Reprinted from
17 October 1969, Volume 166, pp. 338-343

SCIENCE

Identification of a Speaker by Speech Spectrograms

Richard H. Bolt, Franklin S. Cooper, Edward E. David, Jr.,
Peter B. Denes, James M. Pickett, Kenneth N. Stevens

Identification of a Speaker by Speech Spectrograms

How do scientists view its reliability for use as legal evidence?

Richard H. Bolt, Franklin S. Cooper, Edward E. David, Jr.,
Peter B. Denes, James M. Pickett, Kenneth N. Stevens

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.

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 article) to study and report on this issue (*2*). In considering the 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 voiceprinting 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*).

The authors' addresses are: Dr. Bolt, Bolt Beranek and Newman Inc., 50 Moulton Street, Cambridge, Massachusetts 02138; Dr. Cooper, Haskins Laboratories, 305 East 43 Street, New York 10017; Dr. David and Dr. Denes, Bell Telephone Laboratories, Murray Hill, New Jersey 07971; Dr. Pickett, Gallaudet College, Washington, D.C. 20002; Dr. Stevens, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge 02139.

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 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 identify 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 which 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 comparisons of the spectrograms and by subjective judgments about the identity of the speakers represented (see *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*).

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

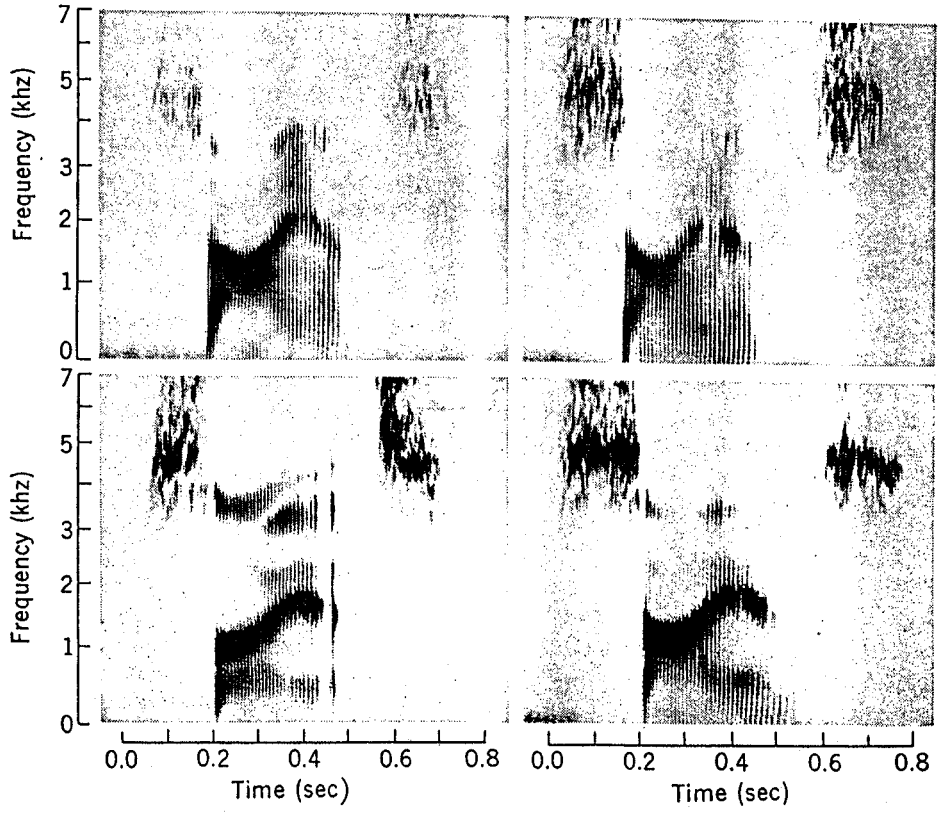

Fig. 1. Four spectrograms of the spoken word "science." The vertical scale represents frequency, the horizontal dimension is time, and darkness represents intensity on a compressed scale. Three of the spectrograms are from three different speakers and the remaining spectrogram is a repetition of the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

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 interruptions. These details are determined mainly by random processes in prenatal skin development. There are a sizeable 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.

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 ten 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 percent (*11, 12*).

Subjective experiments using speech spectrograms have been of two types: (i) sorting experiments in which the observer sorts a set of spectrograms of a test word into individual talker categories; and (ii) 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 to 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 percent for 5 talkers in the set to 1.00 percent 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 percent with a median of 0.42 percent.

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 percent; the range of error rates, over the ten different test words employed, was 0 to 3 percent. 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 percent with a median of 5.7 percent. The error rate for observers working together in pairs ranged from 3.2 percent to 14.3 percent, with a median of 7.7 percent.

In another matching experiment (*15*) using a set of five talkers and trained observers, the average error rate was 22 percent for words spoken in isolation. When the words were spoken in fluent context and matched against the same isolated words in the catalog, the error of talker identification was 63 percent.

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 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 percent for listening and 21 percent 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 to 8 percent of the unknown talkers were incorrectly called known; by visual examination of the spectrograms, 31 to 47 percent 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 exeperiments 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 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 the 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, first, because 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; second, because 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 third, neither legal resolution of a case nor confession of guilt gives reliable information about the correctness of voice identifications that may have been made. It is conceivable that 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)*.

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, 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 procedures 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 specifications 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, nature of the spoken context in both known and unknown samples, and 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.

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 (for example, 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.

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 examintion 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 relationship does not hold for the interpretation of voice patterns.

5) Experimental studies of voice identification by using visual interpretation of spectrograms by human observers indicate false identification rates ranging from zero to as high as 63 percent, depending on the type of task set for the observer, his 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.

References and Notes

1. W. Koenig, H. K. Dunn, L. Y. Lacey, *J. Acoust. Soc. Amer.* **18**, 19 (1946).
2. This article is not a report of the Acoustical Society of America, and the opinions given are those of the authors as individuals.
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 appendices. These will be on file with the Secretary, Acoustical Society of America, or they may be published in the *Journal of the Acoustical Society of America* at the discretion of its editor; a detailed scientific review of voice identification has been prepared by M. Hecker, *Methods for Measuring Speaker Recognition* (Stanford Research Institute, Menlo Park, Calif., April 1969).
5. G. Fant, *Acoustic Theory of Speech Production* (Mouton & Co., 'S-Gravenhage, 1960); K. Stevens and A. S. House, *J. Speech Hear. Res.* 4, 303 (1961).
6. A. M. Liberman, F. S. Cooper, D. P. Shankweiler, M. Studdert-Kennedy, *Amer. Ann. Deaf* **113**, 127 (1968).
7. F. R. Clarke, R. W. Becker, J. C. Nixon, *Characteristics that Determine Speaker Recognition*, Report ESD-TR-66-636 (Decision Sciences Laboratory, Hanscom Field, Bedford, Mass., December 1966); W. D. Voiers, *J. Acoust. Soc. Amer.* **36**, 1065 (1964); C. E. Williams, in *Methods for Psycho-acoustic Evaluation of Speech Communication Systems*, Report ESD-tdr-65-153 (Electronic Systems Division, Air Force Systems Command, Hanscom Field, Bedford, Mass., 1964), sec. III.
8. L. G. Kersta, *Nature* **196**, 1253 (1962); H. Mennen, H. Tillman, G. Ungeheuer, *Entwicklung eines Systems von Beschreibungsmerkmalen für Konturson agramme zum Zwecke der Sprecheridentifikation*, T-888-L-203 (Institut für Phonetik und Kommunikationsforschung Universität Bonn, November 1968).
9. S. Pruzansky, *J. Acoust. Soc. Amer.* **35**, 354 (1963).
10. H. Cummins and C. Midlo, *Fingerprints, Palms and Soles: An Introduction to Dermatoglyphics* (Dover, New York, 1961); F. Galton, *Finger Prints* (Macmillan, London, 1892; facsimile reprint, Da Capo Press, New York, 1965).
11. S. Pruzansky and M. V. Mathews, *J. Acoust. Soc. Amer.* **36**, 2041 (1964).
12. L. G. Kersta, *Preprints of the 1967 Conference on Speech Communication and Processing*, paper B7 (Air Force Cambridge Research Laboratories, Bedford, Mass., November 1967), p. 100.
13. L. G. Kersta, *Nature* **196**, 1253 (1962).
14. O. Tosi, "Speaker identification through acoustic spectography," paper presented at XIV Int.

Congress Logopedics and Phoniatrics, Paris, September 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, B. Woods, *ibid.* **44**, 1596 (1968).
17. L. G. Kersta, personal communication, 1969; O. Tosi, personal communication, 1969.
18. P. Ladefoged and R. Vanderslice, in *Working Papers in Phonetics* (Dept. of Linguistics, Univ. of California, Los Angeles, November 1967), p. 126.
19. A. Fourcin and A. W. F. Huggins, personal communication, 1969.
20. L. G. Kersta, *J. Acoust. Soc. Amer.* **34**, 1978 (1962), abstract.
21. F. R. Clarke, R. W. Becker, J. C. Nixon, *Characteristics that Determine Speaker Recognition*, Report ESD-TR-66-636 (Decision Sciences Laboratory, Hanscom Field, Bedford, Mass., December 1966).
22. Research projects on spectrographic voice identification, sponsored by the U.S. Department of Justice, are currently in progress at Michigan State University (see Tosi, *14*) and at Stanford Research Institute (see Hecker, *4*).