Research on Speaker Verification: Report of Working Group 53 (1971)

Pages
28

Size
8.5 x 10

ISBN
0309340780

Committee on Hearing, Bioacoustics, and Biomechanics; Office of Naval Research; National Research Council

Report of Working Group 53,
Research on Speaker Verification.
NAS-NRC Committee on Hearing, Bio-
acoustics, and Biomechanics, 1971.

NATIONAL ACADEMY OF SCIENCES - NATIONAL RESEARCH COUNCIL
COMMITTEE ON HEARING, BIOACOUSTICS, AND BIOMECHANICS

RESEARCH ON SPEAKER VERIFICATION

Report of Working Group 53
James L. Flanagan, Chairman

Frank R. Clarke
Franklin S. Cooper
Douglas L. Hogan
Arthur S. House
Horace O. Parrack
Irwin Pollack
Kenneth N. Stevens

Office of Naval Research
Contract No. N00014-67-A-0244-0021

March 1971

Order from
National Technical
Information Service,
Springfield, Va.
22161
Order No. AD 721-347

I. INTRODUCTION

Working Group 53 is charged to inform the Surgeons General or the Judge Advocates General of the three military services about current status of research on speaker verification.* Included are auditory, visual, and machine methods for speaker verification. Group 53 is to comment on potential applications of research findings and to recommend directions for continued research efforts.

Privileged and proprietary information has been made available to the working group. Its discussions and reports are therefore private.

II. SUMMARY AND RECOMMENDATIONS

2.1 Auditory verification. Auditory verification is dependent on a large number of poorly understood factors - the manner in which the speech signal is processed; the composition of the words, phrases, or sentences used; the training of the listener or listeners; and the methods used in making decisions. Experiments to date typically have examined the task of identifying a speaker from a group of several speakers (usually 5 to 10) known to the listener. Error scores in such experiments are found to be in the range of 5% to 19%. Being smaller for two- and three-syllable words than for monosyllables, error scores generally decrease with the duration of the speech sample. Most

* Verification and identification are often used with common meaning in this report. They are not identical problems, although their ingredients are similar. In verification an unknown claims an identity. His utterance is compared with a known reference for the claimed individual. A decision is made either to accept or reject the unknown as the claimed individual. In identification, or recognition, an unknown utterance is compared to a library of known reference patterns. A decision is made as to the most likely match from the known reference set. In prescribed instances "no match" is an allowable decision. The experiments discussed here deal both with identification and verification. In each case the experimental conditions specify the nature of the problem.

improvement is achieved over speech durations up to the order of one second, and little improvement is achieved for durations in excess of this.

At the present time methods that require the listener to identify voices rather than to identify attributes of a particular voice provide better recognition scores, other things being equal, and are to be preferred for application. Experimental evidence suggests that auditory verification is more reliable than visual identification from sound spectrograms, at least for subjects that are relatively untrained.

2.2 Visual verification. Visual verification has centered primarily around the use of sound spectrograms to compare voice patterns. It should be recognized, however, that other possibilities exist for visual display and could prove as useful as, or more useful than, sound spectrograms.

Most experiments on visual verification have been carried out in the laboratory with closed sets of speakers, typically 5 to 10, but one test used as many as 50. Experimental tasks have ranged from sorting spectrograms into groups (corresponding to individual speakers) to matching spectrograms of isolated words spoken by a known speaker to the same word in context spoken by an unknown. Error scores in these experiments have ranged from less than 1% to more than 60%. None of the experiments to date has scientifically addressed the question usually encountered in legal identification, namely, whether the spectrograms of a known speaker and an unknown speaker are produced by the same person or by different persons.*

* In the context of the previous definitions, this is strictly a verification problem - not identification.

2.3 Machine verification. Automatic machine methods of authentication and identification, although sophisticated in terms of computer techniques and decision theory, may prove primitive by comparison to the processes involved in human perception. It should not be overlooked, however, that a machine might be able to make use of factors that a human observer cannot assimilate.

In the machine identification studies that have been carried out, results depended strongly upon composition of the training and reference ensemble, the spoken materials, and the signal features used for identification. Typically spectral and other types of analyses have been made of utterances from about 10 to 20 speakers, and reference patterns have been formed. A new utterance from one of the speakers has been analyzed and compared with the reference patterns. The new utterance was identified with the speaker whose reference pattern provided the best match, in a prescribed sense. Error rates of about 10% are typical.

One recent experiment in machine processing has clearly distinguished the problems of verification and identification. The study deals in depth with automatic verification by computer. A population of 40 speakers was used, 32 as impostors and 8 as true. Non-linear time warping techniques, applied to formant, pitch and intensity data, led to error rates less than 2% for automatic operation. Present indications are that this level of performance is likely to be consistently better than that for either visual or auditory methods.

2.4 Directions for research. The verification problem of legal interest generally is the following situation. An evidential sample or pattern is on hand, obtained from a perpetrator; the acused claims not to be the perpetrator. Is the accused's voice pattern the same as the perpetrators?*

Few complete experiments--auditory, visual, or machine--have attacked this question directly. Further, most pertinent experimentation has been under laboratory conditions, usually with small, closed sets of speakers. It is impossible to extrapolate these results to general field conditions of large speaker populations, noisy environments, overt efforts at disguise, and communication and recording facilities of undetermined character. Before speaker recognition techniques can be established with any scientific validity and reliability these questions must be examined by carefully designed research.

Auditory verification depends directly on human perception of speech. The effect of acoustic factors (such as bandwidth, signal-to-noise ratio, non-linear distortion) on quality and intelligibility is still poorly understood. The influence on speaker recognition is even less understood. Research in this area is needed if auditory recognition is to be seriously considered as a routine method of verification.

In visual verification methods the sound spectrogram is likely to remain central in the near future. Any efforts to use visual inspection of spectrograms for verification should certainly quantify the variability of subjects' performance, first in small closed-set cooperative conditions and then in extention to large uncooperative populations. The utility of the sound spectrogram as an exclusion test, rather than as an identification means, ought also to be considered, along with its potential value as a supportive or investigative tool.

* The complement of this situation is also one of interest in defense and security activities. A reference pattern is on hand for speaker S_i. X calls and claims to be S_i. X produces all the requisite voice material to effect a comparison (i.e., a cooperative speaker). A decision must be made to accept or to reject X as S_i.

From the standpoint of defense and military applications, machine methods of speaker verification appear attractive. New techniques for analysis of speech signals to extract appropriate acoustic attributes and to give reliable, continuous measures of formants and fundamental frequency suggest that useful verification by computer may be possible. The important conditions of a limited and cooperative speaker population and of controlled and characterized communication facilities give impetus to this optimism. In terms of foreseeable needs and on-the-horizon technology, machine verification should produce more return per dollar of research investment than visual or auditory methods.

III. DETAILS OF STUDY

CHABA Working Group 53 was convened to advise the Surgeons General on the broad research problems relating to speaker verification, with particular reference to auditory, visual, and machine methods. The group interprets its charge to be to review existing research data, to comment on the potential application of research findings, and to recommend directions for continued research. One specific point of interest among the sponsors is the admissibility of speaker verification data, particularly when based on the sound spectrogram (or voiceprint), as legal evidence. The working group is composed of technical rather than legal experts. It cannot provide authoritative views on the legal standing of such data but will attempt to summarize the views of responsible agencies concerned with legal problems. The discussions and reports of Working Group 53 have profited by the availability of privileged and proprietary information and, therefore, are private and privileged.

At its initial meeting in Washington, D.C., 27 September 1967, Working Group 53 was divided into three subcommittees, as follows:

auditory methods - K.N. Stevens (chairman), F.R. Clarke, A.S. House; visual methods - J.L. Flanagan (chairman), F.S. Cooper, I. Pollack; machine methods - D.L. Hogan (chairman), J.L. Flanagan, H. Parrack. Each subcommittee was charged with the task of drafting preliminary reports for discussion and subsequent compilation into a complete draft document. The group met in Washington, D.C. on 3 May 1968 for discussion of the drafts prepared by the subcommittees. At that time various members were charged with the collection of additional information, the development of background material for the projected report, and with various editing responsibilities.

3.1 Aspects of the problem. One of the major problems faced by the working group in its discussions was the diversity of experimental methods and statistical treatments used in investigations dealing with speaker identification. In all the areas of interest--auditory, visual, and machine--the gap between laboratory and field experimentation is very wide and, in a real sense, not well characterized. Laboratory investigations tend to use a small set of subjects, and the response set is usually well known to the judges (whether human or otherwise). Thus the decision process usually involves the selection of that speaker of the known set of speakers whose vocal characteristics are most similar to those of the speaker to be identified. The question of how similar the two samples must be before both are identified as spoken by the same person does not arise since most decision rules simply choose the most similar pair after making all possible comparisons.

In most field applications the situation is quite different. Here one is typically given two samples of speech and simply asked if both were spoken by the same speaker. This is the problem of verification.

Something may be known of the population from which the individuals were chosen, but this population may be very large and speech samples are not available for all members of this population. Thus, the question becomes one of criterion: How similar must two samples be before they are judged to be spoken by the same person? If the criterion is made very stringent there should be few instances of incorrectly concluding that utterances by two different speakers were spoken by the same person. We refer to this type of error as a false identification. With a very strict criterion in all probability there would be many instances in which two utterances spoken by the same person were not as similar as required by the criterion thereby resulting in an incorrect rejection.

These two types of errors will covary as the criterion is changed, with the more lax criteria resulting in fewer incorrect rejections and more false identifications. The nature of the curve relating these two error rates will differ for different speaker recognition schemes and be dependent upon the parameters entering into the decision procedure. The proper evaluation of any speaker recognition procedure for field application should include specification of these error rates for specified populations. Such specification of error rates has not been accomplished to date for any speaker recognition procedure.

Of possible importance in some applications of speaker recognition procedures is the fact that laboratory experiments usually do not include uncooperative subjects, that is, subjects who wish strongly to remain anonymous. It is clear that the results of small, controlled laboratory studies cannot be extrapolated and interpreted for or against the application of a particular technique to large uncontrolled and possibly uncooperative speaker populations under nonlaboratory (field) conditions.

Furthermore, close examination of the results of laboratory studies has convinced the working group that specification of adequate training and its results has not been accomplished. It seems apparent that considerably more research is needed to establish procedures and techniques that will make experiments dealing with speaker recognition acceptable to large segments of the scientific community.

3.2 Comment on scientific and legal acceptability. Working Group 53's discussions (Washington, September 27, 1967) with representatives of several government agencies brought out one point perhaps not emphasized adequately among technical persons concerned with voice recognition. The point is the apparent difference between scientific and legal criteria for credibility or validity.

The technical person expects research findings to be described in terms of "problem; experimental procedure; results; interpretation" in a way that is unequivocal and repeatable by colleagues. Interpretations that attempt to generalize or extrapolate results to other conditions are expected to be demonstrable by experiment and, even then, are considered speculative until demonstrated. Well-defined, quantitative procedures--not depending upon art or inherent talent--are requisite to results reproducible by others.

By contrast, the legal view is more concerned with what an individual, on the totality of his experience and exposure, believes about a question. That is, his educated opinion is the matter of concern. The credence attaching to this opinion is a weight for the court to decide, again in the frame of the total situation and the individual's attested qualifications. Both defense and prosecution have recourse to expert testimony, and it apparently is not unusual to find experts holding directly opposed views, both purporting to be "beyond a reasonable doubt." In the legal context it is impossible to separate the expert witness from his criteria of judgment. In the scientific situation, this separation is imperative.

3.3 Methods of speaker verification; present status. The present study examines three means for speaker verification: auditory, visual, and machine. The status of research in each of these areas is outlined in the following sections.

3.31 Auditory methods. Experiments and data on the recognition of speakers by listeners exposed to the speakers' voices have generally followed two approaches. In one of these approaches, the procedure was to instruct listeners to rate the voices of speakers on a number of different scales or to assign to the voices a number of different attributes. In other words, the procedure was to listen to various spoken materials and use various scaling techniques in obtaining response data. The objectives of this approach were to determine the number and nature of the ways in which voices are perceived to differ from each other by a typical listener.

In the second approach, the task of the listener was simply to identify or to name a speaker (drawn from an inventory of known speakers) when exposed to a sample of his speech without explicitly stating the perceptual attributes of the voice. Experiments based on this approach required the listener to make his identification either (1) by directly naming a speaker that he knew or whose voice he had learned during a previous training period or (2) by using a matching-from-sample technique in which recorded samples of members of the inventory of voices were available to the listener when he was making his identification. In experiments of this type, the effect of various physical and linguistic characteristics of the speech signal on speaker identifiability was assessed. The objectives were, in part, to determine what aspects of the signal contributed to speaker identifiability, as well as to obtain some absolute measure of the ability of listeners to identify speakers.

Experiments on perceptual attributes of speakers: Voiers (1964) showed that listeners' ratings on semantic-differential scales carried significant information relating to their identity. Voiers obtained listeners' ratings on each of 49 bipolar scales (e.g., clear-hazy, rough-smooth, rumbling-shining, fast-slow) for speech samples of sixteen male speakers. Analysis of variance showed that differences in mean rating assigned to the 16 speakers were significant (p 0.01) for 45 of the 49 scales. Factor analysis was performed with the result that four orthogonal factors were found to account for 88% of the scale variance. Voiers felt that the terms clarity, roughness, magnitude, and animation were descriptive of the four factors. In conclusion Voiers stated:

> "The general inability of listeners to augment their characterizations of voices with additional adjectives of their own choosing constitutes one basis for confidence in the comprehensiveness of the rating form. Further support on this point is provided by the failure of subsequent attempts to discover items that tap additional dimensions of listener response to voices. No item has been found for which the speaker component of rating variance involved any dimensions other than clarity, roughness, magnitude, or animation."

Holmgren (1967) extended this work using three scales to represent each of Voiers' four factors and by obtaining various physical measures that could be compared to the rating data. Twenty listeners rated 10 speakers' voice characteristics. An analysis of variance of each of the 12 bipolar scales revealed that the speaker's main effect was significant in each case (p 0.001). Over 90% of the variance in the semantic-differential data could be accounted for by only two orthogonal factors. When both the scale values and the physical voice measurements (rate of speaking as well as mean variance of amplitude of unvoiced sounds, amplitude of the voiced sounds, and fundamental frequency) were combined in a single factor analysis, there appeared to be three factors represented by both judged and physical voice measures and two factors represented primarily by physical voice measures. Thus Holmgren concludes that some of the physical measures "represent cues to which either (1) the listeners were unable to respond judgmentally or (2) appropriate items were not on the form thus restricting response availability."

This work was further extended by Clarke and Becker (in Clarke, Becker, and Nixon, 1966), who attempted to determine the degree to which information contained in rating responses could be used in actually discriminating among speakers. Speaker discrimination scores obtained using rating-scale data as an imput to a decision algorithm were compared with scores obtained using physical measures on the speech waveform as input to the same decision algorithm and with scores obtained by listeners in direct aural test. This study employed 16 male speakers, 16 listeners, and the same 12 semantic-differential scales employed by Holmgren. A "same-different" test format resulted in 68% correct decisions based on rating data, 83% correct decisions based on power-spectrum data, and 90% correct decisions obtained in direct aural test with listeners.

The three studies reported above show that the listener can rate voices on semantic-differential scales in such a way as to result in significant and reliable differences among speakers. Factor analyses of resultant scale values suggest that three to four "perceptual dimensions" are adequate to account for ratings on semantic-differential scales. While naming such dimensions is rather arbitrary, Voiers' labels (clarity, roughness, mganitude, and animation) are suggestive. While semantic-differential ratings contain significant information pertinent to discriminating among speakers, scores obtained by human listeners in direct aural testing far exceed those scores obtained using scale values to discriminate among speakers. This latter finding would suggest that there are perceptual attributes of speakers that are not adequately reflected in semantic-differential ratings.

Experiments on absolute identification of speakers: A large number of different kinds of experiments on speaker identification have been reported in the literature. Some of the major findings of these experiments are summarized in the following paragraphs.

i. When a listener is exposed to a monosyllabic utterance and given the task of identifying the originating speaker from a group of 8 or 10 voices that are known to him, he obtains an average recognition score of about 84% (Bricker and Pruzansky, 1966; Pollack, Pickett, and Sumby, 1954). When aural identification of bisyllabic words is carried out with an ensemble of 8 speakers using a matching-from-sample technique, an identification score of about 90% is obtained (Carbonell, Grignetti, Stevens, Williams, and Woods, 1965).

ii. Several investigators (Kryter, Williams, and Green, 1962; Williamson, 1961) have shown that good aural recognition scores (90% or more) are obtained when the task is to identify two sequential speech samples as being spoken by the same speaker or by a different speaker. For example, a 93% score has been obtained in such a test using monosyllabic words. In experiments such as these, scores for sequential presentation of stimuli are much higher than scores for simultaneous presentation of material by two speakers.

iii. When the experimental task is to identify a speaker as one of n previously heard speakers, after a relatively short period of training, there is a fairly sharp decrease in score as n increases from six to eight, but the scores for n of 4 and n of 6 are similar (Williams, 1964).

iv. In speaker recognition experiments there is an appreciable difference in scores for a 2-3 syllable sample of speech compared with a one-syllable sample (Williams, 1964; Carbonell, Grignetti, Stevens, Williams, and Woods, 1965). When the recognition score is expressed as percent of correct judgments, scores are fairly steep functions of the duration of the speech sample for durations up to 1.2 sec, but the increase in score above 1.2 sec is rather small (Pollack, Pickett, and Sumby, 1954). Bricker and Pruzansky (1966) showed that the improvement in identification with increased duration seems to be due to the increased sample of the speaker's repertoire.

v. Speaker recognition tests have shown that some parts of the speech frequency range are more important than others in their contributions of identifying cues (Peters, 1954; Compton, 1963). For example, when speech is processed by octave-band filters, the best recognition scores are obtained with the band-pass 1200-2400 Hz condition. Low-pass filtering at 3000 Hz or high-pass filtering at 500 Hz gives little deterioration relative to wide-band speech as far as speaker recognition is concerned.

vi. Noise affects the ability of a listener to recognize a speaker's voice, but there are conflicting data on how much noise gives a substantial decrease in score. For white noise, the decrease in performance seems to occur for signal-to-noise ratios in the range of -4 to +8 dB (Peters, 1954).

vii. Shifting the formant frequencies in connected speech gives sharp drops in speaker recognition scores (Shearme and Holmes, 1959). Removal of the effects of larynx frequency by making the speech monotone also gives deterioration but not as much as shifting F_1 by 100 Hz, F_2 by 300 Hz, and F_3 by 300 Hz.

viii. A speaker can be more readily identified by listeners when the sample of his speech contains front vowels than when it contains back vowels, presumably because front vowels are richer in high-frequency energy (Carbonell, Grignetti, Stevens, Williams, and Woods, 1965).

ix. There is a great variance among listeners in their ability to identify voices (Williams, 1964; Carbonell, Grignetti, Stevens, Williams, and Woods, 1965). Also, there is great variance in average performance of the listeners depending on the group of speakers to be identified (Stuntz, 1963).

3.32 Visual methods. Visual techniques for speaker identification could encompass a variety of signal portrayals, ranging from a simple oscillogram to displays as esoteric as a focal-tract shape or a cochlear transform. In practice, however, most interest has centered on the amplitude-frequency-time display, known as the sound spectrogram.

The technique for making sound spectrograms was developed more than 20 years ago by a Bell Telephone Laboratories group under the direction of R.K. Potter. The method was applied to fundamental studies of the acoustics of speech and to the problem of deaf communication and training, as reported in Visible Speech (Potter, Kopp, and Green, 1947).

During World War II, C.H.G. Gray and G.A. Kopp suggested that the same techniques could be applied to problems of signal intelligence, particularly for speaker identification, but very little practical work was accomplished on the problem.

Active interest in speaker identification by means of sound spectrograms, sometimes referred to as voiceprints, was revived at the Bell Telephone Laboratories about 1961 by L.G. Kersta. This work was motivated by the desire to assess the ability of human judges to identify speakers. The expectation was that quantitative data would evolve that might be useful for development of machine methods of identifying speakers. Early reports of the research in these limited studies quickly came to the notice of various law enforcement agencies. They saw in it a potential aid to identification that might supplement other standard means of identification now in use, such as systems of describing physiognomy, handwriting, fingerprint patterns, and vocal characteristics.

The scientific status of research based on voiceprint techniques is not easily determined and is, indeed, a matter of dispute. The leading exponent of visual identification by means of sound spectrograms, L.G. Kersta, is no longer associated with the Bell Telephone Laboratories but is actively engaged in marketing sound spectrographic equipment and services, through a company he formed for that purpose. To complicate the issue further, Kersta's firm also offers training courses in voiceprint identification for law-enforcement agencies, as well as courtroom services of testimony concerning identifications made by such techniques. The basis for Kersta's claims for the technique, as well as for his own expertise, are not well documented in the scientific press; most of the available citations are to oral presentations at various meetings, to newspaper releases and stories, and to materials made availabe by Voiceprint Laboratories (now a division of Farrington Manufacturing Company). The dearth of literature was verified by a letter addressed to L.G. Kersta by M.A. Whitcomb for this working group. Kersta's response, though cordial, provided only materials of the type mentioned above and descriptions of the firm's products and services.

In general, the known research suffers from the limitations, mentioned above, of extrapolating from the laboratory to field investigations. Moreover, the small number of controlled laboratory studies that have been carried out have reported widely divergent results. An early report (Kersta, 1962) claimed high identifiability of words uttered by up to 12 speakers in a series of experiments; errors in identification of about 1% to 2% were claimed, though the meaning of these percentages is not entirely clear. A later, more detailed experiment (motivated by skepticism of the statistical treatment used by Kersta) obtained substantially poorer results, reporting about 78% identification of isolated words uttered by five speakers and only about 37% identification scores for words excerpted from context (Young and Campbell, 1967). The importance of error-computation techniques in studies reported by Kersta has been argued on qualitative (and sometimes emotional) grounds, particularly by Ladefoged and his associates (Ladefoged and Vanderslice, 1967). These and related studies have recently been examined and compared in depth by Bolt, Cooper, David, Denes, Picket, and Stevens (1970). This paper is recommended for its detail on visual verification and provides valuable points of view.*

* Since two of the members of Working Group 53 coauthored the Bolt, et al. paper, the reader will find some of the material of this report in that paper.

One major reason why visual identification has not demonstrated a high degree of reliability is the fact that an individual speaker shows considerable variability when repeating a given utterance on different occasions and under different conditions of emotion and stress. For example, several studies have shown that the acoustic characteristics of a speaker's voice undergo appreciable changes when he is working under stress or when he is excited, angry, or sad. The kinds of changes that occur under these conditions vary considerably from one speaker to another.

As mentioned above, visual methods of speaker identification have concentrated on the sound spectrogram as a display. There are many other means for displaying the information contained in the speech waveform, some of which may prove to be as useful as the voiceprint. Typical candidates for visual presentation include formant traces, intonation contours, voicing patterns, vocal-tract shapes, and cochlear patterns. Very little basic research has yet been done on these alternatives.

The difficulty in assessing the scientific value of voiceprint identification techniques may be reduced in the future. A relatively large-scale evaluational and experimental study is presently in progress. The study is supported by the U.S. Department of Justice by contract with the Michigan State Department of Police. Subcontracts for various portions of the study have been awarded to the Department of Police Administration, Michigan State University; the Department of Speech, Michigan State University (O. Tosi and H. Oyer*); and Stanford Research Institute (K. Kryter and F.R. Clarke).

3.33 Machine methods. Investigations dealing with machine methods can be grouped primarily according to their specific aims. In general, fundamental investigations have concentrated on describing the speaker-to-speaker variability of various physical parameters of the speech signal. Speaker authentication studies have sought means of verifying the identity of a previously known and cooperative speaker, while studies of speaker identification have sought means to determine whether or not a given (generally uncooperative) speaker is one of a known set of speakers.

* Since this document was prepared, an initial report has been made from the Michigan State study (O. Tosi et al "An experiment on voice identification by visual inspection of spectrograms." Contract No. NI-70-004, U.S. Department of Justice; talk presented to Acoustical Society of America, 80th Meeting, Houston, Texas, November 1970).

The Michigan State study embraces two years of fairly comprehensive testing of visual matching in an identification task; i.e., given an unknown, select the best match from an ensemble of knowns. Known-speaker populations included 10, 20, or 40 young males. Unknown speakers included 250 young males. The experiment included "closed trial" conditions ((in which the judges knew that the unknown speaker was one of the reference (known) library)) and "open trial" conditions (where the judges did not know whether the unknown was among the known set). Recordings of "non-contemporary" samples of a known were made one month later than his reference samples. Judges were given one month of training in spectrogram recognition.

Error rates for the recognition task varied from 1% to 30%. Lowest error rates occurred for matching contemporary spectrograms in small, closed sets using words spoken in isolation. Largest error rates occurred for identification of non-contemporary spectrograms, in large, open sets using clue words excerpted from random context. This range of error rates vs test conditions appears reasonably consistent with other visual experiments mentioned in this report.

An early example (Smith, 1962) of a fundamental study of variability was accomplished at the Lincoln Laboratory. Time-sampled spectral data from vowel portions of speech from 10 speakers were manipulated with standard techniques using discriminant functions, and about 85% accuracy in identification was reported. When a discriminant analysis of the four variables with the highest information content was made, the four variables were all in the frequency range 3 kHz to 8 kHz and the finding was interpreted as meaning the information carrying variables were more closely correlated with speaker characteristics than with phonetic characteristics.* Recent work by Harris tends to support this work and suggests also that the temporal character of spectral change may help to characterize a given speaker.

Considerable work in the area of speaker authentication has been done by research groups at IBM and RCA. The work of the IBM System Development Division was oriented toward the problem of business privacy, as in the case of operating a computerized data store with access by authorized speakers, while the RCA Defense Electronic Products Division studied means of automatically authenticating voices in military communication contexts. Both groups adapted techniques previously devised for automatic speech recognition. The IBM studies used adaptive linear threshold elements and a trainable decision procedure (Li, Dammann, and Chapman, 1966), and the RCA investigations used a set of feature-extraction techniques (Martin, Nelson, and Fadell, 1964). In general, the IBM results, based on 50 speakers saying a single one-second phrase, showed about 90% identification. The RCA results were about the same for

* This interpretation runs contrary to Peters (1954) and Compton (1963) for auditory processing.

a smaller group of speakers. In these studies the groups of speakers included so-called imposters, speakers who were not on the approved list but did not make a conscious attempt to confuse the system.

Two recent experiments at Bell Laboratories (Doddington, 1970) and at IBM (Das, Mohn, and Saleeby, 1970) have treated the machine verification problem. In the Bell work unknown speakers were permitted to claim an identity from a group of true speakers or "customers." The computer, using a non-linear time-warp applied to formant, pitch, and intensity analyses, made a decision to accept or reject the unknown. For a group of 40 speakers, 32 imposters and 8 true, average error rates of 1.5% were achieved. A response of "no decision" was not permitted. The IBM work used male speakers and wide-band, noise-free input. The experiments used a total of about 7,000 utterances of the phrase "Check available terminals" from 118 speakers. An average misclassification rate of 1% with a "no decision" rate of 10% was obtained.

No work is known to have been done on the machine processing of the speech of high-quality mimics.*

In summary, most of the experimental work on machine methods has dealt with relatively small populations and has achieved about 90%-95% correct identification and as high as 98% verification. While this level of accuracy would make machine methods useful as investigative tools, none of the current techniques have been extended to large populations or have achieved completely infallible authentication of speakers. It

* Research on this question is in progress (R.C. Lummis, "Real Time Techniques for Speaker Verification by Computer," to be presented at the 81st Meeting of the Acoustical Society of America, Washington, D.C., April 20-23, 1971. A.E. Rosenberg, "Listener Performance in a Speaker Verification Task," to be presented at the 81st Meeting of the Acoustical Society of America, Washington, D.C., April 20-23, 1971).

is clear, however, that there are many physical characteristics of the speech signal that have not yet received adequate attention as potential indicators of a speaker's identity - even characteristics that a human listener may not utilize. A speaker's prosodic features and his dialectal pecularities are well-known as subjective clues to his identity, but the extraction and evaluation of these factors or of the physical components of the speech signal that underlie these abstract characteristics are not well understood. These questions represent valid and fruitful areas for new research.

REFERENCES

Bolt, R.H., Cooper, F.S., David, E.E., Jr., Denes, P.G., Picket, J.M., and Stevens, K.N. Speaker identification by sound spectrograms: A scientist's view of its reliability for legal purposes. J. Acoust. Soc. Amer., 47, 1970, 597-612.

Bricker, P.D. and Pruzansky, S. Effects of stimulus content and duration on talker identification. J. Acoust. Soc, Amer., 40, 1966, 1441-1449.

Carbonell, J.R., Grignetti, M.C., Stevens, K.N., Williams, C.E., and Wood, B. Speaker authentication techniques. Final Report Contract No. DA-28-043-AMC-00116 (E) with U.S. Army Electronics Laboratories, Fort Monmouth, New Jersey. Bolt Beranek and Newman, Inc., Cambridge, Mass., July 1965; AD 468993.

Clarke, F.R., Becker, R.W., and Nixon, J.C. Characteristics that determine speaker recognition. Stanford Research Institute, Menlo Park, Calif., December 1966; ESD-TR-66-636; AD 646135.

Compton, A.J. Effects of filtering and vocal duration upon the identification of speakers, aurally. J. Acoust. Soc. Amer., 35, 1963, 1748-1752.

Das, S.K., Mohn, W.S., and Saleeby, S.L., Speaker verification experiments, J. Acoust. Soc. Amer., (A), 1970.

Doddington, G., A method for speaker verification, Ph.D. Thesis, Dept. of Elec. Eng., U. of Wisconsin, June 1970. Also, J. Acoust. Soc. Amer., (A), 1970.

Harris, C.M., Informal report, under Contract Nonr-4259(15), 1967.

Harris, C.M., Interim progress report, Contract Nonr-4259(15), April 1968.

Holmgren, G.L. Physical and psychological correlates of speaker recognition. J. Speech Hear. Research, 10, 1967, 57-66.

Kersta, L.G. Voiceprint identification. Nature, 196, 1962, 1253-1257.

Kryter, K.D., Williams, C.E., and Green, D.M. Talker identification test. Unpublished memorandum, Bolt Beranek and Newman, Inc., Cambridge, Mass., 1962.

Li, K.P., Dammann, J.E., and Chapman, W.D. Experimental studies in speaker verification, using an adaptive system. J. Acoust. Soc. Amer., 40, 1966, 966-978.

Martin, T.B., Nelson, A.L., and Fadell, H.J. Speech recognition by feature-abstraction techniques. Avionics Laboratory, Research and Technical Div., AFSC, U.S. Air Force, August 1964; AL-TDR-64-176.

Peters, R.W. Studies in extra-messages: Listener identification of speakers; voices under conditions of certain restrictions imposed upon the voice signal. Project Report No. NM 001 064.01.30, U.S. Naval School of Aviation Medicine, Pensacola, Florida, October 1954.

Pollack, J., Pickett, J.M., and Sumby, W.H. On the identification of speakers by voice. J. Acoust. Soc. Amer., 26, 1954, 403-406.

Potter, R., Kopp, G., and Green, H. Viable Speech, D. Van Nostrand, New York, N.Y., 1947.

Shearne, J.N., and Holmes, J.N. An experiment concerning the recognition of voices. Lanq. Speech, 2, 1959, 123-131.

Stuntz, S.E. Speech intelligibility and talker recognition tests of Air Force communications systems. Technical Documentary Report ESD-TDR-63-224, ESD, AFSC, USAF, L.G. Hanscom Field, Beford, Mass., February 1963.

Vanderslice, R., and Ladefoged, P. "Voiceprint" mystique. J. Acoust. Soc. Amer., 42, (A), 1967; also, Working Papers in Phonetics (U.C.L.A.), 7, November 1967, 126-142.

Voiers, W.D. Perceptual bases of speaker identity. J. Acoust. Soc. Amer., 36, 1964, 1065-1073.

Williams, C.E. The effects of selected factors on the aural identification of speakers. Section III: Methods for psychoacoustic evaluation of speech communication systems, Technical Documentary Report ESD-TDR-65-153, ESD, AFSC, USAF, L.G. Hanscom Field, Bedford, Mass., December 1964.

Williamson, J.A. An investigation of several factors which affect the ability to identify voices as same or different. Unpublished dissertation for diploma in phonetics, University of Edinburgh, 1961.

Young, M.A., and Campbell, R.A. Effects of context on talker identification. J. Acoust. Soc. Amer., 42, 1967, 1250-1254.

Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)*
National Academy of Sciences - National Research Council Committee on Hearing, Bioacoustics, and Biomechanics

2a. REPORT SECURITY CLASSIFICATION
None

2b. GROUP
None

3. REPORT TITLE
Research on Speaker Verification

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

5. AUTHOR(S) *(Last name, first name, initial)*
Dr. James Flanagan, editor

6. REPORT DATE
March 1971

7a. TOTAL NO. OF PAGES
25

7b. NO. OF REFS
24

8a. CONTRACT OR GRANT NO.
N00014-67-A-0244-0021

b. PROJECT NO.

c.

d.

9a. ORIGINATOR'S REPORT NUMBER(S)

9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*

10. AVAILABILITY/LIMITATION NOTICES
Qualified Requesters may obtain copies of this report from DDC.

11. SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY
Office of Naval Research
Code 454
Arlington, Virginia 22217

13. ABSTRACT

This report advised the Judge Advocates General of the military services concern-the relative merits of the three main methods of speaker verification and gives some indication of their absolute value as legal evidence. It then recommends concerning which method has the most promise and thus merits research funding more than the others.

14. KEY WORDS

speaker recognition
speaker verification
sound spectrograms
voiceprint
legal evidence
scientific evidence

DD FORM 1 JAN 64 1473

Security Classification