Proceedings of Meetings on Acoustics

Volume 19, 2013 http://acousticalsociety.org/

ICA 2013 Montreal
Montreal, Canada
2 - 7 June 2013

Speech Communication
Session 1pSCc: Distinguishing Between Science and Pseudoscience in Forensic Acoustics II

1pSCc1. Mismatched distances from speakers to telephone in a forensic-voice-comparison case

Ewald Enzinger*

***Corresponding author's address: Forensic Voice Comparison Laboratory, School of Electrical Engineering and Telecommunications, University of New South Wales, UNSW Sydney, 2052, New South Wales, Australia, e.enzinger@student.unsw.edu.au**

In a forensic-voice-comparison (FVC) case, one speaker (A) was talking on a mobile telephone, and another (B) was standing a short distance away. Later, B moved closer to the telephone. Shortly thereafter, there was a section of speech where the identity of the speaker was disputed. All material for training an FVC-system could be extracted from this single recording, but there was a near-far mismatch: Training data for A were near, training data for B were far, and the disputed speech was near. We describe a procedure for addressing the degree of validity and reliability of an FVC system under such conditions, prior to it being applied to the casework recording: Sections of recordings of pairs of speakers of known identity are used to train an A and a B model; multiple other sections from each of the A and B recordings are used as test data; a likelihood ratio is calculated for each test section; and system validity and reliability are assessed. Prior to training and testing, the A and B recordings were played through loudspeakers and rerecorded via a mobile-telephone network, B was rerecorded twice, once with the loudspeaker near and once with it far from the telephone.

Published by the Acoustical Society of America through the American Institute of Physics

INTRODUCTION

In a forensic-voice-comparison case, one speaker (speaker A) was talking on a mobile telephone, and another (speaker B) was standing a short distance away. Later, speaker B moved closer to the telephone. Shortly thereafter, there was a section of speech where the identity of the speaker was in question. Based on the circumstances of the case, it was determined that the hypotheses to be considered are

- the questioned utterance was spoken by speaker A

  or

- the questioned utterance was spoken by speaker B

and that this is an exhaustive list of hypotheses, i.e., a priori the probability that the speaker of questioned identity could be a speaker other than one of these two is zero. All material for creating models representing these hypotheses and, thus, for training a forensic-voice-comparison system could be extracted from the recording of the conversation. However, there was a near-far mismatch: Training data for speaker A were near, training data for speaker B were far, and the questioned utterance was near.

This paper describes the procedures used to calculate a likelihood ratio for this case, but using data taken from a research database rather than the recordings from the case itself.

METHODOLOGY

Data

Recordings of two male speakers were taken from a database of Australian English speakers. The recordings were of a telephone conversation, with each speaker recorded using a high-quality microphone. To replicate the conditions of the case, the nominal high-quality recordings were rerecorded via a mobile-telephone network. Two loudspeakers were placed in a soundbooth, one approximately 10 cm away from a mobile telephone and the other 1.5 m away. The high-quality recordings were played though the loudspeaker and the acoustic signal picked up by the in-built microphone of the mobile telephone through which a call was established to the receiving landline telephone (Polaris NRX EVO 450), which was connected to an external sound card via a Trillium Telephone Recording Adapter Studio Interface (RECADPT-SI). Each speaker was rerecorded twice, once from the loudspeaker that was near the telephone (*near* condition) and once from the loudspeaker that was far from the telephone (*far* condition). The recordings were played in a conversation-like manner, i.e., a utterance was player from the near loudspeaker, then one from the far loudspeaker, then from the near loudspeaker again.

Forensic-voice-comparison system

Amplitude normalization

In the case there was a mismatch in distances to the mobile telephone's microphone. All else being equal, the signal from a speaker who is further from the microphone is of lower amplitude. Since the extracted features are based on signal amplitude, we normalized the amplitudes of the recordings made in the *near* and *far* conditions. The RMS amplitude was calculated for the signal in each section of each recording. The signal in each section of each recording was then scaled accordingly to match the highest level that did not incur clipping on any of the recording sections.

Feature extraction

Mel frequency cepstral coefficients (MFCCs) were extracted using a MATLAB implementation (Ellis, 2005). First, a pre-emphasis filter (coefficient value 0.97) was applied to the signal. Using a 20-ms-wide Hamming window shifted in 10 ms steps, the signal was divided in a series of frames. The power spectrum for each frame was then multiplied by a filter bank consisting of 26 triangular-shaped filters with a 50% overlap. In the mel-frequency scale the filters all had the same width and overlap. The 26 filters covered the frequency range 300–3300 Hz. A discrete cosine transform (DCT) was fitted to the logarithm of the 26 filter outputs, and the 1st through 11th DCT coefficient values (MFCC values) were used for subsequent

analysis. Finally, liftering (filtering of log-power spectra, Bogert *et al.*, 1963) was applied to the cepstral coefficients. The lifter coefficient was set to 22.

Unlike the analysis performed for the actual case in which all speech portions were manually marked, in this study an automatic speech activity detector was used to select the sections of the recording for acoustic analysis.

Use of data for testing of validity and reliability

The amount of data used for speaker modeling and test segments was matched to that available in the analysis performed on the case by selecting the same number of feature vectors (see Table 1).

TABLE 1: Number of feature vectors used for training data and test segments.

	Number of feature vectors
training data A	4605
training data B	2647
test segment	207

In the present study the near-far mismatch of the data available in the case was accounted for as follows: The training data for speaker A consisted of 90% of feature vectors extracted from recordings in the *far* condition and 10% of feature vectors extracted from recordings in the *near* condition. Data for speaker B were taken from recordings in the *near* condition.

Speaker modeling

Each speaker was modeled using a single multivariate normal distribution. Mean vectors and covariance matrices were calculated from MFCC feature vectors from frames in the training data for each speaker. Histograms and pairwise scatterplots of the training data of each speaker suggested that the assumption of normality was reasonable.

Likelihood ratio calculation

For a questioned utterance, a score was calculated as in

$$\mathrm{s} = \frac{1}{N}\sum_{n=1}^{N}\log\left(\frac{f(x_n|\mu_A,\Sigma_A)}{f(x_n|\mu_B,\Sigma_B)}\right), \tag{1}$$

where x_n is the MFCC feature vector for frame n, N the number of frames, and μ_A, Σ_A, μ_B, and Σ_B are the mean vectors and covariance matrices of speakers A and B, respectively.

The score was then converted to an interpretable likelihood ratio via logistic-regression calibration (Brümmer and du Preez, 2006; van Leeuwen and Brümmer, 2007; Morrison, 2013). Training data for logistic-regression calibration were generated using cross validation. A consecutive set of frames the same length as the number of frames in the questioned utterance was randomly selected from either the training data of speaker A or speaker B, a mock-questioned utterance. All the remaining data were used to build models for speaker A and B. A score was then calculated for the mock-questioned utterance. The extracted data were replaced and the procedure repeated generating 100 scores known to be from speaker A and 100 scores known to be from speaker B. These scores were then used to train a logistic-regression model. The weights from the logistic regression model were then used to convert the score from the actual questioned utterance to a likelihood ratio (calculations were performed using Brümmer, 2005, and Morrison, 2009).

RESULTS AND DISCUSSION

Cross-validated tests were performed in which mock questioned utterances of the same length as the questioned utterance were randomly selected from the training data and used to calculate a likelihood ratio. Calibration parameters were obtained from scores calculated from models and test segments selected from the remaining training data. 1000 mock questioned utterances were extracted from the training data of speaker A and 1000 from the training data of speaker B, resulting in 1000 likelihood-ratio values each.

The cross-validated tests of the forensic-voice-comparison system resulted in a log-likelihood ratio cost (C_{llr}, van Leeuwen and Brümmer, 2007; Morrison, 2011) value of 0.025. A Tippett plot is provided in Figure 1.

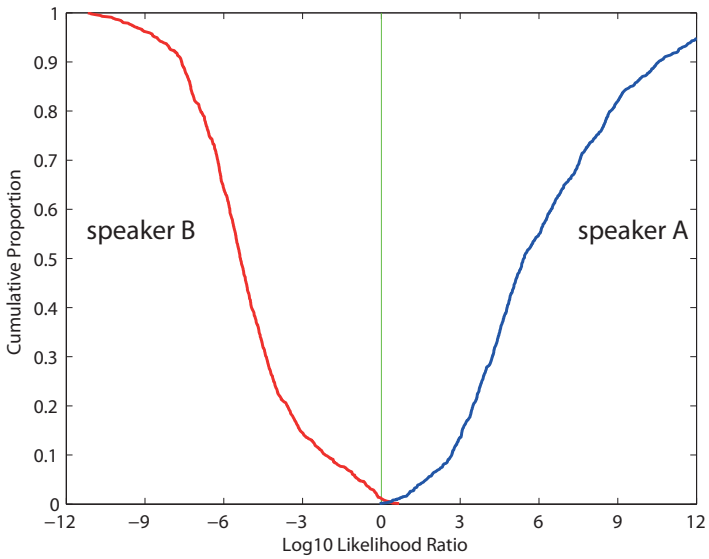

FIGURE 1: Tippett plot indicating system performance when tested on mock questioned utterances the same length as the questioned utterance. Red curve: Proportion of likelihood ratios for test data known to be from speaker B with values greater than or equal to value on the x axis. Blue curve: Proportion of likelihood ratios for test data known to be from speaker A with values less than or equal to value on the x axis.

Two questioned utterances, one from speaker A and one from speaker B, were randomly selected from sections of the recordings in the *near* condition which were not used in the training data. For each of these, the cross-validation calibration procedure was repeated 1000 times, resulting in 1000 likelihood-ratio estimates and the spread of these estimates was examined using a histogram. Figure 2 shows a segment of the Tippett plot given above aligned with a histogram indicating reliability of likelihood-ratio estimates for the questioned utterances selected from speaker A (*magenta*) and from speaker B (*cyan*).

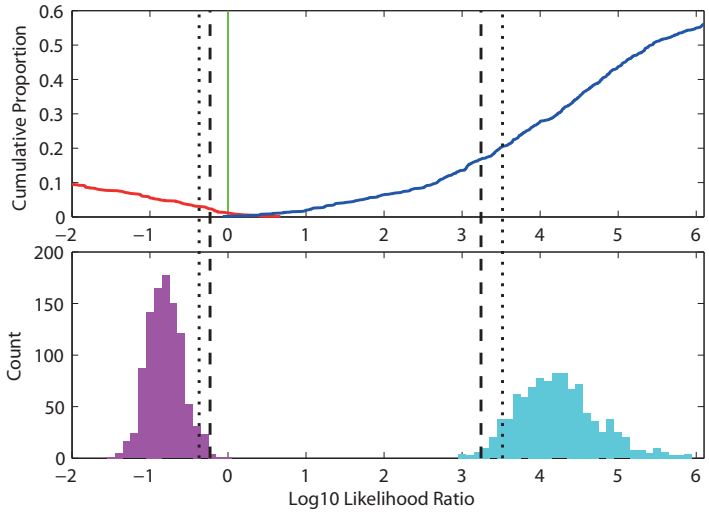

FIGURE 2: Tippett plot and histogram results for the questioned utterances. Both displayed on the same x-axis scale. Speaker A (cyan): The vertical dashed line is at the 1st percentile of likelihood-ratio-estimate values (LR = 1749) and the dotted vertical line is at the 5th percentile (LR = 3294). Speaker B (magenta): The vertical dotted line is at the 95th percentile (LR = 1.7) and the vertical dashed line is at the 99th percentile of likelihood-ratio-estimate values (LR = 2.35).

For the questioned utterance taken from speaker A, we could say that given the results of our analysis we are 99% certain the the probability of getting the acoustic properties of this recording is at least ap-

proximately 1750 times more likely had it been spoken by speaker A versus had it been spoken by speaker B. In tests of our system, none of the 1000 test utterances produced by speaker B had a likelihood ratio as high as this (the highest value obtained was 35.4).

For the questioned utterance taken from speaker B, we could say that given the results of our analysis we are 99% certain the the probability of getting the acoustic properties of this recording is at least approximately 1.7 times more likely had it been spoken by speaker B versus had it been spoken by speaker A. In tests of our system, none of the 1000 test utterances produced by speaker B had a likelihood ratio as high as this (the highest value obtained was 0.7).

Conclusions

A procedure for testing the validity and reliability of a forensic-voice-comparison system under conditions faced in a forensic-voice-comparison analysis was described. Tests were made using recordings simulating a difference in distances to a mobile telephone's microphone. As this case demonstrates, the conditions faced are often highly specific and require testing of validity and reliability on a per-case basis.

Acknowledgments

This research was supported by the Australian Research Council, Australian Federal Police, New South Wales Police, Queensland Police, National Institute of Forensic Science, Australasian Speech Science and Technology Association, and the Guardia Civil through Linkage Project LP100200142. Unless otherwise explicitly attributed, the opinions expressed are those of the authors and do not necessarily represent the policies or opinions of any of the above mentioned organizations.

References

Bogert, B., Healy, M., and Tukey, J. (**1963**). "The quefrency alanysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking", in *Symposium on Time Series Analysis*, 209–243.

Brümmer, N. (**2005**). "Tools for fusion and calibration of automatic speaker detection systems", URL `http://niko.brummer.googlepages.com/focal`.

Brümmer, N. and du Preez, J. (**2006**). "Application-independent evaluation of speaker detection", Computer Speech and Language, **20**, 230–275.

Ellis, D. P. W. (**2005**). "PLP and RASTA (and MFCC, and inversion) in Matlab", URL `http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/`.

Morrison, G. S. (**2009**). "Robust version of train_llr_fusion.m from Niko Brümmer's FoCaL Toolbox (release 2009-07-02)", URL `http://geoff-morrison.net/#TrainFus`.

Morrison, G. S. (**2011**). "Measuring the validity and reliability of forensic likelihood-ratio systems", Science & Justice, **51**, 91–98, doi:10.1016/j.scijus.2011.03.002.

Morrison, G. S. (**2013**). "Tutorial on logistic-regression calibration and fusion: Converting a score to a likelihood ratio", Australian Journal of Forensic Sciences, doi:10.1080/00450618.2012.733025.

van Leeuwen, D. A. and Brümmer, N. (**2007**). "An introduction to application-independent evaluation of speaker recognition systems", in *Speaker Classification I. Fundamentals, Features, and Methods*, edited by C. Müller, 330–353 (Springer).