Speaker identification by speech spectrograms: some further observations*

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

Franklin S. Cooper
Haskins Laboratories, 270 Crown Street, New Haven, Connecticut 06510

Edward E. David, Jr.
Gould Incorporated, 8550 West Bryn Mawr Avenue, Chicago, Illinois 60631

Peter B. Denes
Bell Laboratories, 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

(Received 5 February 1973)

This letter reviews recent research on speaker identification by means of comparisons of speech spectrograms by human observers. Various factors affecting the reliability of identification are discussed, particularly those that would be present in practical forensic situations. Our interpretations of the new data lead us to reiterate our previous conclusion: that the degree of reliability of identification under practical conditions has not been scientifically established.

Subject Classification: 9.1, 9.9.

In 1969 and 1970 we reviewed the scientific basis of speaker identification through the use of speech spectrograms in connection with legal proceedings.[1,2] We cited experimental results that showed, for example, error rates ranging from 6% to 63% false identifications under various conditions that could be encountered in forensic situations. We concluded that the scientific information available at that time was not adequate to provide valid estimates of the degree of reliability of voice identification by examination of spectrograms.

In our papers we suggested and described some experiments required to establish this technique on a scientifically solid basis. The key question to be answered is of the form: What are the odds? What are the probabilities of correct identification, of incorrect identification, or of missed identification, of a person through examination of speech spectrograms representing his voice and the voices of other persons? What are the probabilities under the particular set of conditions involved in the forensic situation? Relevant conditions include the selection and number of persons represented by the spectrograms examined, the methods by which the voice samples were recorded, the time and circumstances when the recordings were made, and the confidence criteria of the examiner in making his decisions.

We concern ourselves only with these scientific questions about the determination of probabilities. Whether any particular value of these probabilities would qualify speech spectrograms as admissible for evidence in court is a legal question about which we offer no judgment.

The present paper updates our review of this subject. We now take into account new information which appeared in 1971 and 1972.

A project on voice identification through acoustic spectrography has been completed by Dr. O. Tosi and his co-workers at Michigan State University; the results have appeared in two laboratory reports[3,4] and in a paper in this JOURNAL.[5] This project simulated in the laboratory some of the conditions present in practice and these are our primary concern. Other workers have investigated the acoustic characteristics that bear on speaker identification[6–8] and the effects of the speech context.[9,10] Hecker has reviewed the basic problems.[11]

Tosi's experiments and similar experiments by others[12] employ the following procedure. Some number of observers participate in the experiment by examining spectrograms and making judgments about them. An observer receives a set of "known" spectrograms and one or more "unknown" spectrograms, which respectively represent a set of "known" persons and a single, "unknown" person whose identity is not divulged to the observer. By examining the spectrograms, the observer judges whether the "unknown" person is the same as one of the "known" persons and, if so, which one. Preliminary training may be provided for the observers; in Tosi's study one month of training was given.

Tosi's experiments involved the effects of five variables: (1) The number of speakers in the known set; (2) open versus closed tests—in open tests the observer was not told whether the unknown speaker was represented in the known set, but in closed tests he was told that the unknown was represented in the set; (3) the context of the speech materials—test words were spoken either in isolation or in sentences; (4) certain characteristics of the speech transmission system; and (5) contemporary versus noncontemporary voice samples—the voice samples to be compared were recorded either on the same occasion (contemporary) or on different occasions (noncontemporary).

Identification errors are of two types: (1) Errors of *false identification*: The observer selects from the known set a speaker who is not the person represented in the unknown spectrograms, and (2) errors of *false rejection* or missed identification: in open tests the observer wrongly decides that the unknown speaker is not represented in the known set.

In the forensic situation, false identification could erroneously single out a particular individual as one of the "suspected" persons. Such errors take on special significance in that they relate to the possible conviction of an innocent person. Errors of false rejection on the other hand, are important in investigative work because they may lead to the elimination of a guilty person from consideration as a suspect.

A wide range of results was obtained in Tosi's study, depending on how the experimental variables were combined.[13] At one extreme, error scores were below one percent for closed tests with words spoken in isolation and with contemporary voice samples. Other results more relevant to forensic applications were found for noncontemporary voice samples (in this case recorded one month later), open rather than closed tests, and words spoken in various sentence contexts rather than in isolation. Several of the results for these forensically relevant conditions provide insight into the problem of speaker identification by spectrograms. When the unknown spectrogram actually represented one of the speakers in the known set, the observers failed to recognize the speaker about 29% of the time.[14] They either identified the unknown as the wrong speaker (5%), or, more frequently, they decided that the unknown speaker was not one of the known speakers (24%).

Table I. Percent errors of false identification reported in the study of Tosi *et al.*, for open tests, words in random context, and trials in which the unknown speaker was not represented in the set of known speakers. Data taken from Appendix B of Ref. 3, rounded off to the nearest 0.1%.

Number of individuals in known set of voices	Contemporary spectrograms	Noncontemporary spectrograms
10	2.0%	4.9%
20	4.0%	8.3%
40	4.5%	9.8%

The results for trials in which the unknown speaker was *not* represented in the known set are given in Table I. For tests in which all unknown spectrograms were *contemporary* the rate of false identifications ranged from 2.0% to 4.5%, depending on the number of voices in the known set. On the other hand, when the unknown spectrograms were *noncontemporary*, this error rate more than doubled to the range of 4.9% to 9.8%.

Why was the percent of false identification higher for noncontemporary voice samples? Speakers' voices might be expected to change from one recording occasion to another, but if the observers used the same average criterion for a match the percent of false identification should not change unless there were differences in similarity between the voices included in the several test sets. Was there some aspect of the experimental design that would account for the result? Was there a change, between the two kinds of tests, in the criterion used by the observers for accepting a match? The reports offer no explanation of this result.

Another result of Tosi's experiments that has substantial consequences for the forensic situation is an increase in the error rate when the context of the test words is changed from words in isolation (an average error rate of 7% for noncontemporary spectrograms) to words embedded in random sentence contexts (16%). Different sentence contexts are known to modify the acoustic characteristics of words. These findings, coupled with the above mentioned increases in error rates for noncontemporary as opposed to contemporary unknown spectrograms, suggest that any experimental condition that is likely to cause a change in the acoustic characteristics of an utterance will lead to an increased probability of error. This increase occurs both for errors of false identification and errors of false rejection.

Not examined in Tosi's study are some other factors that can change the sounds a speaker produces, factors that can increase the intraspeaker variability—these in turn can increase the probability of error. For example it is well known that changes in the psychological state of a talker, induced through emotions or other types of stress, can cause substantial changes in the characteristics of his speech sounds.[15] In a forensic situation additional emotional factors of this kind tend to be present. Other factors that can potentially modify the characteristics of a speaker's voice are the noise level surrounding him, attempts at mimicking or disguise,[7] room acoustics, and recording conditions. Further research is needed to determine the influence of these factors on the reliability of speaker identification.

Because of these sources of increased intraspeaker variability, we regard the 5% to 10% false-identification rates seen in Table I as artificial minima which are likely to increase when conditions depart from the laboratory situation in which the voice samples were recorded. This evaluation of the projection that can safely be made from Tosi's experimental findings differs sharply from his own interpretation[16] and from that expressed in a letter[17] written and circulated by Dr. Peter Ladefoged; further, we question the basis on which claims[18] have been made that the dominant view of the scientific community is now in agreement with those interpretations.

In our previous papers we compared voice identification with fingerprint identification (See Ref. 2, pp. 606–608). We drew a contrast between the stable, anatomical nature of the finger ridge patterns and the changeable patterns of a person's speech. Tosi's results provide direct evidence of the detrimental effect of intraspeaker variability on voice identification and its inherent dissimilarity to fingerprint identification.

The experiments of Hazen[9,10] indicate high identification errors for words from conversation. His first experiment involved closed tests, spectrograms of test words spoken in isolation, and 50 speakers in the known set. In a second experiment with the same observers and speakers, the same test words were taken from conversation, and both closed and open tests were carried out. The closed tests resulted in 3% error with the words spoken in isolation[19] compared with 20% for words taken from conversation.[20] On the open tests with words from conversation[21] the false identifications were 17% and the false rejections were 67%.

The present level of knowledge about personal voice characteristics, their recognition, and how they change under different conditions is still rudimentary. The recent work on speaker identification from spectrograms does not provide any new understanding as to which spectrographic features correlate most clearly or efficiently with the speaker's identity. Research on this question is in progress,[6,8] but results have not yet been applied to the problems of speaker identification from spectrograms. At the present time, therefore, the spectrographic identification of a voice by a trained observer appears to rely on a broad assessment of loosely defined points of similarity rather than on a carefully specified set of objectively defined spectrographic attributes. The Tosi experiments, in fact, show considerable disagreement among different panels of observers as to what constitutes a match when they are given the same matching task. For example, in situations where an incorrect identification is made by one panel, other panels are by no means in agreement with this assessment. In fact, the percent of time that an error of false identification is made by at least one panel (out of nine used in the study) is several times higher than the values given in Table I, which are average data over all panels.

The decision criterion of the observer is of critical importance in personal identification. The Tosi study required the observers to rate their subjective certainty for each decision. A detailed breakdown of decision errors was not given as a function of level of certainty except to say that about 60% of the errors were committed on decisions rated by the observers as uncertain. Further studies are needed to provide a better understanding of the decision process. For example, no explanation is now possible as to why, in open tests, an observer who is uncertain cannot simply reject the unknown spectrogram as not being similar enough to any of the known spectrograms.

Table I illustrates the influence of another variable not fully analyzed in the Tosi reports: the effect of the size of the known set. Increasing the number of persons in the known set from 10 to 40 at least doubled the probability of making a false identification. This result suggests that the use of still larger population sizes would further increase the probability of false identification.

The Tosi reports refer to a field study of voice identifications by the Michigan Department of State Police which employed both spectrographic and auditory comparison of voices in actual police cases.[22] It was found that no positive identification was contradicted by other police evidence. However, we do not consider this type of evidence a reliable criterion of the correctness of identification. The only true criterion of correctness of identification is sure knowledge of the identity of the speaker.

In discussing the forensic application, Tosi and his colleagues say that the error rates may in fact be lower than the values found in their experiments. They reason that a prudent practitioner can exercise caution and can listen to the voice samples as well as view the spectrograms. However, the Tosi reports give no scientific data that define the practitioner's error rate, or show how the rate might vary with his degree of caution, or indicate what improvement can be had by listening.

The Tosi study has improved our understanding of some of the problems of voice identification from spectrograms by indicating the influence of several important variables on the accuracy of identification. In uncovering factors that tend to increase identification errors, however, the study has not given us a definitive answer to the question: "How reliably can a person be identified by examining the spectrographic patterns of his speech sounds?" Under certain laboratory conditions and for some selected sample of the population, the probability of making an error in identification can be stated. But for the less-than-ideal conditions encountered in forensic situations, the indications are that the probability of error will increase substantially. Further studies are needed, with particular attention to the examiner's decision criteria, the selection of speaker population, the time lapse between voice samples, background-noise conditions, and the psychological condition of the speaker.

As scientists rather than lawyers, we offer no judgment as to whether or to what extent speech spectrograms should be used for identification in the courts. We wish only to point out that present methods for such use lack an adequate scientific basis for estimating reliability in many practical situations and that laboratory evaluations of these methods show increasing errors as the conditions for evaluation move toward real-life situations. We hope that our explanations of some of the factors that affect speaker identification will provide the legal profession with helpful information on which to base its own judgments concerning the admissibility of the spectrographic method.[23]

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

[1] R. H. Bolt, F. S. Cooper, E. E. David, Jr., P. B. Denes, J. M. Pickett, and K. N. Stevens, "Identification of a Speaker by Speech Spectrograms," Science **166**, 338–343 (1969).

[2] R. H. Bolt, F. S. Cooper, E. E. David, Jr., P. B. Denes, J. M. Pickett, and K. N. Stevens, "Speaker Identification by Speech Spectrograms: A Scientists' View of its Reliability for Legal Purposes," J. Acoust. Soc. Am. **47**, 597–612 (1970).

[3] O. Tosi, H. J. Oyer, W. B. Lashbrook, C. Pedry, and J. Nichol, "Voice Identification through Acoustic Spectrography," Report SHSLR 171, Department of Audiology and Speech Sciences, Michigan State University, East Lansing, Michigan (Feb. 1971).

[4] O. Tosi, H. J. Oyer, W. Lashbrook, C. Pedry, J. Nichol, and E. Nash, "An Experiment on Voice Identification: Excerpts from Report SHSLR 171," Department of Audiology and Speech Sciences, Michigan State University, East Lansing, Michigan (July 1971).

[5] O. Tosi, H. Oyer, W. Lashbrook, C. Pedrey, J. Nichol, and E. Nash, "Experiment on Voice Identification," J. Acoust. Soc. Am. **51**, 2030–2043 (1972).

[6] J. J. Wolf, "Efficient Acoustic Parameters for Speaker Recognition," J. Acoust. Soc. Am. **51**, 2044–2056 (1972).

[7] W. Endres, W. Bambach, and G. Flosser, "Voice Spectrograms as a Function of Age, Voice Disguise, and Voice Imitation," J. Acoust. Soc. Am. **49**, 1824–1848 (1971).

[8] M. R. Sambur, "Speaker Recognition and Verification Using Linear Prediction Analysis," J. Acoust. Soc. Am. **53**, 354(A) (1973).

[9] B. M. Hazen, "Speaker Identification Using Spectrograms Made on Different Spectrographs," M.S. Thesis, State University of New York at Buffalo (1972).

[10] B. M. Hazen, "The Effects of Changing Phonetic Context on the Voiceprint Identification Technique," Ph.D. Thesis, State University of New York at Buffalo (1972).

[11] M. Hecker, *Speaker Recognition: An Interpretative Survey of the Literature* (American Speech and Hearing Association, Washington, D. C., 1971), ASHA Monograph No. 16.

[12] See, for example, L. G. Kersta, Nature (Lond.) **196**, 1253 (1962); K. N. Stevens, C. E. Williams, J. P. Carbonell, and B. Woods, J. Acoust. Soc. Am. **44**, 1596 (1968).

[13] See Ref. 3, Table 3, p. 34.

[14] See Ref. 3, p. 37.

[15] M. H. L. Hecker, K. N. Stevens, G. von Bismarck, and C. E. Williams, "Manifestations of Task-Induced Stress in the Acoustic Speech Signal," J. Acoust. Soc. Am. **44**, 993–1001 (1968). C. Williams, and K. Stevens, "Emotion and Speech: Some Acoustical Correlates," J. Acoust. Soc. Am. **52**, 1238–1250 (1972).

[16] See Ref. 5, p. 2041–2042.

[17] P. Ladefoged, "An Opinion on "Voiceprints," *Working Papers in Phonetics* (U.C.L.A. Phonetics Laboratory, University of California at Los Angeles, California, 1971), No. 19, p. 84–87.

[18] See, for example, testimony on December 15, 1971, in the case of U.S. versus Raymond, U.S. District Court, District of Columbia, Case #Cr 800-71.

[19] See Ref. 9, Table 4, p. 21.

[20] See Ref. 10, Table 12, p. 41.

[21] See Ref. 10, Table 9, p. 34.

[22] Department of Michigan State Police, *Voice Identification Research, PR 72-1*, (U.S. Department of Justice, Superintendent of Documents, U.S. G.P.O., Washington, D.C., 1972), No. 2700-0144, pp. 77–78.

[23] For a review of the legal questions see J. Crim. Law, Criminol. Police Sci. 63, 343–355 (1972); and Georgetown Law J. 61, 703–745 (1973).