	70
2. Whisper	20
3. Pinched nose	45
4. Slow rate	45
5. Hypernasal	25
6. Falsetto	5
7. Muffled (hand)	35
8. Hoarse	10
9. Free disguise	20
Normal (control)	58

Table 8 for the contrasts). Note, here, the high scores for "pencil in mouth" and the particularly degrading effect on identity resulting from the changes in laryngeal production (falsetto, hoarseness, and whisper).

The TED vector was found to vary in its predictive value in earlier experiments (see previously submitted progress reports) and did not perform as well as did the other vectors even after modification. Accordingly, it was modified yet again and a large experiment carried out that simultaneously evaluated: 1) its reliability (Euclidean; runs 1-4), 2) the effect of distance measure (the four listed in the first column) and 3) effects of sample size (N=5-25). Indeed, the data reported in this table demonstrate that high (correct) identification levels are possible with TED; hence, the newly structured vector in this area now appears reliable -- especially when used with the Steinhaus distance procedure. Moreover, application of the forensic model should lead to good predictions by TED. Most encouraging of all was the very high TED identification level that occurred when this vector was coupled to the Steinhaus distance (re: the N=10 identification experiment).

As stated, research on a possible JIT (jitter or vocal roughness) vector is just being initiated. Table 13 provides some insight into the approaches being taken in this regard. Two different (digital) extraction procedures were contrasted in the first experiment -- and as a function of speaker, vowel and f0. As can be seen, the two approaches are quite similar except in two instances and the findings confirm much of what is known about jitter and its relationship to the acoustic theory of speech production. The JIT vector currently is being evaluated as a speaker identity cue. As can be seen from observation of the data on Table 14, it demonstrates potential in this regard but does not seem to be totally independent of vowel.

A decision was made early in the grant period to evaluate the power of various vowel formant tracking vectors (VFT) on the basis of single vowels and syllables -- at least initially. The results in this regard have been strikingly successful and some of the results of this series of experiments can be found in the tables to follow. As may be seen in Table 15, VFT vector scores were surprisingly high even when identifications were based on only two samples of the vowel /i/. While there is some indication that sample size may be a factor, it appears that position may not be -- that is, if phonemic context is similar and speech is reasonably contemporary (see Table 16).

As stated, both vowels and nasals appear to provide reasonably good articulatory data for use in identifying speakers from their speech. The data found in Table 17 addresses this issue; moreover,

Table 12: Partial results from several experiments with the TED vector. The first set of columns include reliability (discrimination) data based on test/preference samples of 20 sec. (text independent) that are good quality and contemporary. The identification experiments (second set of columns) also utilize good quality speech. All values are in percent correct identification.

	Reliability Experiments (N = 25)				Identification Experiments		
Distance	Run-1	Run-2	Run-3	Run-4	N=25	N=10	N=5
Euclidean	84	100	84	84	44	60	40
Absolute	-	-	-	100	-	-	-
Steinhaus	-	-	90	100	40	90	-
Maximum	-	-	-	64	-	-	-

Table 13. Comparison of fundamental frequency and jitter for four isolated vowels. Data are calculated from FFI-8 output with both the SFF and JIT programs being used -- and (in the second case) by digitizing the signal at 20 kHz and determining pitch periods by axis crossing.

		FFI/SFF/JIT		Digitization/Axis-crossing	
Subject	Vowel	f0	Jitter	f0	Jitter
M1	/u/	135.0	0.46	137.6	0.44
M2	/i/	147.9	0.91	154.7	0.86
M3	/a/	107.9	*	108.0	0.54
M4	/ae/	143.0	*	142.9	0.61

* Reliable data could not be obtained.

Table 14. Percent correct classification for the jitter vector (JIT). Normal speech samples produced by 25 male talkers were used as experimental material.

Test	Reference	1	2	3
First /i/	First /a,ae,u/	41.7	58.3	66.7
First /a/	First /i,ae,u/	33.3	66.7	75.0
First /ae/	First /i,a,u/	25.0	33.3	66.7
First /u/	First /i,ae,a/	16.7	41.7	50.0
First /i/	Second /i/	33.3	58.3	66.7
First /i/	Second /i,a,ae,u/	25.0	41.7	58.3

Table 15. Percent correct classification for the vowel formant tracking (VFT) vector. Normal speech samples only were used; talkers were males.

Speakers	Test	Reference	1	2	3
N=5	Second /i/	Third /i/	40.0	80.0	80.0
N=10	Second /i/	Third /i/	30.0	50.0	60.0
N=25	Second /i/	Third /i/	42.3	50.0	50.0

Table 16. Percent correct classification for the vowel formant tracking vector (VFT). Speech samples consisted of the vowel /i/ isolated from connected speech but with all /i/'s in the same phonemic context. Males (N=25) were used as subjects.

Test	Reference	% Correct 1	2	3
Second /i/	Third /i/	42	50	50
Second /i/	Third /i/, Fourth /i/	35	50	62
Second /i/	Third /i/, Fourth /i/, Fifth /i/	31	46	58
Second /i/	Third /i/, Fourth /i/, Fifth /i/, First /i/	35	39	62

the problem of contemporary/non-contemporary samples is studied simultaneously. Evaluation of the relationships in Table 17 will reveal that contemporariness may be a factor in the identification process. These data also demonstrate that the syllable /na/ (a low vowel coupled to a nasal consonant) may not be quite as good a predictive cue as was expected -- and certainly not as robust as the high front vowel /i/. This conclusion is pretty much confirmed by the results of the 20 experiments reported in Table 18. Here the variables include: 1) the speech sounds /i/ and /na/, 2) four distances and 3) randomly selected foils vs those foils that sound quite similar to the target speaker. These experiments are considered basic in nature even though the forensic model was utilized. In any case, of the four distances, Steinhaus was found to be the most sensitive predictor of identity and (as stated) the high front vowel /i/ a better predictor than /na/. It should be noted also that, when individuals who sound like the talker are present, the identification process is somewhat degraded. To summarize, the VFT vector appears to be a very powerful one; research is underway contrasting our software with the LPC approach provided by ILS. Also underway are studies contrasting five different vowels (see again Table 16).

Research on Vector Combinations -- The SAUSI Profile:

As has been indicated in the quarterly progress reports submitted earlier, a number of attempts have been made to further improve the mathematical decision criteria that support our profile or multiple vector array. The success of these efforts may be best assessed by re-examination of Tables 1-5. Moreover, details of our first attempts to obtain specific detail about relevant relationships associated with our approach may be found in Table 19 (another large experiment of this type -- but with a new "tree" statistical approach currently is underway) -- it is awaiting funding. In this case (i.e., F-19), the power of four and three vector combinations are contrasted (an earlier experiment included the INT vector; those results can be found in a previous progress report). As can be seen by examining the table, LTS is the most powerful vector and in many cases (except for SFF anyway) the addition of data from other vectors does little to improve its predictive ability (even the addition of data from all three other vectors adds only 3%). Even though the same problem does not seem to exist (or is not as debilitating) when the forensic model is applied, we have, nevertheless, revised these procedures and are replicating this experiment.

In summary, we believe that the stated progress made on this grant has been reasonably well reviewed -- even though a number of other single and multiple factor experiments have not been discussed (many were reported in the quarterly reports). Nevertheless, it would appear useful to conclude this report by listing the results of two other related experiments; these data may be found in Tables 20 and 21. In this case, the identification strength of LTS (Table 20) and

Table 17. Percent correct classifications for the VFT vector, utilizing the syllable /na/. The Reference sets consisted of four normal readings, each matched individually to the Test set, and matched in combination. The three nearest neighbors (first, second and third choices) are displayed.

	Test Set			
Reference Set	First Choice	Second Choice	Third Choice	Total
1	16	8	0	24
2	40	20	8	68
3	36	8	8	52
4	12	4	20	36
combined	40	4	16	60

Table 18: Summary table of 20 experiments providing data about the robustness of single vowels/syllables as identification cues. The forensic model is utilized; four "distances" are included. Stimuli include the vowel /i/ and the syllable /na/; subjects for half the experiments were randomly drawn from a homogeneous population, with the other half from a group of "sound-alikes". All values are in percent of correct identifications for five studies.

Distance	Vowel /i/		Syllable /na/		Means
	Random	Sound-alike	Random	Sound-alike	
Euclidean	40	40	20	20	30
Absolute	80	20	80	20	50
Steinhaus	100	40	60	40	60
Maximum	80	40	40	0	40
Means	75	35	50	20	

Table 19: Summary table of two primary verification procedures. The nearest neighbor (serial and weighted) approach was utilized. Subjects were 25 adult males.

VECTOR	EXPERIMENT	
	THREE Vectors	FOUR Vectors
SFF	44	50
LTS	68	74
TED	36	35
VFT*	--	50
SFF/LTS	72	73
SFF/TED	56	54
SFF/VFT	--	54
LTS/TED	56	46
LTS/VFT	--	69
TED/VFT	--	42
SFF/LTS/TED	68	69
SFF/LTS/VFT	--	69
SFF/TED/VFT	--	70
LTS/TED/VFT	--	62
SFF/LTS/TED/VFT	--	77

* A somewhat limited test as the VFT procedure did not provide data for all subject/condition combinations.

SFF (Table 21) were evaluated as a function of channel distortion (noise, passband), disguise (four types) and vector construction (six methods of combining parameters; however, only No. 4 is reported)* -- all on the basis of the three nearest neighbor procedure. As may be seen, the four types of disguise researched degraded LTS but little and except for "muffled with hand," SFF only modestly -- even when these disguises were combined with noise. On the other hand, note the severe degradation that occurs when disguise is combined with telephone transmission (actual not simulated) and that LTS is degraded more seriously (almost to chance levels in most cases) than is SFF. Data such as these are included in order to demonstrate how we are systematically investigating many of the thousands of basic and applied issues associated with speaker recognition. The basic research is necessary as it leads to improved knowledge about vocal function, acoustic phonetics and speaker recognition in general and specific data about the speaker identification process and the strength of our vectors in particular. Even more importantly, however, it leads to possible criminal justice and law enforcement applications of an effective speaker identification profile. The success (see the previous section on the forensic approach) of our partially developed method demonstrates our contentions in this regard.

CONCLUSIONS

A number of conclusions are possible.

1) Our postuation that the natural speech feature approach to speaker identification is superior to other signal processing techniques appears to have been strongly supported by the research carried out under 84-IJ-CX-0014.

2) The vectors selected (as modified) appear suitable for the speaker identification task.

3) The vector combination (or profile) approach to the identification task appears warrented as specific vectors appear to be more robust in certain environments, yet others more viable in yet other situations.

* Three of the approaches were not acceptable; of the other three, individual scores varied +/- 10%. However, since the patterns observable for the No. 4 approach are typical, only these two tables (T-20; T-21) are included in this report.

Table 20: Percent correct classification when disguised speech samples of four types were matched to normal, noise and telephone bandpass samples; LTS was the vector of interest. A test set was compared to a reference set combination of four samples with pooled variance. Males (N=25) were used as subjects. The data reported here are for one evaluation out of six.

Test	Reference	1	2	3
Pencil in mouth	Normal	69.2	76.9	80.8
Pinched nose	Normal	65.4	76.9	84.6
Slow rate	Normal	46.2	53.8	57.7
Muffled with hand	Normal	46.2	50.0	53.8
Pencil in mouth	Noise	65.4	76.9	80.8
Pinched nose	Noise	65.4	88.5	96.2
Slow rate	Noise	53.8	61.5	69.2
Muffled with hand	Noise	30.8	50.0	65.4
Pencil in mouth	Telephone Bandpass	7.7	11.5	19.2
Pinched nose	Telephone Bandpass	3.8	7.7	11.5
Slow rate	Telephone Bandpass	3.8	7.7	7.7
Muffled with hand	Telephone Bandpass	3.8	15.4	19.2

Table 21: Percent correct calssifications for SFF when four different disguise conditions were matched to normal, noise and telephone bandpass samples. A test set was compared to four reference sets with pooled variance. Males (N=25) were used as subjects. The data reported here and for one evaluation (No. 4) out of six.

Test	Reference	1	2	3
Pencil in mouth	Normal	46.2	61.5	61.5
Pinched nose	Normal	34.6	42.3	65.4
Slow rate	Normal	34.6	42.3	50.0
Muffled with hand	Normal	19.2	42.3	53.8
Pencil in mouth	Noise	30.8	53.8	61.5
Pinched nose	Noise	23.1	53.8	69.2
Slow rate	Noise	30.8	46.2	50.0
Muffled with hand	Noise	15.4	26.9	50.0
Pencil in mouth	Telephone Bandpass	15.4	23.1	34.6
Pinched nose	Telephone Bandpass	15.4	15.4	30.8
Slow rate	Telephone Bandpass	7.7	19.2	30.8
Muffled with hand	Telephone Bandpass	11.5	19.2	26.9

4) Our postulate that a two level research approach to the identification task -- i.e., use of both the basic research and forensic modesl -- has been strongly supported by this project. The extremely high levels of correct identification for actual/simulated field evaluations (i.e.., the forensic model) demonstrate this contention.

5) Even though the basic research on the process is not complete, our approach is proving to be field effective in the forensic milieu.

6) The SAUSI (Semiautomatic Speaker Identification) procedure appears of demonstratable merit. However, additional research will be necessary to refine it even though it already appears nearly field-ready. It should be possible to complete the required research in about two years.

A SYSTEM FOR SPEAKER IDENTIFICATION

Investigators

Harry Hollien, Ph.D.
Mark Yang, Ph.D.
Donald G. Childers, Ph.D.
Ruth Huntley, M.A.

FINAL REPORT

Grant: 84-IJ-CX-0014

Report Prepared By:

Harry Hollien, Ph.D.

Institute for Advanced Study of the Communication Processes
University of Florida
Gainesville, Florida 32611

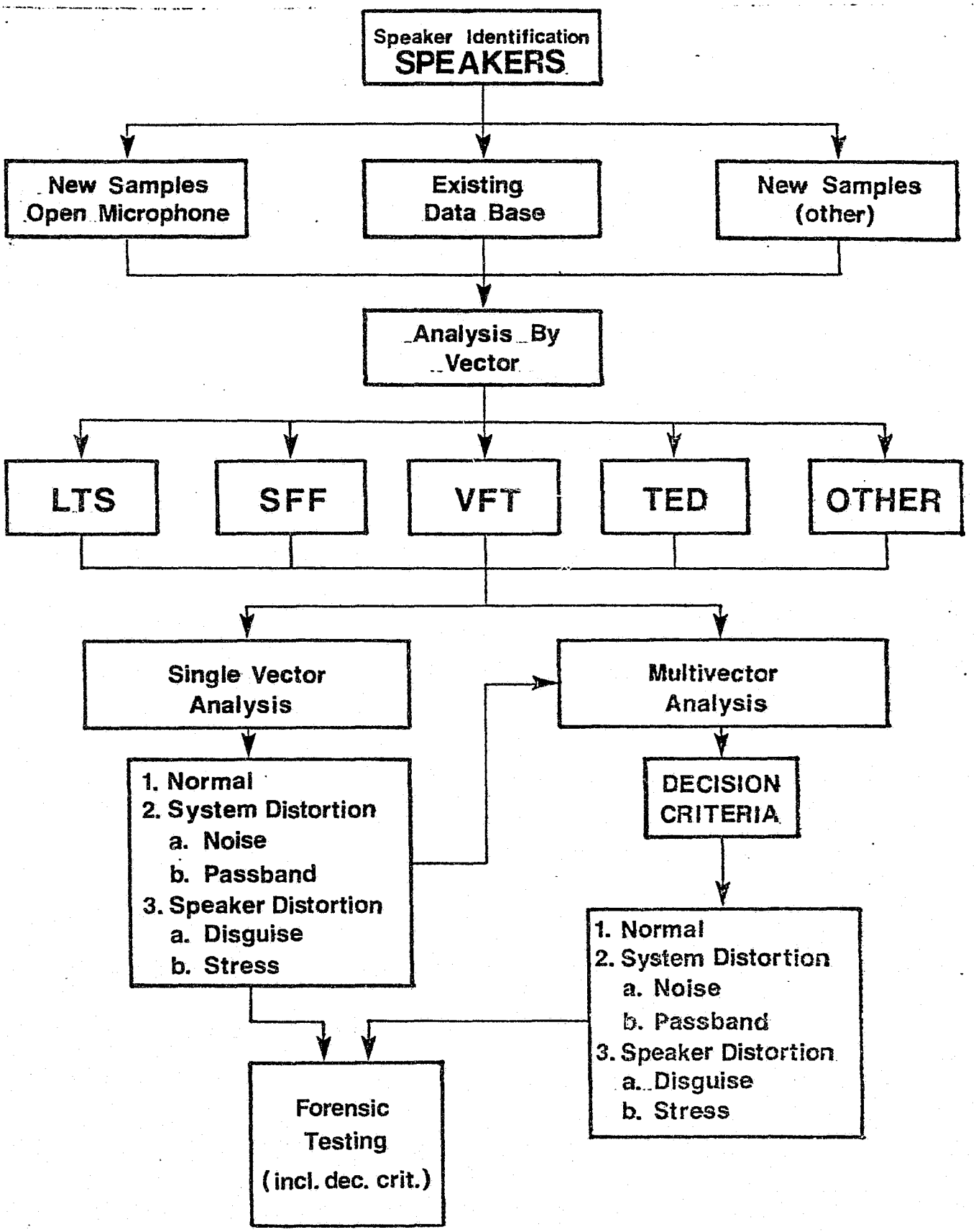
December 31, 1985

ABSTRACT

The basis of this research is the theoretical construct that the speech signal contains features which are sufficiently unique and consistent within a given individual -- and reasonably different among individuals -- to permit successful speaker recognition. That is, both data and logic permit the assumption that certain elements within a talker's speech are relatively idiosyncratic and discriminative as a result of his or her basic glottal and supraglottal structure and the habituation of patterns used in speaking. Social, economic, geographic and educational factors as well as maturation level, psychological/physical states, sex and intelligence all affect speech patterns in specific ways and can combine with those idiosyncratic attributes of the talker's anatomy and physiology to create recognizable features in that person's speech and voice -- ones that can be used in the recognition process. Second, it also is our position that, while there may be no single attribute within a person's speech of sufficient magnitude to permit that individual to be differentiated from all other talkers in all situations, the use of groups of features will permit the recognition process to occur. Third, we postulate that the simple analysis of captured signals (i.e., signal analysis by conventional means) will not permit successful speaker recognition -- especially if channel or speaker distortions are present. Rather, to be successful, a speaker identification system must focus on the natural features within a talker's speech -- that is, if the resulting procedures are to prove robust for criminal justice and law enforcement purposes. Finally, if a speaker identification system is to be employed in the field it must be based on results from systematic research at two levels -- i.e., those that employ 1) the basic research model (investigation of the basic relationships among the parameters and vectors applied to the task) and 2) the forensic or field model (evaluation of the procedures/techniques developed for controlled experiments of this type parallel to, or occuring in, the forensic milieu). Our generalized approach to the problem may be best understood by consideration of Figure 1. As can be seen, human speech samples (drawn as needed from a variety of controlled sources) are analyzed by a variety of methods (i.e., vectors or multidimensional sets of parameters) either singly or in combination, and under a variety of distorting conditions. These procedures are followed for either the basic research or forensic models. Finally, the ultimate goal of this project has been, and is, the development of a valid/reliable (semiautomatic or computer assisted) speaker identification system appropriate for use by criminal justice and law enforcement personnel.

As stated, the approach utilized for the cited purposes incorporated both basic and applied experiments. Subject populations included large (N=25) homogeneous groups of males (i.e., similar dialects, socio-economic backgrounds, education, size, and so on) as well as smaller sets of subjects (known and unknown talkers plus 6-10 foils) both randomly drawn from a large male population and chosen for confusability from prior experiments (i.e., multiple "sound alikes"). As may be seen from the report to follow, yet other groups of male speakers were chosen where warranted and some preliminary research was carried out on female talkers also. Second, the vectors utilized were based on recognizable features in the human voice (also

FIGURE 1. Flow chart of the research approach utilized to develop a speaker identification system for use in the criminal justice and law enforcement milieu. The procedures employed permit appropriate research to be carried out as a function of either the basic research or forensic models.



upon deductive logic and research data); they consisted of 25-40 parameter multi-dimentional -- but related -- sets of speech characteristics. These experimental vectors included 1) general voice quality (long-term spectra or LTS), 2) speaking fundamental frequency level/variability (SFF), 3) vocal intensity level/variability (INT; deferred for improved vector construction), 4) articulation -- both vowels and nasal consonants (vowel formant tracking or VFT), 5) prosodic or time features of speech (rate measures plus time-energy distribution or TED) and 6) vocal roughness (jitter or JIT -- not yet satisfactorily developed). Of the six vectors, the four that have been proven most robust for the speaker identification task are LTS (voice quality), SFF (speaking fundamental frequency), VFT (vowel/syllable features) and TED (speech time/rate). Third, a large number of basic experiments have been carried out as have a number investigations related to the forensic model (either real or simulated "cases") -- plus some highly controlled experiments where field conditions have been rotated/evaluated. Finally, a substantial number of distance measures (simple and multidimensional space), statistical techniques and decision criteria have been studied.

The report to follow provides information about some of the more relevant experiments carried out under the aegis of this grant. While not highly technical, it is somewhat long and detailed. Accordingly, the results of this project -- plus some conclusions and recommendations -- will be briefly summarized in this abstract (please note, however, that necessary clarifications can be found in the body of the report). First, the research to be reported is divided into two sections -- experiments based on the forensic model and, later, studies representative of those carried out in the basic research area. The initial summary is of seven actual civil or criminal cases; all were of field quality and, in several, there was a serious degrading of signal integrity/quality. Positive identifications were made in five cases, eliminations in two. Data from extensive aural-perceptual experiments (large populations; 2-3 groups) proved to be in close agreement with the computer-based machine data in <u>all</u> instances -- and subsequent case disposition appeared to validate the decisions made on the basis of this information. Second, four simulated field experiments were carried out under conditions of severe speech degradation. Correct identifications were realized in all four instances -- and at rather high probability levels. Third, nine seperate experiments were carried out (again under simulated field conditions) and correct identifications were made in all instances. Also reported are 10 experiments where all or nearly all of the foils were individuals who sounded very much like the known talker or had been confused with him in many of the basic experiments. In this case, a second set of variables also was evaluated; i.e., four distance measures. The Steinhaus and absolute distance procedures were found superior to the other two; indeed, 60% correct identification was realized even under these negative conditions. Thus, it can be concluded that the semiautomatic speaker identification procedures resulting from this project display exceptionally good performance in the field even though the basic research experiments (and procedural refinements) are <u>not yet</u> complete<u>!</u>

The results of the several hundred basic experiments funded by the 84-IJ-CX-0014 grant are much more difficult to organize for brief

summary. In this case, protocols were employed (initially anyway) which degraded the entire procedure in such a manner that large response ranges were possible -- thereby permitting the subtle differences among the parameters and the vector strengths and weakness to be efficiently evaluated. Identification of those vectors of demonstrable robustness, modifications for vector improvement and determination of optimum use under specific field conditions also were determined by this approach. Lastly, vector strength under moderate experimental conditions (non-contemporary speech, text independent samples, homogeneous groups of subjects and so on) was carried out.

Briefly, the results demonstrated that the LTS vector is nearly 100% accurate for many conditions and SFF is nearing these levels. Secondly, even though all of the basic research on VFT was carried out on single vowels and nasal syllables, we found this vector to be (potentially) a very powerful one. Indeed, scores here are already so high that it appears that it soon will rank with SFF and perhaps even LTS. On the other hand, while the upgrading of TED has progressed at a somewhat slower rate, the more recent modifications are encouraging as identification scores based on this vector have improved markedly. Moreover, TED is the vector that shows the greatest resistance to bandpass effects and, especially, speaker disguise. To be explicit, the basic experimentational data may be summarized as follows. First, LTS demonstrates 100% correct speaker identification under many conditions; it is resistent to noise, stress and related degradations; the LTS weakness includes decreased effectiveness for telephone bandpass and speaker disguise. Efforts now will be made to improve this vector's performance in these two areas without degrading its strength in the others. Second, SFF is now demonstrating robustness in many areas; however, it is most vulnerable to register/voice changes in disguise -- especially in different psychological environments -- and limited bandpass occasionally results in somewhat unstable SFF data. As with LTS, research focused on these deficiencies is underway (but awaiting funding). Third, the VFT vector appears to be a potentially powerful one but occasionally is difficult to extract from noisy signals. Future thrusts here will be designed to evaluate its strength when multiple stimuli are employed and to see if the vowel formant ratios will permit the high identification levels expected. Also to be researched are several digitization techniques and especially which of them will overcome the occasional instability found in VFT. Fourth, it is apparent that the TED vector should be expanded to include more prosodic parameters. We are especially desirous of improving the general performance of this vector (especially for speaker stress) without degrading its robustness as an identification cue in the areas of disguise and telephone bandpass. Finally, the issue of combining the vectors is a critical one. As it turns out, only modest improvement in identification currently is realized when vectors are combined. However, we are beginning to discover that this situation arises because the strength of a vector varies in different environments and one or two of them prove dominant in most configurations. Thus, we now believe that it is important to conceptualize the SAUSI (semiautomatic speaker identification) profile in <u>two</u> ways. First, it will be necessary to develop profile configurations that fit each of a number of rather specific forensic situations and second, it also will be important to improve the mathematical process of combining the vectors within multi-

multidimensiona space.

A final comment about the basic research thrust carried out under the aegis of this grant seems warranted. As may be seen, a substantial number of studies focused on the four distance measures and on several statistical procedures which have been carried out. This research also should be continued as several new distances must be evaluated and a number of the novel statistical procedures (developed by Dr. Yang) should be evaluated. In any case, currently it appears that the Steinhaus distance measures are superior to the three others. Nevertheless, it is necessary to determine which of the distance measures works best and under which conditions. Once available, these data/procedures can be combined with the vectors in the SAUSI profile and applied differentially to the cited forensic situations. In any case, data outlining many of these relationships can be found detailed in the report to follow.

A few of the more relevant conclusions should be included in this abstract. First, it is now obvious that the natural speech feature approach is one of merit -- and perhaps even the only viable approach currently available to speaker identification (and speaker verification) in the field. Second, the systematic research approaches applied here appear necessary if a good speaker recognition method is to accrue; moreover, the two level (basic/forensic) approach has been found to be quite effective. Third, it now is quite apparent that a profile approach will be necessary if speaker identification is to be carried out successfully in the field. Fourth, the vectors we have chosen appear adequate to the task (they seem especially effective in field situations). Fifth and perhaps most important, the statements made above can be confirmed in a great measure by the exceptionally good (and somewhat unexpected) performance of SAUSI in the field. Finally, even though our method already appears adequate for field use, it has not yet been refined -- and hence additional research is necessary if premature application is to be avoided.

INTRODUCTION

There are three independent yet related areas within the Communication Sciences that are of substantial importance to criminal justice and law enforcement; they are speech recognition, speaker verification and speaker identification. Even though rather substantial progress has been made with respect to the development of on-line methods related to the first of these problems -- and modest progress re: the second -- the fact remains that there are no independent systems currently in existence that permit speech/speaker recognition tasks to be carried out. The problem is complicated by the fact that nearly all of the research being carried out in these three areas is concentrated in the first two. The reasons for this situation are clear. First, the task is a formidable one. Speaker identification -- unlike speaker verification -- always involves an "open" set of suspects (i.e., the criminal may or may not be among subjects in the set), yet one of the subject/suspects is likely to be selected as the individual most similar to the unknown anyway. Second, the signal usually is degraded by system or channel distortions such as noise; limited bandpass and so on and/or by speaker distortions such as disguise, stress and so on. Moreover, the forensic model is one where the process may not be text-independent; it certainly involves non-contemporary matches. Third, there are social implications that sometimes tend to discourage relevant agencies from supporting research in this area.

It is the fourth problem associated with speaker identification that is of greatest importance. Specifically, human speech has been thought to be so variable that there may be no characteristics within the (resultant) acoustic wave which would permit reasonable levels of identification. Yet, it can be observed that, from time-to-time, every normally hearing individual is able to recognize known talkers from the perception of their speech alone; thus, the logic that speaker identification is possible by signal processing would appear irrefutable. Moreover, the results of some completed experiments would suggest the argument that analysis of combinations or groups of speech features will operate powerfully (where single parameter analysis would fail) and permit speaker identification to be carried out on a scale far greater than previously considered possible. Questions remain, however, concerning the potential universality of speaker recognition, the identity of those features that can be most successfully used as cues, the best parameter combinations for each environment and the practical application of those procedures or systems which result from research in this area. In short, two basic questions must be addressed: 1) Is interspeaker variability always greater than intraspeaker variability (i.e., are there certain speech features so idiosyncratic to an individual that identification always is possible) and 2) can these features -- or a profile based on them -- actually be applied to the forensic model. In response, it appeared that both basic and applied research should be pursued in this area -- and such was our approach re: the 84-IJ-CX-0014 project.

Progress has been substantial and, while we concede that there are still some basic (and applied) questions to be answered, it now should be possible to use our procedures in the field -- at least in

a limited way. Moreover, if our procedures are properly applied, there should be a reasonable probability of useful results. We base our position in this regard, in part on the available data and in part on our successful use of the forensic model. That is, while we concede that there may not be a single speech vector which will permit very high levels of correct identification, there appear to be a group of vectors (i.e., a profile) which will permit successful application of the process. Please note, however, that we also concede that additional research will have to be carried out to further validate our procedures, refine them and develop an easily used set of field techniques.

AN APPROACH TO SPEAKER IDENTIFICATION

To date, three general approaches to speaker identification have been utilized; they are: 1) aural/perceptual recognition, 2) spectrogram matching and 3) machine recognition. Perhaps most is known about the aural/perceptual approach. While we concede that definitive data have not yet been reported -- even for small scale sorting and matching experiments of restricted voice samples, we have found it possible to base our efforts on a tentative aural/perceptual model. Indeed, we have found that, in aggregate, most of the strategies used by humans in the identification task have been defined and, thus, coupled to our models, can provide the substructure to our approach. That is, we were able to determine that the following parameters are used by listeners in the speaker identification process: 1) speaking fundamental frequency (f0) level and variability (vocal pitch), 2) speech spectra (voice quality), 3) nasal resonance (articulation), 4) vowel placement (articulation), 5) vocal tract turbulence (quality of speech) and 6) prosody (speech timing/rate). As will be seen (below), the research to follow is organized in response to these observations. That is, we argue that it is only the natural features of speech that will resist channel and speaker distortions. Our results appear to demonstrate the strength of this contention.

It should not be necessary to comment on the second approach to speaker identification -- i.e., the spectrum matching or "voiceprint/gram" approach. As a method, it has been described/defended by a very few authors and studied and/or attacked by numerous scientists. Moreover, when the data from all reported experiments are combined and synthesized, the inescapable conclusion is that the "voiceprint/gram" method of speaker identification is not a valid one. However, data accumulated about vowel/consonant formants suggest that they contain useful identification cues and this relationship is of some importance in our model.

We selected the third approach to speaker identification as the only practical one possible -- and for several reasons. First, we noted that nearly all research being carried out in the area of speaker verification is machine oriented. Second, we believe that there is no practical way that human auditors can be utilized objectively to carry out speaker identification tasks and that modern technology has eliminated the need to do so. In any case, we would argue that appropriate quantitative (machine) processing of human

speech (in order to identify those natural speech features idiosyncratic to the individual) will lead to the most robust method of speaker identification. Indeed, the research to be reported will serve to verify this contention.

To summarize; we now can assume that speech is so unique to the individual that a selected cluster of features can be used as stable, valid identity cues and that natural speech features are among the most powerful signal-contained elements for this purpose. In response to these assumptions, we developed and tested a number of vectors for this purpose.

OBJECTIVES

The basic purposes of this project were to obtain data on the process of human speaker identification and to test the usefulness of a speaker identification system. Also considered important were efforts to test several models of speaker identification, distance evaluation schemes -- and, especially, our postulate that a natural speech feature approach is potentially robust enough to be effective in the field.

To be specific, our objectives were to develop a systematic and sensitive approach to speaker identification by:

1) Identifying, structuring and analyzing those speech features (vectors) that could serve as predictors of a speaker's identity,
2) Evaluating the profile or multiple vector approach to the identification task and determine which configurations were resistant to distortions of various classes and types,
3) Adapting and analyzing various data storage/retrieval and statistical procedures for specific use in decision criteria,
4) Structuring and testing the forensic model as an assessment of the robustness of our techniques for field use.

In short, we believe that our systematic approach to the problem, plus the data generated, have resulted in development of a number of useful constructs and provided sharply upgraded knowledge about the problem. Further, this effort has permitted testing of several speaker identification models and the development of a field system that promises (after additional refinement) to be a reasonably efficient one.

METHOD

As stated, the basis of this project has been the theoretical construct that the speech signal contains features which are sufficiently unique to a given individual -- and enough different among individuals -- so as to permit effective speaker identification to be realized. That is, both data and logic permit the assumption that certain elements within a talker's speech are relatively idiosyncratic -- that they result from habituation of speaking patterns which, in turn, are based on social, econcomic, geographic

and educational factors as well as maturation level, psychological/physical states, sex and intelligence. In any case, we maintain that these factors combine with behaviors related to the talker's anatomy and physiology so as to create recognizable features within a particular person's speech and voice. Further, this position is based on two postulates that: 1) natural speech characteristics provide the most robust identification cues and 2) while no single speech attribute may be powerful enough to permit differentiation from all other talkers, the use of feature groups will permit the recognition process to occur -- even under unfavorable conditions. A brief review of the general methods employed will proceed specification of the results.

The Data-Base

One of the features of this project was that we already had assembled a large portion of the required data-base. In all, recordings were available of 250 men and 136 women (N=386) who produced speech samples of several types (N=4669). This data-base was established so that the effects of speaker and system/channel distortions upon our vectors could be evaluated and identification cues studied in a variety of environments. The primary system distortions included: A) limited passband (including a "telephone data-base") and B) noise, with talker distortions including: 1) several types of stress, 2) disguise (nine types), 3) dialect (two dialects), 4) sex, 5) age and 6) speech materials (four types). Field conditions were simulated by combining system and speaker distortions or by using tape recordings from actual cases. Finally, these 4,669 samples were supplemented as needed (see protocols listed below).

The Experimental Vectors

Even though it was obvious that analysis of certain speech features would lead to the successful identification of talkers, the specific features that would prove most useful had not been identified. Hence, we evaluated available information on fundamental frequency, glottal volume velocity, vowel formant frequencies/bandwidths, turbulent phonemes, nasal consonants, prosodic/timing features and the similar elements; those vectors to follow were selected for intensive evaluation.

Long-Term Speech Spectra (LTS). The use of power spectra as an identification cue has had a relatively long history in this area. We postulated LTS to be an index of general voice quality and that, as such, it is a good cue to a speaker's identity. As will be seen, we have discovered that LTS can correctly predict the identity of speakers at very high levels; at least when data are normalized and for laboratory type research. We also have been able to demonstrate that LTS is relatively resistant to the effects of speaker stress and a varied of noise conditions. However, it does not function well as a predictor of identity when talkers disguise their voices. The LTS data extraction system includes a Princeton 4512, FFT spectrum analyzer coupled to our PDP-11/23 computer; we hope to use ILS (also) in the future. The vector utilizes 40 parameters to generate a power spectrum curve covering a frequency range of 60-10,240 Hz. A number of mathematical "distances" provide the desired statistical

comparisons.

Speaking Fundamental Frequency (SFF). While perception of f0 has been shown to be a reasonable cue for speaker recognition, feature analysis has been only mildly encouraging as researchers have reported varying degrees of success with it. Accordingly we developed a 30 parameter SFF vector and results have been substantially good. The parameters making up this vector include SFF mean, SFF standard deviation, the number of semitone (ST) intervals containing energy plus the number of waves in each of the ST interval "bins." Fundamental frequency data are obtained automatically by means of FFI-8 -- the output of which was fed directly to the PDP-11/23 computer. FFI-8, a digital readout f0 tracking device, consists of a series of successive low-pass filters, with cutoffs at half-octave intervals, coupled with high-speed switching circuits which are controlled by a logic system. FFI measures each wave (it does not "sample") by producing a string of pulses -- each pulse marking a boundary of a fundamental period from complex speech waves -- which are delivered to the computer. An electronic clock marks the time from pulse-to-pulse and these values are processed digitally to yield (among other data) the geometric mean frequency level and standard deviation of the frequency distribution. In order to provide for the new multiple-parameter SFF vector, a subroutine was written to permit the semitone interval analysis; further semitone histograms and frequency or period information were used to compute the actual features such as mean value, modal value, modal frequency, variance and histogram entropy.

Vowel Formant Tracking (VFT). Much use has been made of vowel formant center frequencies, bandwidths and transitions by individuals using time-frequency-amplitude spectrographic techniques in speaker identification. Admittedly, that approach is not a very sophisticated one; however, the research conducted by these individuals has suggested that vowel formants and their nature definitely are important to the speaker identification task. We have upgraded the procedure and are using a Princeton 4512 FFT coupled to our PDP-11/23 computer for this purpose (again we wish to use ILS in the future). In any case, research in both the aural/perceptual area and relative to study of the issue via machine approaches can be used to argue the importance of elements within the formants as speaker identity cues. When selecting the vowel characteristics for this vector, we focused our efforts on several features: 1) the center frequencies of the first three formants (F1, F2, F3) and 2) ratios among these three formants (F1/F2 and F2/F3). As will be seen, it appears that F1, F2 and (perhaps) F3 center frequencies are robust indicators of speaker identity -- (especially when several vowels such as /i,a,u/ plus the syllable /na/ are used as stimuli). Initially, we used spectral analysis protocols (Princeton 4512) with the frequency/intensity of those three energy peaks corresponding to the first three vowel formants used as recognition cues.

As may be seen from the report to follow we have not yet been able to assess the robustness of vowel formant ratios (in process). Nevertheless, we would suggest that they are important for the following reasons. Formant frequencies are generally dependent on

the size and shape of the vocal tract. Therefore, they are based both on anatomy and vocal tract articulatory movements. However, since the range of formant frequency shifts and the F1/F2 and F2/F3 ratios probably cannot be significantly altered at will, they should convey idiosyncratic information about a specific talker. Further, the relative "position" of a speaker's vowels (as determined by both the formant frequencies and ratios) also should be idiosyncratic of an individual's overall speech pattern. Finally, it would appear that formant frequency analysis/ratios appear resistant to many kinds of system distortions and, therefore, should be useful for speaker identification purposes even though those types of interference are present. As stated, however, we have only been able to extract the necessary ratios (by the end of the current grant) and have not analyzed them. Hence, they cannot be used to contribute materially to this report.

Temporal Vector (TED). Very little research on speaker identification has focused on any of the temporal parameters that can be found within the speech wave; and there are but few exceptions here. Nevertheless, there is strong logic that there are prosodic speech elements that can be extracted and used for recognition purposes. For example, given the hypothesis that talkers differ with respect to the durational use of acoustic energy in speech, it is possible that the amount of time a speaker is or is not producing an acoustic signal during a specific amount of connected discourse may be useful in the identification task. Moreover, certain individuals appear to employ a greater number and/or longer silent intervals in producing a linguistic message than do others. In any case, a rather substantial number of temporal speech features appeared amenable to study; we selected the following: a) Total Speech Time (TST) -- defined as the period (in ms) it takes to produce an utterance of a set number of syllables, b) Speaking Time Ratio (S/T) -- defined as a measure of the total time for which acoustic energy is present during a set utterance, c) Silent Interval (SI) -- a reciprocal of speaking time (ST), d) Speech Rate (SR) -- a measure of the speech material completed during a fixed time period (based on syllable rate not word rate) and e) Consonant/Vowel Duration ratios (C/V) -- the (time) ratio between a particular consonant and a vowel in a specified CV utterance. An additional temporal vector also was evaluated; specifically, a time-energy distribution (TED) vector which reflects the total time a talker's speech bursts remain at specific energy levels (relative to his peak amplitude). This also provided an indication of the speaker's speech pattern with respect to speech bursts and pause periods. The TED vector is totally software based; all digitization is obtained utilizing our PDP-11/23 computer with an A/D converter.

Other Vectors: The above cited four vectors provided basis for the primary thrust of this research. However, a number of other vectors also were researched but only on a preliminary basis; they included: 1) vocal intensity (deferred), 2) jitter, 3) shimmer and 4) phonemic unit. However, research here has been preliminary in nature and, to date, no robust vector has been identified from among this group and only a small amount of data will be reported. Actually, a great deal of research was carried out on the vocal intensity vector

(VI). However, it did not prove to be very robust and, hence, research here has been deferred until a more promising VI approach is identified. On the other hand, the jitter vector appears promising -- at least as a subvector to SFF. In any case, some data for each are reported in the results section to follow. In summary, the natural speech vector approach to speaker identification has proved to be effective -- and at least four vectors have been identified as useful.

Experimental Approaches Utilized:

The experimental approach we have employed is considered somewhat unique;. hence it should be emphasized. Fundamentally, research was carried out on two levels; these levels were based on: 1) a basic research model and 2) a forensic model.

The Basic Research Model: The goals of this approach were to cycle the selected vectors through a variety of structured laboratory evaluations in order to study the effects of a large variety of system (channel) and speaker distortions -- and their effect on both the vector and the identification process. Since it was considered desirable to observe differences that could be rather subtle, it was necessary to spread the identification responses over as large a continuum as possible. To be specific, a correct identification range of from 20% to 80% will permit a 60 percentile point dynamic range. Of course, it is conceded that as the vector under study becomes better understood, it also is desirable to assess its maximum potential for correct identification. Nevertheless, the best evaluation of the strengths and weaknesses of a vector can be determined by the process of degrading it in a variety of ways and studying it under these adverse conditions. The maximum identification levels are studied at a later stage (in this case only for LTS so far -- and to some extent SFF).

As would be expected, if an extensive spread of responses is required, the research must be designed appropriately. In this case, a group of 25-30 healthy young males was chosen as the primary experimental population. These males were by-and-large quite similar in education, dialect, size, age, health and so on. Hence, a large, quite homogeneous population was identified and employed in much of the basic research conducted. It should be noted especially that at least five pairs of subjects were relatively close "sound-alikes". A population such as this one will tend to make attempts at speaker identification quite difficult -- especially if channel effects, text, environment and talker status are applied as single or combined environmental effects. It should be stressed again that this approach lends itself equally well to single or multiple vector analysis and evaluation of these vectors under conditions of distortion (i.e., band pass, noise, speaker disguise, etc). In any case, the approach has proved to be quite successful as it has led to: 1) the modification (improvement) of vectors, 2) an understanding of which vectors are most robust under specific speaker/system conditions, 3) the elimination (or deferral) of one vector, 4) the discovery that "wrap-around" (during data

processing) will be a major problem for any computer based approach to the problem and 5) an enhancement of the theory that natural speech features exhibit a robustness not enjoyed by other signal processing approaches.

The Forensic Model: While it is useful to employ a basic research approach to study the conditions under which the experimental vectors will be most powerful and how they can be upgraded, the large/homogeneous population approach cited above is one that will be but very rarely encountered in the criminal justice/law enforcement milieu. Rather, the typical paradigm is one where there is an unknown speaker and a group of 1-N (sometimes 6-10) suspects. An alternate scenario is where there is more than one unknown speaker and perhaps but a single suspect. In this case, it is necessary to provide additional suspects/subjects. In any case, the forensic model is one of match/no match and the most powerful approach is not one-on-one (i.e., not one known vs one suspect) but rather where the unknown voice is compared to the one or more suspects (i.e., the knowns) and the comparisons are made in the milieu of up to 10 foils (or innocent talkers). We included a number of "simulated" studies of this kind among our protocols and, more important yet, have applied the vectors and procedures to several actual cases. In all of these forensic cases, high levels of secondary reliability were included. That is, the actual "unknown" speaker actually was known in the simulated experiments. The results in the actual cases were "validated" by: 1) verification of findings by two or more blind listening panels and/or 2) ultimate knowledge of the actual match between the unknown voice and the suspect. The high level of success in these actual and simulated cases is encouraging. Accordingly, the results based on the forensic model will be considered first.

RESULTS

Forensic Model:

The data generated by the experiments based on the forensic model are presented first. This approach is employed because these results demonstrate that the procedures utilized have merit even though they have not as yet been finalized. Indeed, it still is necessary to 1) identify a final set of vectors for the SAUSI profile, 2) determine the robustness of each under all environmental conditions and combinations (upgrading any that require improvement), 3) evaluate additional distance measures, 4) improve decision criteria and 5) convert all processing to software. Nevertheless, substantial success with this approach appears to have been realized; data from key experiments follow.

Table 1 summarizes the decisions projected for seven actual criminal/civil cases; the tapes for these cases were of field quality but were only poor (noisy) for cases No. 2 and 6. The names have not been revealed for obvious reasons but are available upon request.

Table 1. Summary table of the results of seven actual cases evaluated on the basis of the Forensic Model (details can be provided upon request). Profiles of 3-5 vectors were utilized with LTS, SFF and TED always included. Each decision based on this profile was contrasted with scores obtained from 2-3 listener groups making (blind) ABX judgements. The unknown was compared to all knowns and foils; sometimes there was more than one known suspect. Since information about the suspect (and case disposition) ultimately became available, it is included in the last column.

Case No.	Speakers K-1	K-2	F1	F2	F3	F4	F5	F6	F7	F8	F9	A-P Level/Data Above	Below	%	Apparent* Guilt Yes	No
1	11	NA	-	-	-	-	-	84%	-	NA	NA	X(for F6)		2		X
2	76	-	18	-	-	-	-	-	-	-	NA		X	5	X	
3	88	-	-	-	-	-	-	-	-	-	-	X		3	X	
4	69	61	-	-	-	-	-	-	-	-	NA		X	10	X	
5	-	NA	-	-	57	11	-	43	-	-	-	NA**	NA	NA		X
6	89	82	-	-	-	35	-	-	-	-	NA		X	5	X	
7	86	NA	-	-	-	-	-	-	11	NA	NA		X	7	X	

NA = Not Applicable
\- = Chance levels
A-P = Aural-Perceptual level
* = Apparent guilt is based on case disposition.
** = 70% of all listeners indicated that K and U were different.

The results were accepted by all parties (including the court in question) in four instances, although no trial was held in two. Two of the other cases were internal to a large company and, finally, a court rejected the approach as "premature" in one instance. In five of the cases (2,3,4,6,7), the known and unknown talkers apparently were the same individual (on the basis of external evidence) and in two cases (1,5) they were thought to be two different people. The agreement with the aural-perceptual data was extremely good except for the case No. 5 where listener responses were quite variable. Table 2 provides similar data for simulated field cases; the data are presented in a somewhat different form. Moreover, the known talker was tested against himself in two cases and samples of the so-called "unknown" talker's speech entered twice in two others. By-and-large, these data show a clear cut identification in the proper direction (known with known; unknown with known) except for Experiment No. 2 where one of the foils (F3) was sometimes identified as being the same person as the known talker. It should be of interest to note that this research was completed before work on the intensity vector was deferred and, while data from this parameter set contributed to the decisions made, the contribution was but minimal.

In some ways, the data from Tables 3 and 4 reflect the basic research approach as the experiments are carried out on the 25 (homogeneous) male subjects utilized in that series; i.e., these nine subpopulations are drawn at random from the larger group of men used in the basic experiments. Nevertheless, the forensic model can be seen to be controlling in both sets of studies. Please note also that the data are for a single vector (LTS or power spectra) and good recordings of normal speech were used. In the first of the two sets of investigations (Table 3), speakers were drawn at random with one of them selected as the known. Thus a sample of the "known" or test subjects voice was compared to another sample of his voice plus those of seven other subjects. As can be seen, the known and known were matched in all cases (100% identification). In the second experiment (see again Table 4), an attempt was made to "stress both the model and the vector. In this case, selection for each of the subjects was made as a consequence of evaluation of the basic research data. In all cases the "known" subject was a person whose speech had been seriously confused with at least two other individuals (and sometimes as many as five) and whose identity had been confounded (at least occasionally) with all of the other foils utilized. Thus, the task was to see if the experimental (known) talker could be identified even though other talkers who sounded very much like him were included in the group. Please note also that four distances were compared for robustness as decision approaches. In any case, and as may be seen from examination of Table 4, the absolute and Steinhaus distances resulted in 60% correct identifications (plus several "near misses"); the Euclidean distance approach correctly identified half of the talkers. Admittedly, the speech samples used were of good quality and reasonably contemporary. It is a striking finding, nevertheless, that individuals who sound similar to each other -- and whose identity is often confused -- could be deferentiated in so many cases and on the basis of so little data.

Table 2. Summary table of four field experiments using a four-way profile, i.e., the combined SFF, LTS, INT and TED vectors. In each case, the known talker (K) was matched to each of the suspects (U's) and foils (F's). In Experiment #1 and #3, the known talker was, in addition, matched to himself. Profiles are based on 10 nearest neighbor approach.

Experiment	Subject										
	K	U1	U2	F1	F2	F3	F4	F5	F6	F7	F8
Experiment #1											
Rank	1	3	2	5	4	6	10	9	6	8	NA
4-V Score	1.0	3.8	3.1	6.0	4.5	7.0	8.0	7.4	7.0	7.2	NA
Percent	100	82	86	-	51	-	-	-	-	-	NA
Experiment #2											
Rank	NA	1	2	-	-	3	-	-	-	-	-
Percent	NA	87	63	-	-	53	-	-	-	-	-
Percent profile	NA	91	82	-	-	44	-	-	-	-	-
Experiment #3											
Rank	1	2	NA	-	-	-	3	-	4	NA	NA
4-V Score	1.0	2.1	NA	-	-	-	3.8	-	4.7	NA	NA
Percent	98	88	NA	-	-	-	40	-	36	NA	NA
Experiment #4											
Rank	NA	1	NA	3	-	4	-	-	-	2	NA
Percent	NA	81	NA	36	-	35	-	-	-	44	NA
Percent profile	NA	81	NA	37	-	37	-	-	-	47	NA

K = Known Talker
U = "Suspect" or Unknown
F = Foil

NA = Not Applicable
- = Chance Level

Table 3. Summary table of nine experiments -- utilizing the LTS vector (only) in a typical forensic situation. Subjects were drawn randomly from a homogeneous (age, size, health education) group of young males. A three nearest neighbor approach was used. Note the 100% correct identification.

Test Subject	Experimental Subjects							
1	1/1*	F6 -	F9 -	F11 -	F14 3	F18 2	F21 -	F23 -
5	F2 -	5/1*	F9 -	F14 3	F18 2	F20 -	F21 -	F25 -
9	F4 -	F8 -	9/1*	F10 3	F16 -	F18 -	F22 -	F24 2
12	F1 -	F5 -	F9 -	12/1*	F15 -	F20 2	F24 -	F26 3
15	F3 -	F7 -	F12 -	15/1*	F17 3	F20 -	F23 -	F26 2
17	F2 -	F4 3	F8 -	F15 -	17/1*	F18 -	F21 -	F25 2
23	F1 -	F6 -	F13 2	F14 3	F19 -	F21 -	23/1*	F24 -
24	F2 -	F7 -	F12 -	F15 3	F18 -	F23 -	24/1*	F26 2
25	F3 -	F6 -	F11 -	F14 2	F17 3	F19 -	F22 -	25/1*

1* = correct identification

Table 4. Summary table of LTS (only) matches in a typical forensic situation. Foils were either subjects that sounded quite similar to the test subject or had been systematically confused with him in a number of prior experiments.

Distances	Test Subject	Experimental Subjects							
	2	2	F4	F6	F9	F13	F15	F20	F26
EUC		1*	2	-	-	-	-	-	3
ABS		1*	3	-	-	-	-	-	2
STH		1*	3	-	-	-	-	-	2
MAX		1*	3	2	-	-	-	-	-
	3	F1	3	F6	F9	F10	F14	F19	F21
EUC		-	-	-	-	2	1	-	3
ABS		-	3	-	-	2	1	-	-
STH		-	3	-	-	2	1	-	-
MAX		-	3	-	-	2	1	-	-
	4	F1	4	F5	F7	F8	F13	F19	F25
EUC		-	2	-	-	3	-	-	1
ABS		-	1*	-	-	3	-	-	2
STH		-	1*	-	-	3	-	-	2
MAX		3	-	-	-	2	-	-	1
	6	F2	F3	6	F7	F10	F12	F18	F23
EUC		-	3	1*	-	2	-	-	-
ABS		-	3	1*	-	2	-	-	-
STH		-	3	1*	-	2	-	-	-
MAX		3	-	1*	-	2	-	-	-
	8	F1	F5	F7	8	F16	F19	F24	F26
EUC		2	-	-	-	3	-	-	1
ABS		-	-	-	-	2	3	-	1
STH		-	-	-	-	2	3	-	1
MAX		1	-	-	2	-	-	-	3
	10	F1	F6	F8	10	F13	F21	F24	F25
EUC		-	-	2	1*	-	-	3	-
ABS		-	-	-	1*	2	-	3	-
STH		-	-	-	1*	2	-	3	-
MAX		-	-	3	-	-	-	1	2

Table 4 (Continued)

Distances	Test Subject	Experimental Subjects							
	11	F4	F5	F10	11	F17	F19	F22	F24
EUC		-	-	1	3	-	2	-	-
ABS		-	1	2	-	-	3	-	-
STH		-	1	2	-	-	3	-	-
MAX		-	-	1	2	-	3	-	-
	14	F2	F3	F7	F10	F11	14	F19	F20
EUC		2	-	-	3	-	1*	-	-
ABS		2	-	-	3	-	1*	-	-
STH		2	-	-	3	-	1*	-	-
MAX		1	2	3	-	-	-	-	-
	16	F3	F6	F13	F15	16	F19	F25	F26
EUC		-	-	1	-	2	-	3	-
ABS		-	-	1	-	2	3	-	-
STH		-	-	1	-	2	3	-	-
MAX		-	-	1	-	2	-	3	-
	21	F3	F6	F8	F10	F17	F18	21	F24
EUC		-	-	-	3	-	2	1*	-
ABS		-	-	-	3	-	2	1*	-
STH		-	-	-	3	-	2	1*	-
MAX		3	-	-	2	-	-	1*	-

EUC = Euclidean Distance
ABS = Absolute Distance
STH = Steinhaus Distance
MAX = Maximum Distance

1* = Correct Identification

The final experiment to be reported in this section involves comparisons of the robustness of the same two vectors as identification cues (and generally the same population) for both the basic and forensic research models. Examination of Table 5 will reveal several relationships that should be of interest. First, as expected, LTS was the better predictor of speaker identity (than was SFF at that time anyway). Second, it can be seen how this approach permits various of the experimental distances to be evaluated for predictive strength. For example, the Steinhaus distance appears to be the overall most powerful (at least for these conditions). These data can be used for other evaluations also. However, the main contrast in this case is the one between the two research approaches (basic/forensic). While several comparisons can be made, there are one or two that are rather significant. That is, it can be seen that if the forensic model is utilized with the Steinhaus distance, 96% correct identification can be expected for LTS (this vector reflects general voice quality) and 72% for SFF (a quantitative measure of f0 which, in turn, reflects pitch level). Also demonstrated is the cited fact that, while the basic research approach is appropriate for the evaluation of vector strength, the forensic model permits a far superior prediction of speaker identity.

Basic Research Approach:

So many experiments have been carried out on the basic nature and development of our speaker identification vectors/approaches that a serial listing would be counterproductive and, perhaps, even confusing. Accordingly, we have elected to include only a number of sample experiments -- primarily to demonstrate 1) the investigational approaches utilized, 2) the depth and breadth of this project, 3) how the information derived from the basic research can lead to improved profiles/vectors and 4), of course, a summary of the key results.

LTS: A Sample Vector:

Since more research has been carried out on the long term spectral vector (a vector that is thought to reflect the general voice quality of a speaker), the data here can best be used to describe the cited elements of this research program.

As can be seen from Table 6, a vector can be used to study sample size; in this case, however, the vector was too robust (i.e., the scores were too high for proper evaluation); hence this research currently is being replicated on degraded speech and with other vectors. Table 7 provides data on two variables: 1) the use of different mathematical (distance) approaches in the decision process and (again) the effect of the presence of individuals who sound very much like the individual to be identified. Note that this experiment reflects combination of both basic and forensic protocols, and the control of speech contemporariness/subject/sample quality. In any case, certain relationships are observable (again the scores are a

Table 5. Summary table of means from several parallel experiments carried out on the same two vectors (LTS and SFF) twice: once utilizing the basic research approach (N=25) and the other, the forensic model (N=K + F1 to F7). Speech samples were of good quality but were text independent and non-contemporary. Values are percent correct identifications.

		Basic			
	Distance	1	2	3	Forensic
SFF					
1)	Euclidean	19	39	50	56
2)	Absolute	27	42	58	60
3)	Steinhaus	27	42	58	72
4)	Maximum	19	39	46	56
LTS					
1)	Euclidean	62	69	73	88
2)	Absolute	27	35	46	88
3)	Steinhaus	62	69	73	96
4)	Maximum	54	65	69	88

Basic results are derived from the three-nearest-neighbor approach.

Table 6. Percent correct classification for various length (10-40 sec.) contemporary speech samples. The LTS vector served as basis for the comparisons. All values are in percent; N=25.

A. Long Contemporary Samples

Test Sets	Reference Sets First	Second
First	-	100
Second	96	-

B. Short Contemporary Samples

Test Sets	Reference Sets First	Second	Third
First	-	92	96
Second	96	-	96
Third	92	100	-

little too high but are reasonably useful). First, the Steinhaus distance appears to be somewhat more useful as an indicator of speaker identity -- at least from the multidimensional set of parameters that make up this procedure. Second and as expected, randomly selected subjects were identified somewhat more accurately than those where doubles (sound-alikes) were introduced into the foil group. What was a little startling was the robustness of Steinhaus for this second procedure. If this relationship holds up, it very well may lead to improvement in the predictive ability of the entire system.

The data reported in Table 8 contrast different reference sets and different sample combinations as a function of vocal disguise. As can be seen, the LTS vector (in its present form) is not yet a good predictor of speaker identity when speech disguise is employed. Also apparent is the fact that different types of disguises will affect the identification process by different magnitudes. For example, the "pencil in mouth" and "pinched nose" procedure have the least effect on the data while register shifts (falsetto) appear to be the most detrimental to this vector as an identification cue. Finally (re: this section anyway), Table 9 provides data demonstrating that the approach being utilized in this case is completely text independent -- a result that was totally unexpected. Note also that the four distance measures of interest are again contrasted but in this case, the unexpected high identification scores may have obscured the actual differences among them.

Other Vectors:

As stated, work on the INT vector (INT is thought to reflect the level and variation of vocal intensity and/or perceived loudness) has been deferred until an improved set of parameters can be developed. Thus, the systematic and long-term research that has been carried out was concentrated on SFF or speaking fundamental frequency (pitch of voice), TED or Time vs Energy Distributions (plus speaking rate) and VFT or vowel formant tracking; a JIT or vocal jitter vector has been introduced but the research in this area is just being initiated.

In some cases, it is useful to contrast a developing vector against a more established one. Such contrast can be found in Table 10; here it can be seen that the identification value of SFF can vary between 44% and 78% under various conditions while LTS will vary from 84 to 100% in the same environment. The problem with this set of experiments lay in combining the two vectors of interest (LTS/SFF). In this case, correct identification by LTS was so high that no data relevant to the effects of adding the SFF vector could be obtained. We are now replicating this segment of the research but with the speech samples sufficiently degraded that the cited effects may become apparent. SFF also has been tested for a group of nine disguises in much the same manner as LTS (see Table 11 and see again

Table 7. Comparison of four distance criteria -- Euclidean, Absolute, Steinhaus and Maximum -- based on 16 forensic type experiments (i.e., the unknown matched to the known plus 7 foils). The LTS vector was utilized, as were good samples of contemporary (same day) speech. Eight of the experiments utilized subjects randomly drawn from a homogeneous population of young adult males; only "sound-alikes" and talkers confused with the unknown were utilized in the other eight. All values are in percent.

Distance	Randomly Selected Subjects/Foils	Foils Confused with Unknown	Mean
Euclidean	62	50	(56)
Absolute	88	75	(82)
Steinhaus	88	100	(94)
Maximum	88	75	(82)
All Distances	(82)	(75)	

Table 8. Percent correct classifications for disguise (LTS is the vector of choice). Reference sets consisted of four normal readings, each matched individually to the Test set, and matched in combination; it consisted of approximately 40 seconds of the "Grandfather Passage," whereas the Test sets contained 20 sec.

	Reference Sets					
	1	2	3	4	Combined No.1	Combined No.2
Test Sets						
Normal	52	72	64	64	64	82
Pencil in Mouth	28	40	32	32	40	40
Muffled with Hand	16	24	24	36	28	36
Pinched Nose	44	36	40	40	40	48
Free Disguise	8	20	8	24	16	24
Falsetto	4	12	4	4	8	16

Table 9: Summary table of research focused on text dependency utilizing the LTS vector. All values are correct identifications (in percent) for 25 subjects recorded under good quality conditions; samples were contemporary. Four distances were tested as was text dependency (R-R) and sample size (R=40 sec.; T=20 sec.). The reference level data are obtained in the same manner as studies reported in the early literature.

Distance	Reference Level	R1-R2	R1-R3	R2-R3	T-R1	T-R2	T-R3	Overall
Euclidean	100	88	96	92	88	85	92	90
Absolute	100	92	96	92	85	92	100	93
Steinhaus	100	92	92	81	85	88	100	90
Maximum	81	69	88	62	65	68	85	73

Table 10. Summary table of reliability runs of the verification procedure. Subjects were 25 young adult males; speaking condition was normal.

Analaysis	Percent Correct Verification			
	SFF-1	SFF-2	LTS	LTS/SFF-2
Posterior probability	71	78	100	100
Jackknife test	53	52	96	95
Identification test				
First Run	60	44	96	84
Second Run	60	52	100	100
Third Run	60	68	100	100
Fourth Run	50	48	92	92
Fifth Run	53	52	92	100
Mean	57	53	96	95

Table 11. Summary table of the effects of speaker disguise on the identification ability of the SFF vector. Subjects were 20 young adult males; comparisons were made for nine types of disguise to a normal reference set. Only the SFF vector was utilized.

Condition	Disguise to Normal Comparison percent correct
1. Pencil in mouth	70
2. Whisper	20
3. Pinched nose	45
4. Slow rate	45
5. Hypernasal	25
6. Falsetto	5
7. Muffled (hand)	35
8. Hoarse	10
9. Free disguise	20
Normal (control)	58

Table 8 for the contrasts). Note, here, the high scores for "pencil in mouth" and the particularly degrading effect on identity resulting from the changes in laryngeal production (falsetto, hoarseness, and whisper).

The TED vector was found to vary in its predictive value in earlier experiments (see previously submitted progress reports) and did not perform as well as did the other vectors even after modification. Accordingly, it was modified yet again and a large experiment carried out that simultaneously evaluated: 1) its reliability (Euclidean; runs 1-4), 2) the effect of distance measure (the four listed in the first column) and 3) effects of sample size (N=5-25). Indeed, the data reported in this table demonstrate that high (correct) identification levels are possible with TED; hence, the newly structured vector in this area now appears reliable -- especially when used with the Steinhaus distance procedure. Moreover, application of the forensic model should lead to good predictions by TED. Most encouraging of all was the very high TED identification level that occurred when this vector was coupled to the Steinhaus distance (re: the N=10 identification experiment).

As stated, research on a possible JIT (jitter or vocal roughness) vector is just being initiated. Table 13 provides some insight into the approaches being taken in this regard. Two different (digital) extraction procedures were contrasted in the first experiment -- and as a function of speaker, vowel and f0. As can be seen, the two approaches are quite similar except in two instances and the findings confirm much of what is known about jitter and its relationship to the acoustic theory of speech production. The JIT vector currently is being evaluated as a speaker identity cue. As can be seen from observation of the data on Table 14, it demonstrates potential in this regard but does not seem to be totally independent of vowel.

A decision was made early in the grant period to evaluate the power of various vowel formant tracking vectors (VFT) on the basis of single vowels and syllables -- at least initially. The results in this regard have been strikingly successful and some of the results of this series of experiments can be found in the tables to follow. As may be seen in Table 15, VFT vector scores were surprisingly high even when identifications were based on only two samples of the vowel /i/. While there is some indication that sample size may be a factor, it appears that position may not be -- that is, if phonemic context is similar and speech is reasonably contemporary (see Table 16).

As stated, both vowels and nasals appear to provide reasonably good articulatory data for use in identifying speakers from their speech. The data found in Table 17 addresses this issue; moreover,

Table 12: Partial results from several experiments with the TED vector. The first set of columns include reliability (discrimination) data based on test/preference samples of 20 sec. (text independent) that are good quality and contemporary. The identification experiments (second set of columns) also utilize good quality speech. All values are in percent correct identification.

	Reliability Experiments (N = 25)				Identification Experiments		
Distance	Run-1	Run-2	Run-3	Run-4	N=25	N=10	N=5
Euclidean	84	100	84	84	44	60	40
Absolute	-	-	-	100	-	-	-
Steinhaus	-	-	90	100	40	90	-
Maximum	-	-	-	64	-	-	-

Table 13. Comparison of fundamental frequency and jitter for four isolated vowels. Data are calculated from FFI-8 output with both the SFF and JIT programs being used -- and (in the second case) by digitizing the signal at 20 kHz and determining pitch periods by axis crossing.

		FFI/SFF/JIT		Digitization/Axis-crossing	
Subject	Vowel	f0	Jitter	f0	Jitter
M1	/u/	135.0	0.46	137.6	0.44
M2	/i/	147.9	0.91	154.7	0.86
M3	/a/	107.9	*	108.0	0.54
M4	/ae/	143.0	*	142.9	0.61

* Reliable data could not be obtained.

Table 14. Percent correct classification for the jitter vector (JIT). Normal speech samples produced by 25 male talkers were used as experimental material.

Test	Reference	1	2	3
First /i/	First /a,ae,u/	41.7	58.3	66.7
First /a/	First /i,ae,u/	33.3	66.7	75.0
First /ae/	First /i,a,u/	25.0	33.3	66.7
First /u/	First /i,ae,a/	16.7	41.7	50.0
First /i/	Second /i/	33.3	58.3	66.7
First /i/	Second /i,a,ae,u/	25.0	41.7	58.3

Table 15. Percent correct classification for the vowel formant tracking (VFT) vector. Normal speech samples only were used; talkers were males.

Speakers	Test	Reference	1	2	3
N=5	Second /i/	Third /i/	40.0	80.0	80.0
N=10	Second /i/	Third /i/	30.0	50.0	60.0
N=25	Second /i/	Third /i/	42.3	50.0	50.0

Table 16. Percent correct classification for the vowel formant tracking vector (VFT). Speech samples consisted of the vowel /i/ isolated from connected speech but with all /i/'s in the same phonemic context. Males (N=25) were used as subjects.

Test	Reference	% Correct 1	2	3
Second /i/	Third /i/	42	50	50
Second /i/	Third /i/, Fourth /i/	35	50	62
Second /i/	Third /i/, Fourth /i/, Fifth /i/	31	46	58
Second /i/	Third /i/, Fourth /i/, Fifth /i/, First /i/	35	39	62

the problem of contemporary/non-contemporary samples is studied simultaneously. Evaluation of the relationships in Table 17 will reveal that contemporariness may be a factor in the identification process. These data also demonstrate that the syllable /na/ (a low vowel coupled to a nasal consonant) may not be quite as good a predictive cue as was expected -- and certainly not as robust as the high front vowel /i/. This conclusion is pretty much confirmed by the results of the 20 experiments reported in Table 18. Here the variables include: 1) the speech sounds /i/ and /na/, 2) four distances and 3) randomly selected foils vs those foils that sound quite similar to the target speaker. These experiments are considered basic in nature even though the forensic model was utilized. In any case, of the four distances, Steinhaus was found to be the most sensitive predictor of identity and (as stated) the high front vowel /i/ a better predictor than /na/. It should be noted also that, when individuals who sound like the talker are present, the identification process is somewhat degraded. To summarize, the VFT vector appears to be a very powerful one; research is underway contrasting our software with the LPC approach provided by ILS. Also underway are studies contrasting five different vowels (see again Table 16).

Research on Vector Combinations -- The SAUSI Profile:

As has been indicated in the quarterly progress reports submitted earlier, a number of attempts have been made to further improve the mathematical decision criteria that support our profile or multiple vector array. The success of these efforts may be best assessed by re-examination of Tables 1-5. Moreover, details of our first attempts to obtain specific detail about relevant relationships associated with our approach may be found in Table 19 (another large experiment of this type -- but with a new "tree" statistical approach currently is underway) -- it is awaiting funding. In this case (i.e., F-19), the power of four and three vector combinations are contrasted (an earlier experiment included the INT vector; those results can be found in a previous progress report). As can be seen by examining the table, LTS is the most powerful vector and in many cases (except for SFF anyway) the addition of data from other vectors does little to improve its predictive ability (even the addition of data from all three other vectors adds only 3%). Even though the same problem does not seem to exist (or is not as debilitating) when the forensic model is applied, we have, nevertheless, revised these procedures and are replicating this experiment.

In summary, we believe that the stated progress made on this grant has been reasonably well reviewed -- even though a number of other single and multiple factor experiments have not been discussed (many were reported in the quarterly reports). Nevertheless, it would appear useful to conclude this report by listing the results of two other related experiments; these data may be found in Tables 20 and 21. In this case, the identification strength of LTS (Table 20) and

Table 17. Percent correct classifications for the VFT vector, utilizing the syllable /na/. The Reference sets consisted of four normal readings, each matched individually to the Test set, and matched in combination. The three nearest neighbors (first, second and third choices) are displayed.

	Test Set			
Reference Set	First Choice	Second Choice	Third Choice	Total
1	16	8	0	24
2	40	20	8	68
3	36	8	8	52
4	12	4	20	36
combined	40	4	16	60

Table 18: Summary table of 20 experiments providing data about the robustness of single vowels/syllables as identification cues. The forensic model is utilized; four "distances" are included. Stimuli include the vowel /i/ and the syllable /na/; subjects for half the experiments were randomly drawn from a homogeneous population, with the other half from a group of "sound-alikes". All values are in percent of correct identifications for five studies.

Distance	Vowel /i/		Syllable /na/		Means
	Random	Sound-alike	Random	Sound-alike	
Euclidean	40	40	20	20	30
Absolute	80	20	80	20	50
Steinhaus	100	40	60	40	60
Maximum	80	40	40	0	40
Means	75	35	50	20	

Table 19: Summary table of two primary verification procedures. The nearest neighbor (serial and weighted) approach was utilized. Subjects were 25 adult males.

VECTOR	EXPERIMENT	
	THREE Vectors	FOUR Vectors
SFF	44	50
LTS	68	74
TED	36	35
VFT*	--	50
SFF/LTS	72	73
SFF/TED	56	54
SFF/VFT	--	54
LTS/TED	56	46
LTS/VFT	--	69
TED/VFT	--	42
SFF/LTS/TED	68	69
SFF/LTS/VFT	--	69
SFF/TED/VFT	--	70
LTS/TED/VFT	--	62
SFF/LTS/TED/VFT	--	77

* A somewhat limited test as the VFT procedure did not provide data for all subject/condition combinations.

SFF (Table 21) were evaluated as a function of channel distortion (noise, passband), disguise (four types) and vector construction (six methods of combining parameters; however, only No. 4 is reported)* -- all on the basis of the three nearest neighbor procedure. As may be seen, the four types of disguise researched degraded LTS but little and except for "muffled with hand," SFF only modestly -- even when these disguises were combined with noise. On the other hand, note the severe degradation that occurs when disguise is combined with telephone transmission (actual not simulated) and that LTS is degraded more seriously (almost to chance levels in most cases) than is SFF. Data such as these are included in order to demonstrate how we are systematically investigating many of the thousands of basic and applied issues associated with speaker recognition. The basic research is necessary as it leads to improved knowledge about vocal function, acoustic phonetics and speaker recognition in general and specific data about the speaker identification process and the strength of our vectors in particular. Even more importantly, however, it leads to possible criminal justice and law enforcement applications of an effective speaker identification profile. The success (see the previous section on the forensic approach) of our partially developed method demonstrates our contentions in this regard.

CONCLUSIONS

A number of conclusions are possible.

1) Our postuation that the natural speech feature approach to speaker identification is superior to other signal processing techniques appears to have been strongly supported by the research carried out under 84-IJ-CX-0014.

2) The vectors selected (as modified) appear suitable for the speaker identification task.

3) The vector combination (or profile) approach to the identification task appears warrented as specific vectors appear to be more robust in certain environments, yet others more viable in yet other situations.

* Three of the approaches were not acceptable; of the other three, individual scores varied +/- 10%. However, since the patterns observable for the No. 4 approach are typical, only these two tables (T-20; T-21) are included in this report.

Table 20: Percent correct classification when disguised speech samples of four types were matched to normal, noise and telephone bandpass samples; LTS was the vector of interest. A test set was compared to a reference set combination of four samples with pooled variance. Males (N=25) were used as subjects. The data reported here are for one evaluation out of six.

Test	Reference	1	2	3
Pencil in mouth	Normal	69.2	76.9	80.8
Pinched nose	Normal	65.4	76.9	84.6
Slow rate	Normal	46.2	53.8	57.7
Muffled with hand	Normal	46.2	50.0	53.8
Pencil in mouth	Noise	65.4	76.9	80.8
Pinched nose	Noise	65.4	88.5	96.2
Slow rate	Noise	53.8	61.5	69.2
Muffled with hand	Noise	30.8	50.0	65.4
Pencil in mouth	Telephone Bandpass	7.7	11.5	19.2
Pinched nose	Telephone Bandpass	3.8	7.7	11.5
Slow rate	Telephone Bandpass	3.8	7.7	7.7
Muffled with hand	Telephone Bandpass	3.8	15.4	19.2

Table 21: Percent correct calssifications for SFF when four different disguise conditions were matched to normal, noise and telephone bandpass samples. A test set was compared to four reference sets with pooled variance. Males (N=25) were used as subjects. The data reported here and for one evaluation (No. 4) out of six.

Test	Reference	1	2	3
Pencil in mouth	Normal	46.2	61.5	61.5
Pinched nose	Normal	34.6	42.3	65.4
Slow rate	Normal	34.6	42.3	50.0
Muffled with hand	Normal	19.2	42.3	53.8
Pencil in mouth	Noise	30.8	53.8	61.5
Pinched nose	Noise	23.1	53.8	69.2
Slow rate	Noise	30.8	46.2	50.0
Muffled with hand	Noise	15.4	26.9	50.0
Pencil in mouth	Telephone Bandpass	15.4	23.1	34.6
Pinched nose	Telephone Bandpass	15.4	15.4	30.8
Slow rate	Telephone Bandpass	7.7	19.2	30.8
Muffled with hand	Telephone Bandpass	11.5	19.2	26.9

4) Our postulate that a two level research approach to the identification task -- i.e., use of both the basic research and forensic modesl -- has been strongly supported by this project. The extremely high levels of correct identification for actual/simulated field evaluations (i.e.., the forensic model) demonstrate this contention.

5) Even though the basic research on the process is not complete, our approach is proving to be field effective in the forensic milieu.

6) The SAUSI (Semiautomatic Speaker Identification) procedure appears of demonstratable merit. However, additional research will be necessary to refine it even though it already appears nearly field-ready. It should be possible to complete the required research in about two years.