Spectrographic voice identification: A forensic survey

Bruce E. Koenig

Federal Bureau of Investigation, Engineering Section, Technical Services Division, 8199 Backlick Road, Lorton, Virginia 22079

(Received 25 October 1985; accepted for publication 18 February 1986) - J. Acoust. Soc. Am 79(6) June 1986

A survey of 2000 voice identification comparisons made by Federal Bureau of Investigation (FBI) examiners was used to determine the observed error rate of the spectrographic voice identification technique under actual forensic conditions. The qualifications of the examiners and the comparison procedures are set forth. The survey revealed that decisions were made in 34.8% of the comparisons with a 0.31% false identification error rate and a 0.5 3% false elimination error rate. These error rates are expected to represent the minimum error rates under actual forensic conditions.

PACS numbers: 43.70.Jt

INTRODUCTION

The sound spectrograph is a device which produces a visual graph (spectrogram) of speech as a function of time (horizontal axis), frequency (vertical axis), and voice energy (gray scale or color differences).1,2 It is a well-accepted research tool that is used to study individual vowel characteristics, physiological speech anomalies, etc. However, in the field of forensic voice identification, it has yet to find approval among most scientists in phonetics. linguistics, engineering, and related disciplines as a positive test in comparing voice samples.3-6

Historically, forensic applications were not seriously considered until 1962 when Lawrence Kersta published the results of experiments which reflected error rates of 0% to 3% for one-word spectral comparisons in closed sets (examiner always knows a match exists) of 12 or less speakers.7 In 1972, the findings of a large-scale study at Michigan State University were published in which attempts were made to more closely imitate law enforcement conditions, but only spectral comparisons were made (no aural). The “forensic model” included open set trials (examiner did not know if a match existed), noncontemporary samples (1 month apart), trained examiners, and high-confidence decisions. This resulted in an approximate error rate of 2% for false identification (no match existed but the examiner selected one, or a match existed but the examiner chose the wrong one) and 5% for false elimination (a match existed but the examiner failed to

recognize it). The authors of the study attempted to extend the experimental results to actual law enforcement conditions, which they thought would lower the error rates. They theorized that examiners could aurally compare the voice samples, the number of known suspects would be limited by police investigation, there would be no time limits placed on the examiner, only very high confidence decisions would be used, and additional known voice samples could be obtained.8 Other scientists disagreed on the study's extensions, and stated that in actual forensic conditions the error rate would increase, not decrease.' In 1979, a committee of the National Research Council released its findings and recommendations in a Federal Bureau of Investigation (FBI) -funded study on the reliability of spectrographic voice identification under forensic conditions, which found, in part, that:

(1) Error rates vary from case to case due to the properties of the voices compared, the recording conditions used to obtain voice samples, the skill of the examiner, and the examiner's knowledge about the case. Estimates of error rates are available only for a few situations, and they "do not constitute a generally adequate basis for a judicial or legislative body to use in making judgments concerning the reliability and acceptability of aural-visual voice identification in forensic applications."10

(2) Examiners should fully use all available knowledge and techniques that could improve the voice identification method.'0

(3) Spectrographic voice identification assumes that intraspeaker variability (differences in the same utterance repeated by the same speaker) is discernable from interspeaker variability (differences in the same utterance by different speakers); however, that "assumption is not adequately supported by scientific theory and data." Viewpoints on actual error rates are presently based only on "various professional judgments and fragmentary experimental results rather than from objective data representative of results in forensic applications."''

FBI examiners have used the spectrographic technique since the 1950s for investigative support, but have not provided expert court testimony on comparison results.'2

This paper presents the results of 2000 forensic comparisons, under actual law enforcement conditions, by FBI examiners.

. SURVEY PROCEDURES

The FBI conducts forensic voice identification examinations using the spectrographic or voiceprint technique for the FBI, other Federal agencies, state and local law enforcement authorities, and many foreign governments. After each examination is conducted, a written report of findings is mailed to the contributor with the name of the examiner and the disposition of the submitted voice samples. If an identification or elimination is made, the contributor is contacted by telephone and asked if the results are consistent with interviews and other evidence in the investigation. If other information strongly supports the voice comparison result, then the contributor is told to contact the FBI if later developed evidence contradicts the finding. If the voice comparison results contradict other evidence, the matter is closely followed until legally adjudicated or investigatively closed. In the few occurrences where no final determination was possible, the voice comparison result was considered a “no decision” in the survey.

The results of the last 2000 requested comparisons, spanning 15 years, were compiled and organized into total identification and elimination decisions, known errors, and no or low confidence decisions.

II. QUALIFICATIONS OF EXAMINERS

All of the individuals conducting the voice comparison examinations were FBI employees with the following qualifications: (I) at least two years of full-time experience in voice identification and analysis of tape recorded voice signals using sophisticated digital and analog analysis and filtering equipment; (2) completion of over 100 voice comparisons in actual cases; (3) completion of a basic two week course in spectrographic analysis, or equivalent; (4) passing a yearly hearing test; (5) formal approval by other trained examiners; and (6) a minimum of a Bachelor of Science Degree in a basic scientific field.

III. COMPARISON PROCEDURES

The following procedures were used, if at all possible, on every attempted voice comparison in the survey.

(1) Only original recordings of voice samples were accepted for examination, unless the original recording had been erased and a high-quality copy was still available.

(2) The recordings were played back on appropriate professional tape recorders and recorded on a professional full-track tape recorder at 7 1/2 ips. When possible, playback speed was adjusted to correct for original recording speed errors by analyzing the recorded telephone and AC line tones on spectrum analysis equipment. When necessary, special recorders were used to allow proper playback of original recordings that had incorrect track placement or azimuth misalignment.

(3) Spectrograms were produced on Voice Identification, Inc., Sound Spectrographs, model 700. in the linear expand frequency range (0-4000 Hz), wideband filter (300 Hz) and bar display mode. All spectrograms for each separate comparison were prepared on the same spectrograph. The spectrograms were phonetically marked below each voice sound.

(4) When necessary, enhanced tape copies were also prepared from the original recordings using equalizers, notch filters, and digital adaptive predictive deconvolution programs13,14 to reduce extraneous noise and correct telephone and recording channel effects. A second set of spectrograms was then prepared from the enhanced copies and was used together with the unprocessed spectrograms for comparison.

(5) Similarly pronounced words were compared between two voice samples, with most known voice samples being verbatim with the unknown voice recording. Normally, 20 or more different words were needed for a meaningful comparison. Less than 20 words usually resulted in a less conclusive opinion, such as possibly instead of probably.

(6) The examiners made a spectral pattern comparison between the two voice samples by comparing beginning, mean and end formant frequency, formant shaping, pitch, timing, etc., of each individual word. When available, similarly pronounced words within each sample were compared to insure voice sample consistency. Words with spectral patterns that were distorted, masked ‘by extraneous sounds, too faint, or lacked adequate identifying characteristics were not used

(7) An aural examination was made of each voice sample to determine if pattern similarities or dissimilarities noted were the product of pronunciation differences, voice disguise, obvious drug or alcohol use, altered psychological state, electronic manipulation, etc.

(8) An aural comparison was then made by repeatedly playing two voice samples simultaneously on separate tape recorders, and electronically switching back and forth between the samples while listening on high-quality headphones. When one sample had a wider frequency response than the other, bandpass filters were used to compensate during at least some of the aural listening tests.

(9) The examiner then had to resolve any differences found between the aural and spectral results, usually by repeating all or some of the comparison steps.

(10) If the examiner found the samples to be very similar (identification) or very dissimilar (elimination), an independent evaluation was always conducted by at least one, but usually two other examiners to confirm the results. If differences of opinions occurred between the examiners, they were then resolved through additional comparisons and

discussions by all the examiners involved. No or low confidence decisions were usually not reviewed by another examiner.

IV. SURVEY RESULTS

The survey found that in 2000 voice comparisons, the following decisions and errors were observed:

Decisions	Number	Percent (%)
No or low confidence	1304	65.2
Eliminations	378	18.9
Identifications	318	15.9
Errors		
False eliminations	2	0.53
False identification	1	0.31

Most of the no or low confidence decisions were due to poor recording quality and/or an insufficient number of comparable words. Decisions were also affected by high-pitched voices (usually female) and some forms of voice disguise.

V. CONCLUSIONS

(1) The observed identification and elimination errors probably represent the minimum error rates expected under actual forensic conditions, since

investigators are not always correct in their evaluation of a suspect's involvement, due to limited physical evidence, faulty eyewitness statements, etc.

(2) The stated results should only be considered valid when compared with examiners having the same qualifications and using the same comparison procedures.

(3) The FBI has emphasized signal analysis and pattern recognition skills for conducting voice identification examinations, more than formal training in speech physiology, linguistics, phonetics, etc., though a basic knowledge of these fields is considered important. ACKNOWLEDGMENTS

Thanks are due to the following colleagues who were involved in conducting the comparisons used in this survey:

Steven A. Killion, Barbara Ann Kohus, Dale Gene Linden, Gregory J. Major, Artese Savoy Kelly, Keith W. Sponholtz, Ernest Terrazas, Richard L. Todd, and Charles Wilmore, Jr.

1 w. Koenig, H. K. Dunn, and L. Y. Lacey, J. Acoust. Soc. Am. 18, 244(1946)
2G. M. Kuhn, J. Acoust. Soc. Am. 76, 682—685 (1984).
3R. H. Bolt, F. S. Cooper, E. E. David, Jr., P. B. Denes, J. M. Pickett, and K. N. Stevens, J. Acoust. Soc. Am. 47, 591—612 (1970).
4R. H. Bolt, F. S. Cooper, E. E. David, Jr., P. B. Denes, J. M. Pickett, and K. N. Stevens, J. Acoust. Soc. Am. 54, 531—534 (1974).
5K. N. Stevens, C. E. Williams, J. R. Carbonell, and B. Woods, J. Acoust Soc. Am. 44, 1596—1607 (1968).
6R. H. Bolt, F. S. Cooper, D. M. Green, S. L. Hamlet, J. G. McKnight, J. M. Pickett, 0.1. Tosi, and B. D. Underwood, "On the Theory and Practice of Voice Identification," N.A.S.N.R.C. Publ. (1979).
7L. G. Kersta, Nature 196, 1253—1257 (1962).
80. Tosi, H. Oyer, w. Lashbrook, C. Pedrey, J. Nicol, and E. Nash, J. Acoust. Soc. Am. 51, 2030—2043 (1972).
9R. H. Bolt, F. S. Cooper, E. E. David, Jr., P. B. Deres, J. M. Pickett, and K. N. Stevens, J. Acoust. Soc. Am. 54, 53 1—534 (1974).
10R. H. Bolt, F. S. Cooper. D. M. Green, S. L. Hamlet, J. G. McKnight, J. M. Pickett, 0. I. Tosi, and B. D. Underwood, N.A.S.N.R.C. Publ., 60 (1979).
11R. H. Bolt, F. 5: Cooper, D. M. Green, S. L. Hamlet, J. G. McKnight, J. M. Pickett, 0. 1. Tosi, and B. D. Underwood, N.A.S.N.R.C. Publ., 2 (1979).
12B. E. Koenig, FBI Law Enforcement Bulletin (January and February, 1980).
13J. E. Paul, IEEE Circuits and Systems Magazine 1, 2—7 (1979).
4J. E. Paul, paper presented at Voice Interactive Systems Subtag, Orlando, FL (Oct. 1984); hosted by U. S. Army Avionics Research and Development Activity, Ft. Monmouth, NJ.