Reprinted from THE JOURNAL OF THE ACOUSTICAL SOCIETY OF AMERICA, Vol. 47, No. 2, (Part 2), 597–612, February 1970

Speaker Identification by Speech Spectrograms: A Scientists' View of its Reliability for Legal Purposes

RICHARD H. BOLT

Bolt Beranek and Newman Incorporated, 50 Moulton Street, Cambridge, Massachusetts 02138

FRANKLIN S. COOPER

Haskins Laboratories, 305 East 43rd Street, New York, New York 10017

EDWARD E. DAVID, JR.

Bell Telephone Laboratories, Incorporated, Murray Hill, New Jersey 07974

PETER B. DENES

Bell Telephone Laboratories, Incorporated, Murray Hill, New Jersey 07974

JAMES M. PICKETT

Gallaudet College, Washington, D. C. 20002

KENNETH N. STEVENS

Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

Can you reliably identify a person by examining the spectrographic patterns of his speech sounds? This is a scientific problem of social consequence because of the interest of the courts in this question. The Technical Committee on Speech Communication of the Acoustical Society of America has asked some of its members to review the matter from a scientific point of view. The topics they considered included the nature of speech information as it relates to speaker identification, a comparison of voice patterns and fingerprint patterns, experimental evidence on voice identification, and requirements for validation of such identification methods. Findings and conclusions are reported; supporting information is given in appendixes.

INTRODUCTION

The sound spectrograph is an instrument that finds widespread use in current research on speech sounds. It portrays, in graphical form, the time variations of the short-term spectrum of the speech wave.[1] Examples of such speech spectrograms are shown in Fig. 1 for four instances of the word "science." In each spectrogram, the horizontal dimension is time, the vertical dimension represents frequency, and the darkness represents intensity on a compressed scale. This representation of the sound patterns of speech has proved to be extremely powerful in research on the phonetic aspects of speech because the spectrogram gives valuable information about speech articulation. In the examples of Fig. 1, the middle portions of the patterns show effects of the articulations corresponding to the vowels of "science." The initial and final portions of each spectrogram show sudden changes in the frequency pattern where consonants and vowels join.

[1] W. Koenig, H. K. Dunn, and L. Y. Lacey, J. Acoust. Soc. Amer. 18, 19 (1946).

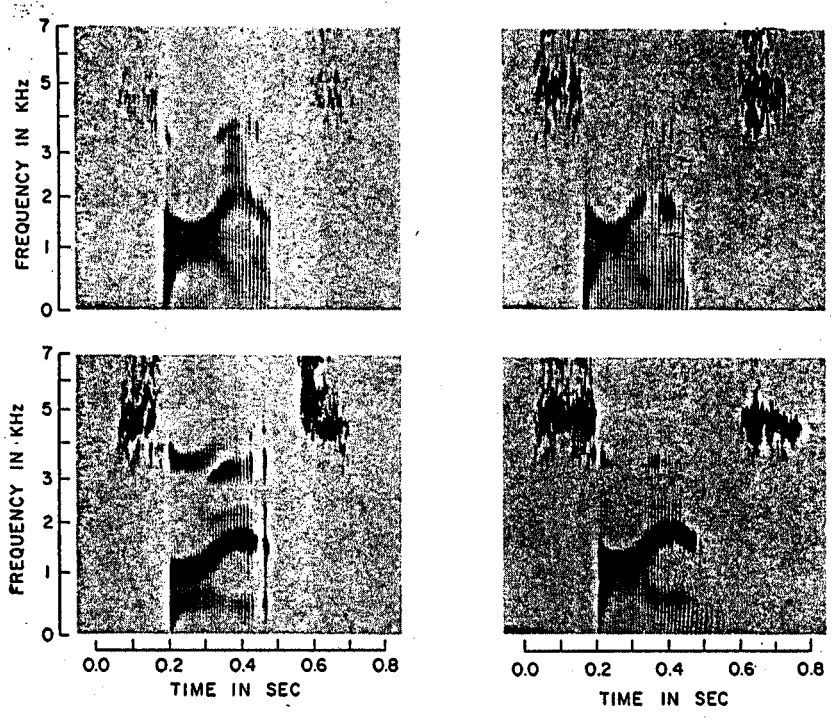

FIG. 1. Four spectrograms of the spoken word "science." The vertical scale represents frequency, the horizontal dimension is time, and darkness represents intensity on a compressed scale. Three of the spectrograms are from three different speakers and the remaining spectrogram is a repetition of the word by one of the speakers (see text). The spectrograms were made on a Voiceprint Laboratories sound spectrograph.

When two persons speak the same word, their articulation is similar but not identical; therefore, spectrograms of these words will be similar but not identical. There are also similarities and differences even when the same speaker repeats the same word. These facts are apparent in the spectrograms of Fig. 1. The two spectrograms at the top were made by the same speaker on two different occasions; the two spectrograms at the bottom were made by two other speakers.

Speech scientists have found spectrograms very useful in studying how people pronounce different words. Can spectrograms also be applied to distinguishing one person from another? In several recent court hearings, evidence has been presented both for and against the use of speech spectrograms, or "voiceprints," for personal identification. Scientists in speech research have been concerned, for reasons of social importance and scientific credibility, about such use of speech spectrograms; the Technical Committee on Speech Communication of the Acoustical Society of America asked six members of the Society (the authors of this paper) to study and report on this issue.[2] In considering this problem, we asked questions such as the following: When two voice spectrograms look alike, do the similarities mean "same speaker" or merely "same word spoken"? Are the irrelevant similarities likely to mislead a lay jury in assessing conflicting testimony from opposing experts? How permanent are voice patterns? How distinctive are they for the individual? Can they be successfully disguised or faked?

Whatever the future may hold for voice printing as a method of identification, expert witnesses at the present time do not agree as to its reliability, and various courts of law have ruled both for and against the admission of such evidence.[3] These differences of opinion are, however, only the surface reflections of deep-lying difficulties, inherent in the nature of spoken language, that serve to make voice identification equivocal for the expert and confusing to the layman.

It is against this background that we have undertaken to point up the difficulties inherent in voice identification, to review and assess the relevant scientific knowledge available today, and to examine the problem of scientific validation for the use of voiceprint identification as legal evidence.[4]

I. THE NATURE OF SPEECH INFORMATION AND VOICE IDENTIFICATION

The aim of speech is communication. For this purpose, speakers of a given language use a common code and a common set of speech sounds. Thus, the same message produced by different speakers uses basically the same sequence of sounds; when a person speaks a

[2] The views given here are those of the authors as individuals. Additional background about this report will be found in J. Acoust. Soc. Amer. 46, 867–868 (1969).

[3] State v. Cary, 49 N. J. 343 (1967); 99 N. J. Super. 323 (L. D 1968); N. J. Supreme Court, Docket C-207; People v. King, Calif. App. 2nd Dist., 2nd Crim. 13588; United States v. Wright, 17 U.S.C.M.A. 183 (1967).

[4] Technical details in support of the discussions are contained in appendixes. A detailed scientific review of voice identification has been prepared by M. Hecker, "Methods for Measuring Speaker Recognition," Stanford Res. Inst., Menlo Park, Calif. Apr. 1969.

word or phrase, he tries to produce sound patterns like those of other speakers of his dialect. In fact, however, only certain aspects of the sounds are the same when two speakers produce the same word or when one speaker says the same word on different occasions.

There are several reasons why some aspects of the sound pattern of a word are different on different occasions. For different speakers, the vocal anatomy may be different. Regardless of the speaker, some aspects of the sounds are nonessential in the sense that they are not used to identity words, and speakers are free to produce them in various ways. Different speakers may develop characteristically different habits in using these nonessential features, or a single speaker may show considerable variation in their use from one utterance to another. Thus, the speech sounds carry several submessages, including information about the speaker's identity, his mood, and his manner of speaking, as well as the words he says. At present, we do not have a clear understanding of which sound features are likely to be invariant for a given speaker, and which are likely to show variation from one speaker to another.[5]

A further complication is that the sound features do not fall neatly into separate sets that refer to the various submessages carried in speech. All these submessages are merged into a complex sound stream; moreover, all of them can affect all the sound features, so that there is no simple, obvious relation between messages and features.[6]

Yet, recovering one of these submessages is the essence of speaker identification: the task is to tease out from the sound patterns those features that correspond to the talker's vocal anatomy and his habits of forming speech sounds, since these might characterize him as a speaker. This is usually attempted by comparing different utterances of the same word or phrase, one from a known speaker, and interpreting the similarities and differences. There will be many similarities because the same words were used; there will also be differences that may be due either to a difference in speakers, or to the free variations of a single speaker.

The correct assignment of the differences, given all these complexities, is a difficult matter. Yet, we know that almost everyone can identify some voices just by listening to them. We know also, from controlled experiments, that identification by ear alone is not highly reliable.[7]

A newer method of voice identification uses visual comparison of the graphic patterns resulting from a gross acoustic analysis using the sound spectrograph. Not all details of the acoustic patterns are presented in this graphic display; moreover, the display is designed to emphasize those features that characterize the words of the spoken message. Speech-sound spectrograms of this type are the primary material used forensically for voice identification. The identification is done, not by the spectrograph, but by means of visual comparison of the spectrograms and by subjective judgments about the identity of the speakers represented.[8]

Could a better instrument be developed? One possibility would be a device with a display emphasizing those sound features that are most dependent on the speaker. The patterns could then be judged with greater confidence by human experts. We do not yet know how to design such an instrument, primarily because of the inherent complexity of speech sounds. We are even farther from having a fully objective procedure by which the features that characterize an individual speaker could be extracted and evaluated automatically.[9]

II. VOICE PATTERNS AND FINGERPRINT PATTERNS

How similar is voice identification by spectrogram to fingerprint identification? The differences between them seem to exceed the similarities, as the following comparative summary shows:

Fingerprints show directly the physical patterns of the fingers producing them and these patterns are readily discernible. Spectrographic patterns and the sound waves that they represent are not, however, related so simply and directly to vocal anatomy; moreover, the spectrogram is not the primary evidence, but only a graphic means for examining the sounds that a speaker makes.

In fingerprint identification, the gross types of ridge patterns, such as loops and whorls, are used for classification and indexing; these types are determined mainly by heredity and thus have only limited power in differentiating persons. The minute details of the ridges are then compared for final identification and all points of similarity strongly imply a match, while any point of dissimilarity strongly implies a mismatch. In comparing voice patterns, we are not able to interpret similarities and differences in such simple ways.

The fingerprint features that are ultimately used for identification are the most minute details of the skin ridge patterns such as bifurcations, terminations, and

[5] G. Fant, *Acoustic Theory of Speech Production* (Mouton and Company, 's-Gravenhage, 1960); K. N. Stevens and A. S. House, J. Speech Hearing Res. 4, 303 (1961).

[6] A. M. Liberman, F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy, Amer. Ann. Deaf 113, 127 (1968); Psych. Rev. 74, 431–461, 1967.

[7] F. R. Clarke, R. W. Becker, and J. C. Nixon, "Characteristics that Determine Speaker Recognition," Rep. ESD-TR-66-636, Decision Sciences Laboratory, Hanscom Field, Bedford, Mass., Dec. 1966 (report under contract to Stanford Res. Inst., Menlo Park, Calif.); W. D. Voiers, J. Acoust. Soc. Amer. 36, 1065 (1964); C. E. Williams, "The Effects of Selected Factors on the Aural Identification of Speakers," Sec. 111 in "Methods for Psychoacoustic Evaluation of Speech Communication Systems," Rep. ESD-tdr-65-153, Electronic Syst. Div., Air Force Syst. Command, Hanscom Field, Mass., 1964.

[8] L. G. Kersta, Nature 196, 1253 (1962); H. Mennen, H. Tillman, and G. Ungeheuer, "Entwicklung eines Systems von Beschreibungsmerkmalen für Kontursonagramme zum Zwecke der Sprecheridentifikation," T-888-L-203, Institut für Phonetik und Kommunikationsforschung, Universität Bonn, Nov. 1968.

[9] S. Pruzansky, J. Acoust. Soc. Amer. 35, 354 (1963).

interruptions. These details are determined mainly by random processes in prenatal skin development. There are a sizable number of these minute anatomical features on each finger. There are an enormous number of possible combinations of these features and it is known that their patterns remain unchanged throughout life.[10] Comparable voice features for identification, if they exist, have not been established; moreover, changes with growth and environmental influences could be expected.

Whereas fingerprint patterns cannot easily be faked or disguised, a speaker can learn to alter his voice and imitate, with some success, the speech of other persons.

Variations found in fingerprint patterns do not consist of changes in patterns from one type to another, but rather in expansions (with growth), obliterations (of some features), smudges, or incompleteness. Spectrographic patterns are affected in a more fundamental way by the distortions of frequency, energy, and time that are commonly encountered in the transmission, recording, and analysis of sound. The very dimensions of the pattern are those that are changed by such sound distortions.

In view of basic differences between fingerprints and voice patterns, and the inherent complexity of spoken language, we doubt that the reliability of voice identification can ever match that of fingerprint identification.

III. EXPERIMENTAL EVIDENCE ON VOICE IDENTIFICATION

Both objective and subjective methods have been used to try to identify voices. In objective methods, a piece of equipment makes all the decisions. Subjective methods may also involve equipment, such as a sound spectrograph, to display the acoustic information, but the final decision—the judgment—is made by a man.

Objective methods of voice identification have used automatic pattern matching, applied to voice patterns. In one study, average spectral patterns were obtained for each of 10 talkers and stored in a computer. To make identifications, a new pattern from each of the talkers was compared with each of the stored patterns to find the one most similar; identification errors were about 10%.[11,12]

Subjective experiments using speech spectrograms have been of two types: (1) sorting experiments, in which the observer sorts a set of spectrograms of a test word into individual talker categories, and (2) matching experiments, in which the observer identifies spectrograms of single talkers by matching them against spectrograms in a catalog of talkers, all speaking the same word or set of words.

In the sorting experiments, the observers knew how many talkers there were and how many examples were taken from each talker. In these experiments, test sets of 5–12 talkers were drawn at random from a pool of 123 male talkers selected to be homogeneous in regional accent[12,13] In a test, there were four examples of each test word from each talker. With 12 talkers, for example, 48 spectrograms were given to the observer and his task was to sort them into 12 categories corresponding to the individual talkers. Trained observers were used. In one such experiment,[13] which used test sets of 5, 9, or 12 talkers, the average error rates, pooled over observers, ranged from 0.35% for 5 talkers in the set to 1% for 12 talkers in the set. In another sorting experiment,[14] the observers were nine law-enforcement officers, of whom seven were fingerprint experts; all were first trained in voice identification from spectrograms; test sets of 12 talkers were used; the observers' error rates ranged from 0 to 3.48%, with a median of 0.42%.

The matching experiments reported to date have employed test sets of talkers ranging in size from 5 to 50. In one matching experiment,[13] nine talkers were used in the catalog; the catalog contained two examples of a test word as spoken by each talker and the observer's task was to match a third example of the word spoken by one of the nine talkers; the average error rate was 1%; the range of error rates, over the 10 different test words employed, was 0 to 3%. In another matching experiment[14] with 50 talkers drawn from the pool of talkers mentioned above, the catalog of spectrograms consisted of two examples of each of five words spoken in context by each talker. Nine trained observers matched new sets of the same five words, each set spoken by a talker who was one of the 50 talkers in the catalog. The error rate for observers working individually ranged from 0 to 11.1%, with a median of 5.7%. The error rate for observers working together in pairs ranged from 3.2% to 14.3%, with a median of 7.7%.

In another matching experiment[15] using a set of five talkers and trained observers, the average error rate was 21.6% for words spoken in isolation. When the words were spoken in fluent context and matched against a catalog taken from context, the error of talker identification was 62.7%.

In still another matching experiment,[16] the results obtained from listening only were compared with the results obtained solely by visual examination of spectrograms, using the same set of utterances for the two

[10] H. Cummins and C. Midlo, *Fingerprints, Palms and Soles: An Introduction to Dermatoglyphics* (Dover Publications, New York, 1961); Francis Galton, *Finger Prints* (Macmillan and Company, Ltd., London, 1892; facsimile reprint, Da Capo Press, New York, 1965, with historical introduction by H. Cummins).

[11] S. Pruzansky and M. V. Mathews, J. Acoust. Soc. Amer. 36, 2041 (1964).

[12] L. G. Kersta, Paper B7, Preprints of 1967 Conf. on Speech Communication and Processing, Air Force Cambridge Res. Lab., Bedford, Mass., Nov. 1967, pp. 100–103.

[13] L. G. Kersta, Nature 196, 1253 (1962).

[14] O. Tosi, "Speaker Identification through Acoustic Spectrography," Paper presented at XIV Int. Congr. Logopedics and Phoniatrics, Paris, Sept. 1968.

[15] M. A. Young and R. A. Campbell, J. Acoust. Soc. Amer. 42, 1250 (1967).

[16] K. N. Stevens, C. E. Williams, J. P. Carbonell, and B. Woods, J. Acoust. Soc. Amer. 44, 1596 (1968).

methods of identification. A set of eight talkers was used and a series of 14 identification tests was carried out. The performance of the observers improved over the series. The error rate for listening was always lower than for visual identification; at the best levels of performance, the average error rate was 6% for listening and 21% for visual identification. In further tests using new unknown talkers among the test samples, the observers were asked to judge whether a sample was spoken by any of the eight known talkers in the catalog. By listening, 6%–8% of the unknown talkers were incorrectly called known; by visual examination of the spectrograms, 31%–47% of the unknown talkers were called known. This result indicated that visual comparisons between spectrograms of talkers were less reliable than auditory comparisons.

The wide differences in error rate seen in these experiments reflect the strong dependence of voice identification judgments on specific conditions, in particular, on the experimental test procedures, but also on the experience and training of the observers, on the speaking conditions under which the speech samples are collected, and on instrumentation.

How relevant are these experiments to voice identification as used in legal trials? The task of the expert witness usually consists of judging the identity of a speaker from two sets of spectrograms, one from a known speaker (the accused) and the other from a speech sample associated with the case but produced by an unidentified speaker. This is neither a sorting nor a matching task. It is not matching because there is only one entry in the catalog of known speakers and the unknown speaker may not even be in this catalog. It is not sorting, because the spectrograms are already sorted into two categories: known and unknown. Further, all matching and sorting experiments reported in the literature employed a closed set of known size; the unknown sample with which the expert witness is confronted is drawn from an indefinitely large set of unidentified speakers. *None of the experiments in the literature has employed a comparable task.*

In addition to the results of controlled experiments, there are essentially anecdotal accounts of experiences in applying the methods of spectrographic voice identification to law enforcement problems. For example, we are informed that ". . . over 250 cases were processed for over 48 different law enforcement agencies in the United States and Europe which [is believed to be] a considerable body of practical proof, since no report of an error has occurred"; also, that a police officer has ". . . produced approximately 25 verified identifications where suspected persons admitted their guilt. In 37 cases the suspected persons were eliminated and released from any charges. . . ."[17] The question of what interpretation or reliance to put on reports of this general kind is a difficult one, because (1) the relevant facts may not be publicly available in some types of investigations, or the facts may be fragmentary and disputed, as in courtroom proceedings; (2) actual cases usually involve other kinds of evidence, so that the contribution of voice identification to the resolution of the case cannot be determined; and (3) neither legal resolution of a case nor confession of guilt gives reliable information about the correctness of voice identifications that may have been made. Perhaps a careful analysis of experience with the investigative uses of spectrographic voice identification could lead to dependable estimates of the practical reliability of the method as applied to courtroom proceedings; however, other methods using controlled experiments could be far more direct and would gain credibility by full disclosure of data and procedures.

Situations in which one speaker attempts to mimic another have not been examined in depth, but speech scientists have noted cases in which spectrograms of different talkers are very similar[18] and in which an experienced mimic with special playback aids can produce speech sequences whose spectrographic patterns are capable of being confounded with those of another talker.[19] There have also been reports of instances in which the speech spectrograms of a mimic appeared quite different from those of the individual being mimicked.[20]

IV. REQUIREMENTS FOR VALIDATION OF VOICE-IDENTIFICATION METHODS

What kinds of evidence would convince scientists of the reliability of speaker identification based on voice patterns?

The usual basis for the scientific acceptance of any new procedure is an explicit description of experimental methods and of results of relevant tests. The description must be sufficient to allow the replication of experiments and results by other scientists. We have seen, in the preceding section, two doubts that arise when we apply this criterion to voice identification based on spectrograms. First, fully reliable identifications were not the usual result, even in small-scale sorting and matching experiments. Second, even when experimental methods were explicit, they differed in kind and complexity, as well as in scale, from the practical task of positively identifying a man solely on the basis of voice patterns.

Lacking explicit knowledge and procedures, can individuals nevertheless acquire such expertise in identification from voice patterns that their opinions could be accepted as reliable? This possibility may exist, for the human eye and brain are superb instruments.

[17] L. G. Kersta (private communication), 1969; O. Tosi (private communication), 1969.

[18] P. Ladefoged and R. Vanderslice, "The 'Voiceprint' Mystique," Working Papers in Phonetics, Dep. Linguistics, Univ. Calif. Los Angeles, Nov. 1967, pp. 126–142.

[19] A. Fourcin and A. W. F. Huggins (private communication), 1969.

[20] L. G. Kersta, J. Acoust. Soc. Amer. 34, 1978(A) (1962).

But it cannot be assumed without proof. Validation of this approach to voice identification becomes a matter of replicable experiments on the expert himself, considered as a voice identifying machine.

Thus, voice identification might be accomplished either on the basis of explicit knowledge and procedure available to anyone, or on the basis of the unexplained expertise of individuals. In either case, validation requires experimental assessment of performance on relevant tasks.

Explicit procedures might be developed, based on specification of voice features useful for identification. Once the features were known, it would be important to learn how such features were distributed in the population. These distributions would permit an estimate of the size of the population of discriminable voices, and so give an indication of the reliability that would be theoretically attainable in specific situations.[21]

What would we need to know about the performance of the expert whose procedures are not fully explicit? First, the experiments with experts should be statistically valid models of the practical task. The tests should include judgments of whether two speakers are identical when one spectrogram is available from each speaker, and when more than one spectrogram is available. It may also be appropriate to perform tests in which the unknown talker, whose identity is to be determined from a spectrogram, may be drawn from a set of known talkers or may not be a member of this set. Test formats should yield information about the probabilities of missed identification as well as false identification, and the trade-off between them; also, about the effects of size of population, the nature of the spoken context in both known and unknown samples, and the type of display of voice pattern and its sensitivity to noise, distortion, or deliberate attempts to disguise the unknown voice.[22]

It may be objected that this minimal set of tests is unreasonably arduous. We do not believe that it is. As scientists, we could accept no less in checking the reliability of a "black box" supposed to perform speaker identification. This is how we must view the expert until he can provide an explicit and testable explanation of his methods.

V. SCIENTIFIC CRITERIA AND LEGAL ACCEPTANCE

Scientific and legal judgments differ in this basic respect: scientific acceptance is closely tied to technical evidence, whereas court determinations may rely heavily on the opinions of expert witnesses. When experts in recognized specialties differ in their opinions, the court may leave to a jury the assessment of conflicting opinions and of the relative expertise of witnesses. When new kinds of expert testimony are offered (e.g., speaker identification by spectrographic voice patterns), the court, before accepting such evidence, may first scrutinize the nature of the proffered expertise in relation to the consensus of informed scientific opinion. Today's consensus suggests that speaker identification by voice patterns is subject to error at a high, and as yet undetermined, rate.

Court determinations may also depend on the apparent validity of exhibits brought in evidence. Spectrographic evidence may often display features that are overwhelmingly influenced by the words spoken rather than by the speaker's identity. Judge and jury may therefore be misled in understanding the evidence and in assessing an expert's testimony.

VI. SUMMARY AND CONCLUSIONS

(1) Speech carries many simultaneous messages interwoven in a complex of words and phrases, moods, and individual voice characteristics. In their acoustic realization as speech, these messages are highly interdependent and thus difficult to disentangle. However, human observers can, to a limited extent, identify voices by ear or by visual examination of the acoustic patterns of speech.

(2) The acoustic speech signal can be analyzed in frequency, energy, and time, and recorded graphically to produce a spectrogram. Neither the spectrogram nor any other known process can directly display an individual's voice traits, because of the intermixing of these traits with the features that characterize words and phrases. At present, a human observer must examine the patterns of spectrograms and decide subjectively about the identities of talkers.

(3) Similarities and differences among spectrographic patterns are ambiguous and may be misleading. Prominent similarities usually indicate that similar sounds were spoken, but do not necessarily imply that they were spoken by the same person; differences in pattern, when the words are the same, may reflect differences of speaker or only normal variations in the utterances of a single speaker.

(4) Speech spectrograms, when used for voice identification, are not analogous to fingerprints, primarily because of fundamental differences in the sources of the patterns and consequent differences in their interpretation. For example, fingerprint patterns are a direct representation of anatomical traits. Vocal anatomy, on the contrary, is not represented in any direct way in voice spectrograms. In the interpretation of fingerprints, all points of similarity imply a match, although some more strongly than others; this simple

[21] F. R. Clarke, R. W. Becker, and J. C. Nixon, "Characteristics that Determine Speaker Recognition," Rep. ESD-TR-66-636, Decision Sciences Laboratory, Hanscom Field, Bedford, Mass., Dec. 1966 (report under contract to Stanford Res. Inst., Menlo Park, Calif.).

[22] Research projects on spectrographic voice identification, under sponsorship of the U. S. Department of Justice, are currently in progress at Michigan State Univ. (see Tosi, Ref. 14) and at Stanford Res. Inst. (see Hecker, Ref. 4).

relationship does not hold for the interpretation of voice patterns.

(5) Experimental studies of voice identification using visual interpretation of spectrograms by human observers indicate false identification rates ranging from zero to as high as 63%, depending on the type of task set for the observer, the observer's training, and other factors. Reliable machine methods for voice identification have not yet been established.

(6) Experience in applying spectrographic voice identification in law enforcement has led proponents of the method to express confidence in its reliability. The basis for this confidence is not, however, accessible to objective assessment.

(7) Experimental studies to assess the reliability of voice identification under practical conditions, whether by experts or by explicit procedures, have not yet been made, but the requirements for such studies have been outlined.

We find, in brief, that spectrographic voice identification has inherent difficulties and uncertainties. Anecdotal evidence given in support of the method is not scientifically convincing. The controlled experiments that have been reported give conflicting results. Furthermore, the experiments reported thus far do not provide a direct test of the practical task of determining whether two spoken passages were uttered by the same speaker or by two different speakers, one of whom may be a person unknown.

We conclude that the available results are inadequate to establish the reliability of voice identification by spectrograms. We believe this conclusion is shared by most scientists who are knowledgeable about speech; hence, many of them are deeply concerned about the use of spectrographic evidence in the courts. Procedures exist, as we have suggested, by which the reliability of voice identification methods can be evaluated. We believe that such validation is urgently required.

Appendix A: The Nature of Speech Information and Voice Identification

The root of the voice identification problem lies in the nature of speech itself. This may not be immediately apparent, because speech is so familiar and so seemingly simple, but intensive research, especially over the past 20 years, has shown it to be a very complex process that is only partially understood even now. We do know, however, one important way in which this complexity affects the problem of voice identification: speech carries many different messages simultaneously, about the speaker as well as about the words spoken, and it does this by blending them into a single acoustic stream. Thus, the speaker combines and encodes these messages for acoustic transmission and the listener must decode them. But now the component messages are no longer distinct, nor are message elements represented in any simple, direct way by acoustic elements of the speech.

The encoding operation, at least for the primary message of words and sentences, is a straightforward consequence of the articulatory process. As an example, in saying the word "tag," the vocal tract is first closed momentarily by bringing the tip of the tongue up to the dental ridge; then the tract is open wide by depressing the tongue tip rapidly while lowering the tongue body and jaw and again closed by lifting the back of the tongue. At every moment of this smoothly flowing gesture, the vocal cavities can serve as resonators to reinforce selectively certain frequency regions (the formants) of whatever sound is produced at the vocal cords or by turbulent air flow at narrow constrictions. The resulting acoustic stream has an over-all spectral pattern that is distinctive for the word "tag," but it does not have distinctive parts for the component gestures of "t" or "a" or "g." Each of the component gestures affects most of the tract for much of the time and generates a correspondingly encoded acoustic signal. A listener can, to be sure, hear "tag," and can identify the component speech sounds, but the perceptual processes by which he decodes the acoustic signal are far more involved than introspection would suggest. Indeed, much remains obscure about the perception of all the messages carried in speech.

What are these messages, and how do they relate to voice identification? There is, of course, the primary message comprising the words and sentences that were said. But *how* they were said (in what manner and mood and with what emphasis), by what class of speaker (man or woman, of what dialect and occupational group), in what assumed interpersonal relationship to one or many listeners, and even in what state of the speaker's health? These are some of the other messages carried by the speech and encoded into it. In addition, there is information about who is speaking. This is, for voice identification, the one message that must be teased apart from all the others.

By ear, identification seems an easy task, at least for the speech of friends and regular acquaintances. But wrong identifications are not uncommon, and there is no way to scrutinize directly the basis on which auditory identifications are made. An indirect method, employing the conversion of speech sounds into graphic patterns for recognition by eye, would seem to be an attractive alternative. Such a procedure, using sound spectrograms (voiceprints), was described by Kersta[A1] in 1962 and has been used by him and others since that time. There was earlier work dating back to the 1940's, in conjunction with military communications, but few records and no publications resulted.

The proponents of voiceprint identification have advanced a two-part theoretical basis for their methods:

[A1] L. G. Kersta, Nature 196, 1253 (1962).

(1) people differ anatomically, e.g., in the size and shape of their oral and nasal cavities, in the structure of the larynx, etc., and (2) people have different but stable habit patterns in the ways they use this vocal apparatus in speaking. All of these affect their speech spectrograms. Thus, it may be supposed that the combination of so many factors would uniquely characterize a particular speaker and set him apart from all others. The sound spectrogram, or voiceprint, is then used as an objective display in which to seek for enough points of similarity between the patterns of two spectrograms (one from a known speaker) to conclude that the suspected and known speakers are (or are not) the same person.

Clearly, these points of similarity should reflect only, or mainly, those things that characterize the speaker, i.e., his anatomy and his individualistic speech habits; otherwise, the information about the identity of the speaker is contaminated by the other messages in his speech. This is a major difficulty for, as we have seen, it is inherent in the nature of speech that the messages are acoustically interdependent and, in fact, the spectrographic patterns do show the mixed influences of quite different kinds of information.

An example of these mixed influences is provided by the vowels, which are among the simplest of speech sounds. A steady "ah," made with the vocal tract in a relaxed open-mouth position, has peaks, or formants, in the spectrogram that are determined mainly by the length and shape of the pharynx and mouth, and so reflect directly these anatomical and physiological dimensions of the speaker. But other vowels are characterized by different formants, which are produced by changing the positions and shape of jaw, tongue, and lips. Some effects of the speaker's body dimensions will remain, but these effects are changed in unobvious ways by how he moves his articulators to approximate the vowel formants that are typical of the dialect he speaks. Thus, formant frequencies are affected by at least three factors: the speaker's dialect, his tactics and precision in matching its norms, and the set of bodily dimensions that set him apart in some degree from other speakers.

Actually, the situation even for vowels is more complex. Vowels rarely occur as steady sounds; usually the formants are changing continuously. Often vowel formants do not reach their "target" frequencies (corresponding to the steady state of the vowel) at any point in a spoken syllable. Always, they are much affected by the particular consonants that precede and follow them. Speech is, in short, a highly dynamic set of gestures that so contorts the vocal tract as to make difficult and uncertain any precise inferences about its "normal" dimensions.

The gestures of speech are, however, highly repetitive, which should lead to stable habit patterns characteristic of the individual. But here, also, there is interaction between personal traits and the need to communicate. The latter imperative implies an agreed set of speech sounds that each speaker of a dialect must use if he is to be understood. But not every sound feature is so tightly specified; some may be varied at will without violating the phonetic rules, thus permitting each speaker some freedom to be individualistic in his speech gestures. Nevertheless, the fact that he and every other speaker of the same dialect must conform to its sound patterns implies that there will be more similarities than individual differences among their speech patterns. Clearly, such points of similarity, as they are reflected in spectrograms, must be carefully avoided in attempting to establish speaker identity. But differences in the spectrographic patterns can also be misleading. They may indicate that the sounds came from different speakers, or only that a single speaker was free to vary this sound feature and had done so. Thus, it would appear that neither points of similarity nor points of difference between two spectrograms will, in general, give unambiguous evidence as to whether the speech they represent came from the same speaker or from different speakers.

Are there, nevertheless, some points of similarity or difference that are not equivocal? Some that could serve as reliable "markers" of individual voices? As a technique for speaker identification, voiceprinting appears to be dependent on the existence of such similarities in spectrograms, and on human ability to recognize them by eye.

In assessing the method, it will be important to note what functions are performed by the spectrograph and what by the eye, and to weigh their respective roles in arriving at a subjective judgment about the speaker. Let us start with the instrument, with what it does, and with what it can tell the eye about speech.

The principles underlying the operation of the sound spectrograph were well known long before they were adequately implemented; indeed, a crude hand-drawn spectrogram, based on months of laborious analysis, had been published in the early thirties.[A2] The announcement by Bell Telephone Laboratories[A3] of an easy-to-use instrument that could make high-quality spectrograms, plus a brilliant exposition by Martin Joos[A4] of their relevance to acoustic phonetics, excited much interest and motivated a number of efforts to interpret spectrograms, to teach people to "read" them, and to devise automatic methods of speech recognition. It is a measure of the inherent difficulty, as well as the challenge, of these problems that the efforts to solve them have continued to the present time.

Although various types of sound spectrograph have been built for special purposes, the original instrument—and most of those now in common use with speech—are similar in principle and make very similar spectrograms. They analyze a recording of speech that

[A2] J. C. Steinberg, J. Acoust. Soc. Amer. 6, 16–24 (1934).
[A3] R. K. Potter *et al.*, J. Acoust. Soc. Amer. 18, 1–89 (1946); (A series of six articles).
[A4] M. Joos, "Acoustic Phonetics," Language Monogr., No. 23 (1948).

is about 2 sec in duration by replaying it repeatedly through a band-pass filter that has its effective center frequency swept slowly through the audio frequency range, typically about 0–8 kHz. The short-time-average amplitude of the filter output is recorded along the horizontal (time) axis of a sheet of electrosensitive paper, in step with the speech sample. Recordings of successive analyses are displaced upward along the vertical (frequency) axis in proportion to incremental increases in the center frequency of the filter. Thus, the spectrogram is generated as a raster of closely spaced horizontal lines. When average amplitude is recorded as degree of blackening, the spectrogram constitutes a "picture" of the spoken words (reading from left to right), with the dominant frequency components shown as the darker areas.

Originally, two types of spectrogram could be recorded, using either a wide-band or a narrow-band filter for the analysis. A third type, known as a contour spectrogram, was developed later and can be made on most present-day spectrographs. One might wonder, since all three types of spectrogram contain essentially the same information, what reasons there may be for this variety. Actually, the justification for three different displays lies partly in the nature of speech and partly in the characteristics of visual pattern perception. The relationship between these factors will emerge from descriptions of the different types of spectrograms.

Contour spectrograms capitalize on the preference of the eyes for closed figures with sharp boundaries, as do conventional contour maps, which they resemble. In a contour spectrogram, it is the transition from one amplitude level to another that is recorded as a dark line. The line will then enclose an area on the spectrogram for which all amplitudes lie above (or below) this particular transition amplitude. The areas between adjacent contour lines is shaded to give an approximate indication of absolute levels. Typically, successive contour lines (and shadings) change by 6-dB steps. The visual patterns of contour spectrograms are far more striking than those of narrow-band or wide-band spectrograms, and quantitative estimates of spectrum level can be made directly by eye with greater ease and accuracy than for the other two types. The visual patterns change substantially, though, with changes in amplitude distributions small enough that the ear would scarcely notice them; thus, two utterances that sound very alike may look quite different. It is not surprising, then, that contour spectrograms have gained little favor in speech research, despite their striking appearance and quantitative advantages. In the simplest terms, they emphasize to the eye a dimension of speech to which the ear is largely inattentive.[A5]

[A5] The ear is even less attentive to the relative phases of the component frequencies of the speech. Since relative phases are ignored in spectrograms but strongly affect the wave shapes of oscillograms, it may be that spectrograms owe a substantial part of their superiority as visual patterns to the fact that they omit phase information.

In generating narrow-band spectrograms, the effective width of the analyzing filter is about 45–50 Hz, and so resolves the individual harmonics of all but the deepest male voices. Consequently, the spectrograms consist of many equally spaced horizontal lines that may sweep upward and diverge, or may turn downward and converge, as voice pitch rises or falls in a voiced passage. Those harmonics that are close in frequency to the vocal cavity resonances are selectively reinforced and so appear as darker regions on the general background of (more or less) horizontal bands. The temporal resolution of the impulsive sounds of speech is smeared somewhat, though not excessively. The primary effect of the narrow-band filter is, however, to emphasize voice pitch and, accordingly, to deemphasize other features of the utterance.

Wide-band spectrograms, for which the analyzing filter has a bandwidth of about 300 Hz, do not resolve the voice harmonics (except for very high-pitched voices), but merge all of the harmonics within a formant region into one solid "bar" that will tilt up or down as the shape of the vocal tract is changed. The emphasis, for the eye, has been shifted to the changing shapes of the vocal tract, whereas voice-pitch information has been minimized though not removed, since it can still be seen in the spacing of the vertical striations during voiced sounds. The over-all patterns for wide-band spectrograms look simpler to the eye than do the corresponding patterns for narrow-band spectrograms, although they contain approximately the same amount of information.

The advantage of wide-band spectrograms lies only partly in their apparent simplicity; another and more important factor is that they emphasize the information that characterizes the primary message, i.e., the words or phrases that the speaker articulates. Indeed, the primary objective in developing the spectrograph was to find a way of recording speech that would emphasize the similarities and differences among words. It follows that other differences, and in particular the differences between speakers, have been subordinated to this primary objective. Clearly, we need—but do not know how to build—an instrument that would give good graphic patterns that emphasize the distinctive characteristics of speakers. It may be that, when we have learned much more about the sound features that characterize individual speakers, it will be possible to design an instrument that can be a powerful aid to the eye in voice identification, or even one that can operate automatically in a completely objective manner.

BIBLIOGRAPHY

P. B. Denes and E. N. Pinson, "The Speech Chain," Bell Telephone Laboratories, 1963 (obtainable through Bell System educational representatives).

G. Fant, *Acoustic Theory of Speech Production* (Mouton and Company, s'-Gravenhage, 1960).

J. L. Flanagan, *Speech Analysis Synthesis and Perception* (Academic Press Inc., New York, 1965).

A. M. Liberman, F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy, "Why are Speech Spectrograms Hard to Read?" Amer. Ann. Deaf 113, 127–133 (1968).

A. M. Liberman, F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy, "Perception of the Speech Code," Psychol. Rev. 74, 431–461 (1967).

R. K. Potter, G. A. Kopp, and H. C. Green, *Visible Speech* (D. Van Nostrand, Inc., New York, 1947).

K. N. Stevens and A. S. House, "An Acoustical Theory of Vowel Production and Some of its Implications," J. Speech Hearing Res. 4, 303–320 (1961).

Appendix B: Voice Patterns and Fingerprint Patterns

The term *voiceprint* implies an analogy with fingerprints. In this appendix we compare voice identification and fingerprint identification in their details, to see just how far they may be analogous. First, we review briefly the history and procedures of fingerprint identification.[B1,B2]

The first classification of fingerprint patterns, i.e., the ridge patterns on the underside of the finger tips, were made by J. Purkinje in 1823. He was concerned with these patterns only as part of the tactile sensory system and did not mention them as a basis for personal identification. In 1880, H. Faulds, a British medical missionary in Japan, first reported to the use of fingerprints as a device for identification in two criminal cases. From 1886 to 1888, Faulds attempted to interest Scotland Yard, but he was unsuccessful, probably owing to the lack of a workable classification scheme for files. Upon Fauld's publication, W. Herschel, an administrator in India, published an account of his 20-year use of fingerprints against impersonation and repudiation of signatures, and in registering prisoners. In this paper, Herschel also observed the permanence of fingerprint patterns, basing his conclusion on repeated prints of the same persons over intervals as long as 30 years. His use of fingerprints began in 1858 and official use in a few Indian government departments began in 1877.

Francis Galton, the great British geneticist and anthropologist, carried out the first extensive scientific study of fingerprints as compared with other methods of personal identification. His work began in 1888 and culminated in 1894 with the adoption of fingerprint classification as a main filing method used by the British government and as the ultimate proof of a person's identity. About 1901, fingerprint classification was adopted as the basis of the filing method. Subsequent developments in classification systems stemmed from those adopted in India by E. Henry (published in 1900) and in Argentina by Vucetich (beginning in 1891). Henry, during his early work, consulted Galton extensively.

Galton's studies on identification by fingerprints covered the following problems: (1) print persistance with growth, (2) probability of false identification, (3) indexing methods, (4) practical uses, and (5) comparison with other methods.

The work on *persistence* used Herschel's material, which was taken from 15 persons over age ranges beginning at various stages in life from infancy onward. The longest period studied was 30 years, the shortest 12 years. Repeated prints from about 25 different fingers were available. Certain general features of a print were found to change somewhat with age; for example, the sizes and distances increased with growth and the print texture or grain might change because of calluses or old age. However, the pattern type and the minute details remained remarkably constant. The details noted were the points of appearance or disappearance of a ridge, the occurrences of islands and enclosures, and the number of ridges involved in typical parts of the patterns, such as the whorls, loops, and deltas. The exact imprinted textures of these details would vary somewhat with changes in skin condition and with print-inking conditions. However, the details, when compared over age were always identifiable by their unchanging pattern, ridge count, and positions relative to each other. Six hundred ninety-nine out of 700 such details, when compared with the same print at the later age, remained unchanged as to type of feature, ridge count, and relative location. The one point changed from a divided ridge to a single ridge.

In his study of the *probability of false identification*, Galton first classified the thumb prints of 1000 different persons into 100 classes based on general pattern types, attempting to make the 100 classes appear equally different from each other. However, two-thirds of the prints fell onto only 12 classes; thus, there was the chance of about one in 18 that two patterns of the same class would be from different persons. This error rate (about 5.6%) was not acceptable in relation to the low rates to be had by considering the minute details of each print, which, conservatively estimated, numbered about 30 per print. Galton noted that a disagreement in general pattern type for the same finger would be conclusive evidence that the two prints were from different persons.

To estimate the probability of false identification based on the minute details, Galton first estimated the size of a square (in terms of ridge intervals) within which the detailed ridge pattern was 50% predictable if only ridge patterns outside the square are known; i.e., he then estimated that there was an equal chance of being

[B1] Francis Galton, *Finger Prints* (Macmillan and Company, Ltd., London, 1892; facsimile reprint, Da Capo Press, New York, 1965, with historical introduction by H. Cummins).

[B2] H. Cummins and C. Midlo, *Finger Prints, Palms and Soles: An Introduction to Dermatoglyphics* (Dover Publications, Inc., New York, 1961).

right or wrong as to the pattern within the square. Conservatively estimated, there were 24 such independent squares in a single print, and thus the probability that two different prints would correspond exactly within the independent squares was estimated as 1 in 2^{24}. When Galton included the chances of guessing the number and course of the ridges outside each independent square, the chance of false identity was estimated as 1 in 2^{36}, or about 1 in 64 000 000 000. For two fingerprints, this estimate is squared, for three cubed, etc. The chance of three fingerprints of two persons being identical in their minute details was estimated as 1 in $64^3 \times 10^{27}$.

Usually, there were about 35 details in a print. Not all might agree in two different prints of the same finger. If, for example, there are 35 points of comparison and all but four agree, the chances are estimated as 1 in 2^{29} that they are prints from different fingers.

In terms of modern theory a possible decision matrix would be as shown in Table B-I, where P_{SS} is the probability of the same print corresponding to the same person (i.e., the hit rate) and P_{SD} is the probability of the same print corresponding to different persons (i.e., the false alarm rate). Galton's probability calculations are used in the "Same prints" row. The probabilities in the "Different prints" row would reflect the fact that it is very easy to detect different prints by general pattern type, and thus the chances would be very small for an incorrect rejection or miss and correspondingly very high for a correct rejection.

Galton worked with 3000 prints to develop index methods for establishing files. This was important for his demonstrations and tests in his campaign for official adoption. His index system was the one initially used in 1894.

Another important aspect of Galton's work was a study of the relative power of fingerprint identification and the then current identification method of Bertillon, which employed a standard set of body measurements. The theoretical error rate of the basic body set was about 1 in 1 000 000; but Galton sought to see if, as with his general types of fingerprint patterns, there was significant correlation between classes of the measures. He made the set of body measurements on 500 persons and then distributed the measures into 243 Bertillon classes; two or three additional measures were then necessary to resolve the 24 cases falling in the largest class. He concluded that about 2000 persons could thus be distinguished. As an example of a Bertillon-indexed register of persons, a total of 20 000 persons were assumed, and then Galton pointed out that, with fingerprints added, this register could be expanded to 10 000 000, assuming that only 500 persons could be perfectly distinguished by fingerprints alone.

Thus, Galton modestly looked on fingerprints as a powerful supplement to current identification classes of body measurements. But only seven years elapsed before fingerprint classification in England officially superseded classification by body characteristics.

We note that, in contrast to the current practice of voice identification, the use of fingerprint identification in the 1890's rested on a rather large body of experience, on scientific studies of pattern permanence over time and of the statistical distribution of the pattern features, and on the development of a classification system.

The current American classification system for registering complete sets of fingerprints is virtually the same as the original one of Henry. The primary classes are based on the over-all pattern types, such as whorls,

TABLE B-I. Decision matrix showing possible correspondences of fingerprint patterns to subjects to be identified.

	S (Same person)	D (Different persons)
S (Same prints, one finger)	Correct identification $P_{SS}=1-\frac{1}{64\times10^9}$	False identification $P_{SD}=\frac{1}{64\times10^9}$
D (Different prints)	Incorrect rejection	Correct rejection

TABLE B-II. Comparison of fingerprints and voiceprints. This comparison shows that fingerprint identification is basically different from voice-pattern identification in several essential and critical aspects; further, it indicates that voice identification is inherently more complex and more susceptible to error than fingerprint identification.

Finger ridge patterns	Voice patterns
1. Patterns are inherent in anatomy, not changeable in kind, i.e., they cannot be changed from one pattern to another. Parts of pattern, large or small, can only be obliterated.	1. Patterns are only partially dependent on anatomy and are changed by the articulatory movements needed to realize the language code.
2. Details of patterns: (a) are permanent; (b) are not affected by growth (aging merely changes the size or the print grain); (c) are not affected by habits (calluses merely change the print grain).	2. Details of patterns: (a) are just as variable as the over-all patterns; (b) are affected by growth; (c) are affected by habits (learning new dialects and voice qualities).
3. Pattern similarity depends entirely on underlying anatomical structure.	3. Pattern similarity depends primarily on acquired movement patterns used to produce language code and only partially on anatomical structure.
4. Patterns result from a direct transfer from the skin of the finger to the surface touched by it.	4. Patterns result from an analysis of voice sounds which, in turn, are related only indirectly to the vocal anatomy of the speaker. Moreover, the transmission channel from speaker to spectrograph is vulnerable to acoustical and electrical distortions.

arches, and loops, rather than on the minute details such as enclosures, ridge counts, islands, deltas, terminations, origins, and bifurcations. Each set is first classed according to the pattern, over the 10 fingers, of the presence or absence of a whorl in each print. There are thus 2^{10} (i.e., 1024) such classes; they are called primary classes. As Galton found, some of these classes contain many more cases than others; for example, the class *no whorls* contains about 25% of all the cases. Thus, the primary classes must be subdivided into secondary classes based on patterns, over the 10 prints, of occurrences and types of arches and loops. Further subordinate divisions are made by ridge counts of loops and whorls.

Anatomical studies of the ridge patterns have established that the general pattern types, such as loops and whorls, are due to hereditary factors. However, the minute details such as the bifurcations, terminations, and interruptions of ridges, are the final bases for fingerprint identification. These minutiae are determined mainly by random processes in the prenatal development of the skin.

The randomness of the minutiae and their number per print form a statistical basis for individual identification that is virtually unequivocal. Cummins and Midlo[B2] (p. 151) estimated that the chance that prints from two different fingers will match in all details is about 1 in 3×10^{42}. This probability is based on the number of the combinations of 25 binary events falling in all the possible patterns at 25 locations in the print [1 in $(\frac{1}{2}\times25)^{25}$].

The matching of a single chance print depends much more on the pattern of minute features than on the overall pattern. Classification of single prints, where such files are kept, generally starts with a police category related to the crime at which the print was collected.

Partial prints are sometimes the only print samples available for a fingerprint identification. In these cases, the fingerprint expert first attempts to decide which finger of which hand each print belongs to. Then the partial prints are compared with filed prints. As far as we could determine, there are no published experimental studies of the statistical success of identification by partial fingerprints. Experts agree that perfect correspondence of 12 points of comparison, and no discordances, proves that two prints originate from the same finger (Ref. B2, pp. 152–155, 182–183).

Using the foregoing information, we have compared fingerprint identification and voice-spectrogram identification as to their basic pattern sources. We have arranged the essential points of our comparison in Table B-II.

Appendix C: Experimental Evidence on Voice Identification

I. IDENTIFICATION BY SUBJECTIVE EXAMINATION OF SPECTROGRAMS

Several experimenters have investigated the ability of observers to identify talkers by visual examination of voice spectrograms. In most of these experiments, a limited group of talkers was used. Spectrograms were available of each of the talkers uttering a fixed set of words or phrases several times. The exact nature of the observers' tasks differed considerably among the various experiments.

The first experiments were reported by Kersta.[C1,C2] His observers were female high school students who worked in pairs in the final test experiments. They were given about one week of training in the interpretation of voice spectrograms. In some of the experiments reported, their task consisted of sorting an array of spectrograms into individual talker categories; they knew the number of different talkers and the number of examples from each talker. The experiments involved 5, 9, and 12 talkers who were drawn at random from a pool of 123 male talkers. In a typical experiment (say with 12 talkers), spectrograms were made of four utterances of a one-syllable word, such as "you," spoken in isolation, by each talker. The observers were given the 48 spectrograms and were asked to sort them into 12 piles corresponding to the 12 different talkers. No further details of the training and test procedures, nor of the instructions to the observers are reported. Error scores of individual observers ranged from zero to 2%; average error rates, pooled over all observers, ranged from 0.35% for five talkers in the set to 1% for 12 talkers in the set. In another task, the observers had to identify a talker by matching his spectrograms against those in a catalog of a set of talkers; the size of the set ranged from 9 to 15 talkers; the test words were spoken in context; the average error score was 1%.

Experiments with a matching procedure were carried out by Stevens *et al.*[C3] For the most part, these experiments employed a closed set of eight talkers. The observers were college students with no previous training in reading spectrograms. In a number of experimental sessions, observers working individually were given a set of eight "standard" spectrograms, representing a given utterance spoken by each of the eight talkers. They were then presented "unknown" spectrograms, one at a time, each representing the same utterance by

[C1] L. G. Kersta, "Voiceprint Identification," Nature 196, 1253–1257 (1962).

[C2] L. G. Kersta, "Voiceprint-Identification Infallibility," J. Acoust. Soc. Amer. 34, 1978(A) (1962). See also, L. G. Kersta, "Voiceprint Identification," J. Acoust. Soc. Amer. 34, 725(A) (1962); L. G. Kersta, "Voiceprint Classification," J. Acoust. Soc. Amer. 37, 1217(A) (1965); L. G. Kersta, "Voiceprint Classification for an Extended Population," J. Acoust. Soc. Amer. 39, 1239(A) (1966).

[C3] K. N. Stevens, C. E. Williams, J. P. Carbonell, and B. Woods, "Speaker Authentication and Identification: A Comparison of Spectrographic and Auditory Presentations of Speech Material," J. Acoust. Soc. Amer. 44, 1596–1607 (1968).

one of the talkers, and were asked to make an identification of the talker. Some learning and consequent improvement of error scores occurred in the first few sessions, and then the scores reached a relatively uniform level. Error scores ranged from 18% to 50%, depending upon the utterance; error scores were generally higher for brief monosyllabic utterances than for utterances with more syllables.

A procedure similar to that of Stevens *et al.*, was followed in an experiment by Young and Campbell.[C4] Their observers, however, had some training in spectrogram interpretation for purposes of matching talkers; also, the experiment was designed to compare identification using words spoken in context *versus* identification using words spoken in isolation. The observers examined spectrograms of two different words (*it* and *you*) spoken by five talkers. The catalog for matching consisted of all five talkers. The observers were first trained and tested with spectrograms of the isolated words; then they were tested on the words taken from context. With the words spoken in isolation, the average error score was 21.6%. With the words spoken in context, the average error score was 62.7%. Young and Campbell point out that words spoken in context have shorter durations and different acoustic patterns than words spoken in isolation and that these differences may have been responsible for the large difference in identification.

An experiment similar to that of Young and Campbell just described was carried out by R. Bruce.[C5] He used six talkers, and his "standard" spectrograms consisted of 10 key words (*the, to, and, me, on, is, you, I, it,* and *a*) spoken in isolation. He had one test utterance, which was a long sentence containing all of these words, and again the task of the observer was to determine the speaker of the test utterance, using all 10 spectrograms of key words as standards. The error rate for this task was about 50%.

Other evidence with regard to identification of speakers from spectrograms has been reported by Ladefoged and Vanderslice.[C6] Although these investigators did not carry out formal experiments, they provide examples to show that spectrograms of the same speaker producing the same utterance twice can appear to be quite different, whereas two different speakers can produce very similar spectrograms. Other investigators have reported informally on situations in which one speaker attempts to mimic another; in one of these situations, an experienced mimic with special aids produced speech sequences whose spectrographic patterns appeared capable of being confounded with those of another talker being mimicked; in another case, the spectrograms of a famous mimic appeared to be different from those of the speaker being mimicked.[C7,C2]

Tosi[C8] recently reported confirmation of the average error rate, about 1%, in the spectrogram sorting task originally used by Kersta. Further, in a matching task, Tosi used two utterances of each of five test words arranged in a catalog of 50 different talkers selected from a larger set of 123 talkers. The spectrograms were the same as those used for the work reported by Kersta.[C2] Seven of Tosi's observers were police fingerprint technicians. Several days of training preceded the experimental tests. The error rate for observers working individually ranged from zero to 11.1% with a median of 5.7%. The error rate for observers working together in pairs ranged from 3.2% to 14.3% with a median of 7.7%.

None of the experiments reported to date has employed an observer's task that simulates the task commonly encountered by the expert in voiceprint identification, namely, the task of deciding the identity of a speaker of a known key utterance and the same utterance spoken by an unknown speaker. However, Stevens *et al.*[C3] carried out some tests where the observers were asked to judge whether a sample was produced by any of the eight known talkers in a catalog. New "unknown" talkers among the test samples were incorrectly called "known" 6%–8% of the time by listening and 31%–47% of the time by visual examination of the corresponding spectrograms.

II. IDENTIFICATION BY LISTENING

A number of experiments have examined how well a listener can identify a talker from the sound of his voice. In most of these experiments, the task of the listener was to identify a talker from an ensemble of several (typically, 5–10) talkers that were known to him or whose voices were avaliable to him on recordings. Error scores in such experiments are in the range of 5% to 19%, depending on the conditions of the experiment.[C3,C9,C10] Stevens *et al.*[C3] found that listening gave better talker identification than the examination of spectrograms of the same utterances. The observers also reported higher confidence in listening than in their visual identifications.

There is an appreciable decrease in error scores for a two- or three-syllable sample of speech compared

[C4] M. A. Young and R. A. Campbell, "Effects of Context on Talker Identification," J. Acoust. Soc. Amer. 42, 1250–1254 (1967).

[C5] R. Bruce, unpublished study, MIT, 1966.

[C6] P. Ladefoged, and R. Vanderslice, "The 'Voiceprint' Mystique," Working Papers in Phonetics, UCLA, 126–142, Nov. 1967.

[C7] A. Fourcin and A. W. F. Huggins (private communication), 1969.

[C8] O. Tosi, "Speaker Identification Through Acoustic Spectrography," Paper presented at XIV Int. Congr. Logopedics and Phoniatrics, Paris, Sept., 1968.

[C9] I. Pollack, J. M. Pickett, and W. H. Sumby, "On the Identification of Speakers by Voice," J. Acoust. Soc. Amer. 26, 403–406 (1954).

[C10] P. D. Bricker and S. Pruzansky, "Effects of Stimulus Content and Duration on Talker Identification," J. Acoust. Soc. Amer. 40, 1441–1449 (1966).

with a one-syllable sample.[C8,C11] Recognition 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.[C9] Several investigators have shown that fairly high aural recognition scores are obtained when the task is to identify two sequential speech samples as being spoken by the same talker or by different talkers. Correct identification was 93% in such tests using one-syllable words.[C12] With five-syllable speech samples, Clarke *et al.*[C13] found about 89% correct same–different decisions for two samples, about 85% correct for matching an unknown sample to one of two different samples, and about 60% correct matching to four samples.

It has been observed that some parts of the speech frequency range are more important than others in their contributions of cues for speaker recognition. For example, when speech is processed by octave-band filters, the best recognition scores are obtained with the filter 1200–2400 Hz. Low-pass filtering at 3000 Hz or high-pass filtering at 500 Hz caused little deterioration relative to wide-band speech. Noise affects the ability of a listener to recognize a talker'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 −4 to +8 dB.[C13,C14]

[C11] C. E. Williams, "The Effect of Selected Factors on the Aural Identification of Speakers," in *Methods for Psycho-acoustic Evaluation of Speech Communication Systems*, Rep. ESD-TDR-65-153, Electronic Syst. Div., Air Force Syst. Command, Hanscom Field, Mass., 1964.

[C12] J. A. Williamson, "An Investigation of Several Factors Which Affect the Ability to Identify Voices as Same or Different," unpublished dissertation, Univ. Edinburgh, 1961.

[C13] F. R. Clarke, R. W. Becker, and J. C. Nixon, "Characteristics that Determine Speaker Recognition," Rep. ESD-TR-66-636, Decision Sci. Lab., Hanscom Field, Bedford, Mass., Dec., 1966 (report under contract to Stanford Res. Inst., Menlo Park, California).

III. IDENTIFICATION BY OBJECTIVE METHODS

There have been a number of experiments on the design and evaluation of methods for objective voice identification using completely automatic procedures.[C2,C15,C16–C20] Typically these studies have employed sets of about 10 talkers and a process of spectrum analysis to produce the voice patterns. The patterns were fed to a computer where they were processed statistically to yield a reference pattern for each talker. A new utterance from one of the talkers is then analyzed and fed to the computer for comparison with each reference pattern. A measure of the similarity of the new pattern to each reference pattern is then computed and the reference talker yielding the highest similarity is chosen as the identity of the talker of the new utterance. Error rates in these experiments were about 10%.

[C14] R. W. Peters, "Studies in Extra-messages: Listener Identification of Speakers' Voices under Conditions of Certain Restrictions Imposed on the Voice Signal," Project Rep. No. NM 001-064.01.03, U. S. Naval School of Aviation Medicine, Pensacola, Fla., Oct. 1954.

[C15] S. Pruzansky, "Pattern-Matching Procedure for Automatic Talker Recognition," J. Acoust. Soc. Amer. 35, 354–358 (1963).

[C16] S. Pruzansky and M. V. Mathews, "Talker-Recognition Procedure Based on Analysis of Variance," J. Acoust. Soc. Amer. 36, 2041–2047 (1964).

[C17] W. A. Hargreaves and J. A. Starkweather, "Recognition of Speaker Identity," Language and Speech, 63–67 (1963).

[C18] L. G. Kersta, Paper B7, Preprints of 1967 Conf. on Speech Commun. and Process., Air Force Cambridge Res. Labs., Bedford, Mass., 100–103, Nov. 1967.

[C19] K. P. Li, J. E. Dammann, and W. D. Chapman, "Experimental Studies in Speaker Verification Using an Adaptive System," J. Acoust. Soc. Amer. 40, 966–978 (1966).

[C20] S. K. Das, "A Method of Decision Making in Pattern Recognition," IEEE Trans. on Computers C-18, 329–333 (1969).

Appendix D: Requirements for Validation of Voice-Identification Methods

Any identification method that is highly reliable will depend ultimately on the use of a sufficient number of independent identifying features that are stable for the individual and have a wide variation in the population. Voice features having these characteristics may eventually be found, but we can expect that considerable research will be necessary because of the complexity of speech.

Several investigators have studied identification features of voice that may be useful.[D1,D2,D3] The statistical methods they[D1,D3,D4] describe seem to be promising approaches to identification from a population of talkers. Research on individual voice features is still in its early stages and there is no large body of data and experience on which to base a system of features for identification.

An alternative to using a well-defined set of identification features is to train a human observer to perform voice identification, using whatever procedures seem to yield the best results. Experimental tests of such observers, as summarized in Appendix C, have not been extensive, and some have not employed appropriate statistical designs. All the reported tests have used sorting or matching tasks and closed (usually small) sets of speakers and subjects. None of the tests has covered identification tasks that resemble real-life situations.

[D1] S. Pruzansky, "Pattern-Matching Procedure for Automatic Talker Recognition," J. Acoust. Soc. Amer. 35, 354–358 (1963).

[D2] W. D. Voiers, "Perceptual Bases of Speaker Identity," J. Acoust. Soc. Amer. 36, 1065–1073 (1964).

[D3] F. R. Clarke, R. W. Becker, and J. C. Nixon "Characteristics that Determine Speaker Recognition," Rep. ESD-TR-66-636, Decision Sciences Lab., Hanscom Field, Bedford, Mass., Dec., 1966 (report under contract to Stanford Res. Inst., Menlo Park, Calif.).

[D4] S. Pruzansky and M. V. Mathews, "Talker-Recognition Procedure Based on Analysis of Variance," J. Acoust. Soc. Amer. 36, 2041–2047 (1964).

Validation of any method of identification poses important problems in statistical decision making. Some of the important questions are these: What is the probability that the same person will be judged as different (missed identification) and the probability that different persons will be called the same (false identification)? How do these errors trade off against each other, and what is the influence of population size? It is instructive to formulate these questions in terms of statistical decision theory.

An appropriate model assumes that each observation of a talker's voice defines a point in a multidimensional space. Successive utterances of the talker define a distribution of points in that space. Comparable utterances by other people define other distributions. The reliability of identification is tied up with the degree of overlap between the distributions. The dimensions of the multidimensional space concern properties of speech used in the identification scheme. The dimensions may be either subjective (pitch or loudness for example) or objective (fundamental voice frequency or sound intensity, for example), depending upon whether the identification judgment is made entirely subjectively by a human observer or is based upon physical measurements taken from a spectrogram or some other physical analysis. This decision model can form the basis for validation of a voice-identification scheme.

The problem of determining the probability that two samples are drawn from the same distribution, i.e., that they came from the same speaker, is well known in statistical theory. There are well-documented decision techniques for deciding (with stated probabilities of error) whether two observations come from the same or different multidimensional distributions. However, the application of these techniques depends upon knowledge of the distributions involved as well as on the dimensions in which they are defined. These dimensions are closely tied to the "features" in a pattern which distinguish it from patterns drawn from other distributions. In the case of speech spectrograms, we know neither the dimensions nor their distributions. Thus, current experiments run the risk of producing atypical results because of selection of test materials and subjects.

The applicable distributions and dimensions depend not only on the "natural" voice characteristics of the statistical population, but also on the instrumentation used to display the voice, on factors in the recording environment, on the context in which key words, phrases, or sounds are spoken, and on attempts to disguise the voice. Also, the experimental paradigm of a voice-identification test must be such that the data collected can be readily interpreted in terms of the statistical model discussed above.

With these considerations in mind, specific investigations can be proposed to determine the validity of personal identification using speech spectrograms. We assume that the relevant task is to decide from samples of two utterances if these were spoken by the same or different talkers. The various types of investigations fall into four classes: first, those concerned with understanding and characterizing the normal subjective process of comparing speech spectrograms for identification; second, experiments comparing the subjective findings with comparable results from other more "objective" methods of talker identification; third, studies to assess the performance of experts in voice identification by means of speech spectrograms; and fourth, studies to define statistical procedures for validation of a specific identification.

I. UNDERSTANDING AND CHARACTERIZING THE PROCESS OF SUBJECTIVE IDENTIFICATION USING SPEECH SPECTROGRAMS

One way to achieve quantitative results from the subjective process is the following. Observers are presented all the possible talker pairs of test spectrograms, one pair at a time. Pairs of separate utterances of the test words by the same talker are included, but the observer is not told how many talkers there are. For each pair, the observer's task is to judge whether they are from the same talker or from different talkers. A matrix is then constructed to show the percentages of "same" judgments for all the voice pairs. The confusion rates can then be used as input measures of similarity to multidimensional scaling, cluster analysis, and other similar techniques. These provide metric representations that, as quantitative descriptions of the subjective process of identification, *may* lead to some insight into physical dimensions underlying the process. For further analyses, one could also calculate various summary statistics (e.g., interpoint distances, measures of "tightness" of clusters, etc.) and derive statistical properties, such as their distributions. Further experiments could vary the number of talkers and thereby study the behavior of the patterns in the confusion matrices, their summary statistics and statistical distributions, and their correlations with physical dimensions. This would clarify the dependence of these statistical measures on the number of talkers and help to evaluate their stability.

II. COMPARISONS WITH OBJECTIVE METHODS OF TALKER IDENTIFICATION

One could compare the confusion matrices and the values and distributions of the summary statistics, as obtained in the subjective experiments, with the corresponding entities obtained in more directly quantifiable and objectively describable processes of talker identification applied to an acoustic analysis of the same set of utterances. Such comparisons would help to detect the existence and nature of differences between the subjective process and these more objective procedures. Also, the objective procedures may be employed to guide the

selection of the acoustical samples for further subjective experiments.

III. ASSESSMENT OF EXPERTISE

To compare performance in the task of voice identification of "experts" with that of "average" persons, the experiments described above (which would ordinarily use "average" persons as subjects) could be repeated with a group of "experts." The resultant data summaries could then be compared with the data from "average" subjects to determine if, for example, (a) the confusion matrix for experts is more nearly "diagonal" in structure (i.e., sparse almost everywhere except for the same-speaker pairs) than the matrix for "average" subjects; (b) the clustering is "tighter" for the experts; etc.

IV. VALIDATION OF A SPECIFIC IDENTIFICATION

For "validating" an identification based on a single pair of speech spectrograms, one possibility is to obtain direct measures of the physical dimensions (uncovered by the approach described above) for the pair of spectrograms in question. One can, in such a case, calculate a statistic to measure the closeness of the pair (e.g., the "intervoice" distance in multidimensional space) and see where its value falls in the reference distribution for these physical dimensions.

Investigations of the four types suggested are minimal for judging the validity of voice identification using speech spectrograms. In practice, it seems likely (and desirable) that acceptance of any scheme for general use would be based upon both an analysis of laboratory data *and* usage tests under realistic conditions undertaken to confirm the results from the experiments.

Apart from the statistical model suggested above, one might well ask whether it is possible to validate entirely through actual use. The question in this form is difficult to answer in more than a "common sense" fashion. From this viewpoint, it would seem that validation would require experience with some substantial fraction, say 1%–10%, of the population of the community in question (the United States, for example). Experience might be limited to a smaller fraction of the population, but there would be the nagging doubt that troublesome cases might show up until a larger number of cases had been examined. This doubt could not be removed by any sampling procedure since the relevant voice characteristics are unknown. Granted that a long string of successes under these conditions might eventually convince many people, the point at which the case was "proved" would remain very much a value judgment based on inference, and could not be said to be established in a scientific sense.