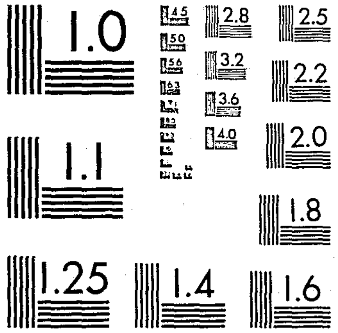

5/23/84

92658

MACHINE AIDED SPEAKER IDENTIFICATION

(Phase-1)

Semiautomatic Speaker Identification System (SAUSI)

Investigators

Harry Hollien, Ph.D.
James W. Hicks, Jr., Ph.D.
Leslie H. Oliver, Ph.D.

Final Report of Phase-1 Research: IASCP/DJ/006

Report Prepared By:

Harry Hollien
J.W. Hicks, Jr.

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

Contract #82-IJ-CX-0034

November 1, 1983

MACHINE AIDED SPEAKER IDENTIFICATION

H. Hollien, J.W. Hicks, Jr. and L.H. Oliver
IASCP: University of Florida

Introduction

The report to follow serves two functions. First, it constitutes a Final Report for NIJ grant 82-IJ-CX-0034. It also provides a brief review of the progress that was made during the Phase-1, or first year, of what is hoped will be a three-year program of research. As it turns out, a number of accomplishments have been realized. Moreover, the research has proven to be interactive in nature. Specifically, interpretation of the data being produced has led to the direct modification of the subsequent research protocols and modification, in turn, has improved vector robustness. For example, all ot the vectors under study have been modified at least twice -- and in less than a single year. These interactions have led to an improved experimental approach. Thus, in order that a reasonable understanding of the nature and impact of this project be established, this Introductory section will be followed by seven others. They include Project History, Project Perspectives, The Vectors Utilized in the Research Program, Progress to Date, The Phase-2 Experiments, The Phase-3 Thrust and Epilogue; these sections should provide perspectives/details about the nature and extent of the project, what has been accomplished -- and what it is hoped will be accomplished.

Project History

This project actually was initiated over 15 years ago. Prior to the time of first NIJ support (1982-3), the research conducted by our group was carried out on a relatively piecemeal basis -- primarily due to a lack of funds. Although the project appeared to be quite central to the NIJ mission (especially relative to the needs of law enforcement and security agencies, the courts and the judicial system) we did not apply for assistance because we were out of phase with LEAA at that time and we knew it (i.e., we did not accept "voiceprints" as valid and we predicted that a useful speaker identification system probably could not be designed and implemented until the late 1970's). Accordingly, we sought support from other sources. Grants or contracts were obtained from the following.

1) Our respective universities -- especially the University of Florida. We received 6-7 small grants (four from UF). However, none of the administrators at these universities were in a position to offer large grants; hence, this support was in the nature of small seed-money type awards. 2) The Max and Victoria Dreyfus Foundation assisted us from time-to-time. However, this organization was not in a position to provide either large or long-term grants. 3) Spinoff funds from NIH grants. The research in speaker identification was (and is) rather tangential to the NIH mission. Accordingly, research support could not be obtained from this source -- either directly or on a continuing basis. 4) The Army Research Office provided us with temporary, short-term support. While this support was extremely helpful (and much appreciated), it was not systematic enough to permit us time to make the appropriate advances. In any case, ARO is similar to most military groups. They perceive their needs (in this area) to be speaker verification (not identification) and it is quite difficult for them to justify contractor support of an identification program. 5) The Polish Government supported this project to some extent. One of the early investigators was a Polish national who served a postdoctoral apprenticeship with the principal investigator. Further research was carried out during exchange visits. At that time, offers were made for continued support of a major program in speaker identification -- but all research was to be carried out in Poland! The reasons why this offer was respectfully refused are obvious!

In short, over the years, a total of 12 scientists struggled with the issue and over 40 experiments were completed. In turn, these projects led to nearly 30 publications. The breakthrough came in 1982 when a subset within our scientific group judged that it was time to carry out major research on the approaches that had been developed. Application was made to NIJ and the present one-year contract was awarded. Thus, although many individuals and agencies contributed materially to this thrust; the current successes we are having with the approach is due to the forsightedness of NIJ and its administrators. Phase-1 now is being completed (and successfully); Phase-2 is structured and awaiting support. Moreover, the Phase-3 potential now is obvious (see below) and should be seriously considered relative to implementation.

Project Perspectives

Most approaches to the speaker identification problem employ some form of signal analysis. While no systems currently are efficient enough to be placed on-line, a number show promise -- at least under optimum conditions. However, very few of the existing purely signal analysis approaches (i.e., those that ignore speech features) are robust enough to withstand the dehabilitating effects of noise and signal distortions -- nor would they be practical for the forensic situation. Admittedly, techniques are available that permit digitalization of the entire signal with the speech portions reconstructed. However, these approaches are time and energy consuming, costly and, anyway, only result in an approximation (often very crude) of the original signal. Thus, while those (resulting) approximations may be adequate for message (intelligibility) decoding, they simply are not powerful enough to permit the identification task to be carried out.

The solution to this problem appears to be a speaker identification system based upon a speech feature approach to signal analysis. The relevant research literature suggests that humans develop perceptual strategies when they attempt to make verification judgments. These include (among others) the processing of (1) vocal pitch level and variability information, (2) talker speaking time and rate patterns, (3) vocal intensity data, (4) subjective analyses of the talker's speech quality, phoneme usage, coarticulation -- and so on. Our response was to develop a speaker <u>identification</u> method which is based on multiple-parameter speech (feature) vectors. We then were able to test them under the most stringent of the forensic models (published and submitted reports are on file at NIJ; additional copies are available on request). To do so, we used single points in multidimensional space, short (sometimes very short) samples, open sets and signal distortions (noise, telephone bandpass , disguise). We found that the procedures tested (i.e., four of the five vectors listed below) were remarkably robust in the face of a

very difficult test situation.

To be specific, our general approach to the problem, is to evaluate selected acoustic and temporal parameters in order to study the <u>basic</u> identification problem (i.e., the relationship between intra- and interspeaker variability) and, if possible, to advance the status of our particular speaker recognition system. The investigations we outlined in our Phase-2 (continuation) proposal would build on existing research and generate information that currently is not available about the issue of interest. Moreover, it is expected that, ultimately, they would result in an operational system. In any case, our stated approach is to investigate a selected group of <u>natural</u> features (within the speech signal) which are thought to be idiosyncratic of an individual speaker. Test vectors are generated on this basis and evaluated singly and in combination: 1) in the laboratory (including simulated field conditions) and 2) in the field, where attempts are being made (and would continue to be made) to evaluate the approach as a function of a number of typical law enforcement scenarios. It is by the second process that this particular speaker authentication system would be refined for operational use. We are persuaded (by available data) that only "natural" speech analysis will be robust enough to meet the constraints imposed by the forensic modal -- and that many forms of basic signal processing would not.

<u>Vectors Utilized in this Research Program</u>

First, it should be noted that one of the features of this research is that our data-base already has been established -- it now consists of nearly 2000 recordings of 435 speakers (264 males; 171 females) variously producing 27 classes of utterances. Experimental samples include (depending upon the subset) normal reading/speaking, digits, free/controlled disguise, telephone transmissions, varying dialects, induced stress (two types) and speech involving covarying fundamental frequency-vocal intensity; moreover, any type of noise or system distortion can be added to the recordings.

As would be expected, these experimental samples also will be utilized in the Phase-2 research. Specifically, the five vectors that contribute to our overall approach have been chosen on the basis of: 1) high probability of discriminating among speakers, 2) enhanced utility when combined with other factors, 3) resistance to distortion, 4) availability, 5) convenience in modification and 6) compatability with computer processing. They are as follows:

a) The Speaker Fundamental Frequency Vector (SFF). Since the perception of vocal pitch has been shown to be a reasonable cue for speaker recognition, we have developed an acoustic analysis approach based on this speech feature. We are experiencing some success with SFF -- especially a modified version (a second modification is about to be carried out -- i.e., in Phase-2). In any case, the current vector is based on measures of central tendency and variability plus the frequency of f0 occurrence within semitone intervals. SFF data is obtained by analysis of the speech signal via the IASCP Fundamental Frequency Indicator (FFI-8) coupled to our PDP-11/23 computer.

b) The Long-Term Speech Spectra Vector (LTS). We have found, through our research, that LTS can predict the identity of speakers at very high levels (98-100%) even for relatively large subject groups --

at least in the laboratory. We also have been able to demonstrate that LTS is resistant to the effects of speaker stress, limited bandpass conditions and other distorting conditions. The approach utilized provides 40 parameters generated by a Princeton Model 4512 FFT spectrum analyzer coupled to our computer. As expected, it is proving (see below) to be a robust vector.

c) The Vowel Formant Tracking Vector (VFT). Our VFT vector consists of an 80 parameter cluster made up of three center frequency and two transition measures for each of the first two formants of four selected vowels (/i, a, ae, u/) within the test utterance (repeated for a second set). A high-speed Fourier analysis hybrid system currently is being evaluated. Specific vowel formant frequency windows are preprogrammed so as to make this system operable. The VFT vector is our newest one; it has only recently been placed on-line. Since prior research suggests that this vector should be a powerful one, and we wish to include it among our vectors, we already are working on a backup procedure -- that of the LPC approach to vowel formant identification.

d) The Temporal Analysis Vector (TA). Only modest research in the speaker identification area has focused on any of the temporal parameters that can be extracted from the speech wave. Nevertheless, there is strong logic that there are elements within this domain that can be utilized for recognition purposes. Our TA vector is composed of several sets of temporal parameters. It is developed by segmenting the speech signal into 10 equal intensity levels from which three distributions are obtained; specifically, 1) mean speech time (MST), 2) number of speech periods (NSP) and 3) percent speech time (PST). In addition, speech rate, total number of pauses and pause/time ratio are determined for the sample; a total of 33 parameters results. Data for this vector are obtained by use of a rectifier-integrator circuit coupled to the A/D converter of the PDP 11/23 computer. As will be seen from the progress report (below), this vector has enjoyed only mixed success; it is in the process of being upgraded. The dynamic "timing" characteristics of speakers now will be stressed rather than simply "time" on a static basis.

e) <u>Vocal Intensity Vector (VI)</u>. Extraction system: the PDP-11/23 and related software. The VI vector utilizes 40 (relative) intensity parameters including mean pressure level, range, variance and a slope algorithm consisting of (1) mean (intensity) rise rate (IRR), (2) mean intensity fall rate (IFR) and (3) the variance (i.e., standard deviation) of both. During the first phase of our speaker identification research, the predictive value of this vector was raised from chance to about 25% for both normal and distorted conditions. Since that time, we have further modified the vector configuration and expect that it will provide even more robust determinations relative to the recognition task. However, it may be necessary to completely reconstruct this vector and we are planning for this possibility.

f) <u>Other Vectors</u>. At present, additional vectors are being developed, placed on-line and subjected to pilot-level evaluations. These vectors include vocal jitter (now on-line), voice shimmer, phoneme analysis, vowel/consonant ratios and other speech features that could be idiosyncratic to the individual speaker. Moreover, we now will begin to reorganize the older vectors, and develop the newer ones, on the

basis of speaker's structural anatomy and functional physiology. Although we will defer the study of women for a period, we now realize that these size/proportion relationships may explain why the identification of women talkers was so much poorer than it was for men. Finally, while the main thrust of this project will focus on the existing vectors, the cited (new vector) experiments will be carried out on a systematic basis also.

<u>Progress to Date</u>

Many of the contributions made by the present investigators, of course, came prior to the initiation of this NIJ contract. To be specific, this research effort has led to nearly 50 publications (of all types) in over 15 journals, proceedings and books (such as the <u>Journal of the Acoustical Society of America</u>, <u>Acustica</u>, <u>The Journal of Phonetics</u>, IEEE journals, crime countermeasures proceedings and so on); over half of the publications (nearly 30) are focused <u>directly</u> upon the development of this speaker identification method. The balance are directed at basic or tangential issues, and/or the evaluation of parallel techniques. Listings of these publications are available upon request (as are most of the reprints).

The accomplishments during this, the first or Phase-1, year of the grant have been outlined in our five progress reports. They can be summarized as follows. Since we were aware that this year was the critical one, we chose to test our vectors as vigorously as possible -- rather than to use procedures that would enhance the obtained scores but leave us vulnerable to the possibility that we would find out subsequently that the approach was not robust enough to use in the forensic milieu. Accordingly, we established a procedure where only one data-point would result for each parameter (we employed a multidimensional space procedure but one with only a single value at each reference junction). Moreover, we used very short speech samples -- only one for each subject reference or test -- and distorted them (noise, telephone passband). We do not believe that a more rigorous test could have been developed. Fortunately this thrust was successful. While no vector attained a 95% or better correct identification level, the more robust ones maintained high identification rates (75%-9Ø%) and (by means of parameter upgrading) we were able to double -- and in one instance quadruple -- the predictive power of the others.

As stated, the specific results obtained from the many experiments completed prior to August, 1983 have been submitted to NIJ in the form of five quarterly reports. However, a brief review would apper relevant at this juncture. In order to summarize a number of them, mean data from 24Ø procedures (drawn from over 8ØØ) will be found in Table 1. The procedure of choice is the three nearest neighbor approach; the comparisons and types of test/reference contrasts are listed. The scores resulted from the first set of modifications only. In many cases, the second set of changes has further raised the identification levels -- as do certain of the multiple vector procedures (not shown). In any case, the following statements may be made; they are based on all data accumulated and analyzed.

1) Reasonable identification power (excellent in some instances) was found for all vectors even in the face of a severely structured

Table 1. Cumulative percent correct classification within one, two and three nearest neighbors.

Vector	Condition*	Males			Run**	Females			Run**
		1	2	3	#	1	2	3	#
LTS:	Nm/Nm	85	85	85	(#1)	62	81	85	(#4)
	Ns/Nm	19	35	35	(#1)	5Ø	73	85	(#1)
	Bp/Nm	8	12	19	(#1)	15	23	27	(#6)
	Ns/Ns	73	81	85	(#1)	38	54	62	(#6)
	Bp/Bp	19	35	42	(#1)	46	5Ø	65	(#1)
SFF:	Nm/Nm	39	62	73	(#4)	27	42	42	(#4)
	Ns/Nm	27	5Ø	58	(#6)	46	62	65	(#6)
	Bp/Nm	8	12	19	(#1)	12	15	35	(#1)
	Ns/Ns	38	65	77	(#1)	35	58	73	(#6)
	Bp/Bp	46	58	69	(#1)	31	42	58	(#6)
TED:	Nm/Nm	35	42	46	(#4)	27	31	39	(#2)
	Ns/Nm	12	23	27	(#1)	8	15	23	(#1)
	Bp/Nm	4	8	19	(#6)	4	4	8	(#6)
	Ns/Ns	12	15	23	(#6)	12	27	31	(#1)
	Bp/Bp	8	15	23	(#6)	4	23	31	(#6)
INT:	Nm/Nm	12	23	35	(#6)	8	8	15	(#4)
	Ns/Nm	4	8	23	(#1)	19	23	31	(#1)
	Bp/Nm	8	8	8	(#1)	8	11	15	(#1)
	Ns/Ns	4	12	19	(#6)	4	4	15	(#1)
	Bp/Bp	11	27	35	(#1)	15	19	19	(#1)

* Nm/Nm: Normal Test/Normal Reference
Ns/Nm: Noise Test/Normal Reference
Bp/Nm: Bandpass Test/Normal Reference
Ns/Ns: Noise Test/Noise Reference
Bp/Bp: Bandpass Test/Bandpass Reference

** Run #1: One reference set and one test set.
Run #2: Two reference sets (averaged) and one test set.
Run #3: Two reference sets (three nearest neighbor weighting) and one test set.
Run #4: Four reference sets (pooled variance) and one test set.
Run #5: Four reference sets (individual variances) and one test set.
Run #6: Four reference sets (pooled variance and 1Ø-nearest neighbor weighting) and one test set.

forensic evaluation -- i.e., where distortions were present, short samples were used and noncontemporary speech was employed.

2) The LTS vector (long-term power spectra) maintained very high predictive rates varying up to nearly 90% identification; the level of the TA (temporal) vector (approaching 40%) was found even though the original levels were established under less than reasonable conditions. The SFF (speaking fundamental frequency) vector was modified prior to the first series of trials and then again prior to the second. It improved from just above chance levels (under optimum conditions) to levels approaching or exceeding the TA vector. Shifts of these magnitudes simply were not expected.

3) Over 800 subexperiments were completed during the contract year. This accomplishment was made possible due to our simplified approach and our new computer.

4) There were sex linked identification differences. That is, scores for women generally were lower (sometimes dramatically so) than those for men; certain vectors were less sensitive to the voices of females, and others to males (see Table). The patterns were found to be systematic and these relationships must be studied further for, if the specific nature/causes of these differences are not determined, erroneous identifications could occur in the future. However, experiments on female subjects will be deferred until later and all Phase 2/3 research will be carried out exclusively on males.

5) The (machine processed) vowel formant tracking vector was completed and placed on-line. It replaces the cumbersome (and marginally accurate) hand processed vector used previously (and by all

other investigators). While we have not, as yet, fully evaluated our VFT scores, this vector promises to be a very powerful one (especially if previous research is to be believed). It would seem logical that testing should continue here.

6) As it turned out, we often employed too many parameters in certain of our vectors. Pattern matching accuracy dropped when the number of parameters exceed the number of events. Changes are being made here also (during Phase-2).

The Phase-2 Experiments

The planned experimental procedures are as follows. Each of the five vectors now have been (or are being) evaluated alone in a laboratory discrimination task utilizing relatively large subject populations and in the presence of both system (limited bandpass, noise, etc.) and talker (disguise, stress, etc.) distortions. As will be seen, we have found that the identification levels have approached 100% only for the LTS vector and for multiple vector analyses -- but all data were obtained relative to the very limited and difficult conditions relating to the forensic situation. These first experiments will be replicated; however, relative to a somewhat more favorable model. That is, the vectors would first be tested singly and then in combination but only after data clusters for each of the multidimensional referents had been developed on each subject used in the research -- a feature that legitimately can be applied to the forensic model. The experiments would be replicated again but with these subsequent trials carried out under noise/distortion conditions. At this point, special consideration will be given to the decision criteria utilized. That is, in any recognition situation, data usually are based on potentially open sets. Hence, if the test talker happens not to be within the reference group, the method still will "make a choice" (i.e., the person who is most like the test subject). Thus, the testing of various decision criteria would constitute an element of the research program for, in our program, the operator (not the "machine") must make the final decision. Moreover, we now are beoming interested in alternate methods of analysis -- such as consideration of growth dynamics, pattern matching, naturalness, correlations with body structure and so on (rather than just multiple distance measures). Finally, the five vectors will be analyzed in all possible combinations, simply by the application of appropriate statistical procedures, and this phase of the research would continue (with modifications) until the most effective system was identified.

Field Experiments. Forensic related field research would be carried out during the next two phases of the project; in these experiments we would attempt to establish the identity of talkers when their spoken messages were received over standard law enforcement communications gear and the received signal was mixed with "typical" noise (music, automotive, wind or similar). Foils (in the initial experiments anyway) would include 25-4Ø individuals; a multiple vector procedure also would be conducted.

To be specific, the main thrust of the Phase-2 research is to evaluate all vectors (both present and new) under more realistic operating conditions. These modifications are as follows. The multidimensional points will be expanded to areas by establishing multidimensional "spaces" (i.e., clusters of data "points"). We will do so by analyzing

more than one exemplar (per subject or suspect) and developing the target areas or fields mathamatically. We believe that this approach is a legitimate one because large numbers of voice exemplars (certainly 3-5) are routinely obtained from any person suspected of a crime where speaker identification is at issue. Since our approach is a potentially effective one, we suggest that it can be tested under more realistic circumstances and, hence, avoid having its predictive power artificially depressed. As stated, the vectors used will be modified (and simplified) on the basis of the Phase-1 experiments, and these modifications should improve vector robustness. A final issue needs to be reviewed. If the data resulting from the four previous field tests (see main proposal) are to be believed, our approach works better in the field than it does in the laboratory -- it is not clear why these differences occur (even though they are encouraging). Accordingly, a series of field tests would be initiated during the Phase-2 period. As with the experiments listed above, these investigations are detailed in the main proposal. However, here we would use tape recordings from previous cases or those furnished by law enforcement agencies (of course there will be no invasion of the talker's privacy). Also, techniques specific to the forensic milieu would be employed (8-1Ø foils, "sound- alike" foils, etc.). Finally, these data also would be contrasted to the mean scores obtained from a variety of listeners' panels -- a technique used in the field (also it constitutes a test of our vectors against listeners' strategies). It is by these experiments that we would expect to begin discovering why the procedure appears to be more sensative in the field than it is in the laboratory.

The Phase-3 Thrust

The main focus of the third phase of this project would be on "packaging" the speaker identification system that results from Phases 1 and 2. Currently, we utilize hybrid systems in the processing of two of the five main vectors and for all of the experimental vectors; even one of the three remaining vectors interfaces with the computer's A/D converter by means of a rectifier-integrator. We do so because the speed and accuracy of processing are enhanced by our equipment. However, other laboratories cannot be expected to have hardware arrays similar to ours -- especially since certain of our systems were invented and fabricated at IASCP. On the other hand, most major laboratories now have reasonably good computer support. Hence, the main thrust during Phase-3 of this project either would be to convert our vector processing to totally software routines or develop a hybrid type system in the form of computer type conversion boards plus software (especially for PDP 11, NOVA and Apple type computers). However, the software approach will be tried first. That our expectations are reasonable here is supported by our experience with the features employed. Initially all processing was carried out on hardware systems; as can be seen, we already are over half way complete relative to our conversion to software. Further, the project enjoys the services of one of the foremost young computer scientists in the country (L. Oliver) -- as well as his many associates and students.

Finally, it should be noted that experimental research will be carried out during Phase-3. For example, the field experiments (cited above) would be completed. Of course, we have not provided experimental protocols for this second series of investigations simply because we do not know what the first set of 5-10 Phase-2 experiments will reveal. Suffice to indicate that a major follow-up of the observed relationships would be carried out as we must identify the reasons our approach functions better in the field than in the laboratory. A final point must be made. It has been asked "What does semiautomatic mean; how automatic do you expect this method to be?". First, since we utilize hybrid systems, it is obvious that assistants are needed to operate the hardware and continually check on its accuracy (calibration, operational procedures, built-in error detection algorithms). However, our reference to "semiautomatic" extends beyond the simple checking for data validity -- which would be continued, of course. Specifically, it is our belief that machines should not make decisions about humans. Accordingly, all of our procedures, the decision criteria utilized, and the resulting vector/combined vector probabilities and variances have been developed so that reasonably intelligent judgements can be made by relevant decision-makers at all levels. Included are technicians and appropriate forensic scientists, prosecutors and defense attorneys and, ultimately, the courts, jurists and juries. Accordingly, our entire thrust -- as technically sophisticated as it is -- is designed to permit functional understanding. Indeed, this relationship must hold if the resultant data are to be useful to those individuals who must interpret and apply them.

Epilogue

We believe that an effective speaker identification system can be developed only: 1) by utilization of multiple vectors, 2) by the use of vectors that are inherently and directly related to human speech features and 3) by utilization of both laboratory and field evaluation techniques.

Our system is one that is based on appropriate theoretical constructs and successful experiments. As can be seen in the progress-to-date section, it already has been tested in identification type experiments and the results have been most encouraging. It now would appear appropriate to continue its development as a speaker identification technique. Second, this project is particularly cost-effective. It builds upon 12 years of research and the efforts of a large group of scientists. Further, nearly all equipment, the data-base and the feature vectors are available. Most important, however, is the fact that the value of the approach was recognized by NIJ administrators -- and funding initiated by the present contract. Hence, it would appear logical to continue this modest research program to its reasonable conclusion rather than dropping it now and not receiving reasonable compensation for the first grant. That is, since the success of this program has been greater than expected (during the present contract), it is hoped that NIJ will continue support for the Phase 2/3 portions of the research and that a effective operating system will result.

END