Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 70 (2015) 28–41

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Mismatched distances from speakers to telephone in a forensic-voice-comparison case [☆]

Ewald Enzinger [a,b,c,*], Geoffrey Stewart Morrison [a,d]

[a] School of Electrical Engineering & Telecommunications, University of New South Wales, Sydney, Australia
[b] National ICT Australia (NICTA), Australian Technology Park, Sydney, Australia
[c] Acoustics Research Institute, Austrian Academy of Sciences, Vienna, Austria
[d] Department of Linguistics, University of Alberta, Edmonton, Canada

Received 8 September 2014; received in revised form 10 March 2015; accepted 11 March 2015
Available online 19 March 2015

Abstract

In a forensic-voice-comparison case, one speaker (*A*) was standing a short distance away from another speaker (*B*) who was talking on a mobile telephone. Later, speaker *A* moved closer to the telephone. Shortly thereafter, there was a section of speech where the identity of the speaker was in question – the prosecution claiming that it was speaker *A* and the defense claiming it was speaker *B*. All material for training a forensic-voice-comparison system could be extracted from this single recording, but there was a near-far mismatch: Training data for speaker *A* were mostly far, training data for speaker *B* were near, and the disputed speech was near. Based on the conditions of this case we demonstrate a methodology for handling forensic casework using relevant data, quantitative measurements, and statistical models to calculate likelihood ratios. A procedure is described for addressing the degree of validity and reliability of a forensic-voice-comparison system under such conditions. Using a set of development speakers we investigate the effect of mismatched distances to the microphone and demonstrate and assess three methods for compensation.

Keywords: Forensic voice comparison; Likelihood ratio; Validity; Reliability; Distance mismatch; Mismatch compensation

1. Introduction

Although there remain some dissenting voices, there is wide support for the position that the logically correct way for a forensic scientist to evaluate the strength of forensic evidence is using a likelihood ratio (Evett et al., 2011; Berger et al., 2011; Redmayne et al., 2011; Robertson et al., 2011). A likelihood ratio is the probability of the observed evidence if the prosecution hypothesis were true versus if the defense hypothesis were true (Robertson and Vignaux, 1995; Aitken et al., 2010). Over the last half century there have also been calls for forensic-analysis methodologies to be empirically tested under conditions reflecting those found in casework (see Morrison, 2014 for a review). Morrison and Stoel (2014) have also argued in favor of the calculation of forensic likelihood ratios on the basis of relevant data, quantitative measurements, and statistical models. Morrison (2014) has described a paradigm for the evaluation of the strength of forensic evidence consisting of the following components:

1. use of the likelihood-ratio framework

[☆] A preliminary version of portions of this paper was published as Enzinger, E. “Mismatched distances from speakers to telephone in a forensic-voice-comparison case,” Proceedings of the 21st International Congress on Acoustics (ICA), June 2–7, Montréal, Canada (POMA Volume 19, pp. 060039). 10.1121/1.4805425.

[*] Corresponding author at: School of Electrical Engineering & Telecommunications, University of New South Wales, UNSW Sydney, NSW 2052, Australia.

E-mail address: ewald.enzinger@oeaw.ac.at (E. Enzinger).

http://dx.doi.org/10.1016/j.specom.2015.03.001
0167-6393/© 2015 Elsevier B.V. All rights reserved.

2. use of approaches based on data representative of the relevant population, quantitative measurements, and statistical models
3. testing of validity and reliability under conditions reflecting those of the case under investigation.

In this paper we illustrate a methodology for implementing this paradigm based on the conditions of a particular forensic-voice-comparison case: One speaker (speaker *A*) was standing a short distance away from another (speaker *B*) who was holding a mobile telephone through which a call had been established to an emergency call center. Both speakers spoke in a loud voice, and their speech was recorded off the telephone system at the emergency call center.[1] At a particular point in time speaker *A* moved closer to the telephone. Shortly thereafter, there was a short section of the recording where the identity of the speaker was in question – the prosecution claimed that it was speaker *A* and the defense claimed it was speaker *B* (henceforth this section of the recording is referred to as the "questioned utterance").[2] 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* (*prosecution hypothesis*)
- the questioned utterance was spoken by speaker *B* (*defense hypothesis*)

and that this is an exhaustive list of hypotheses, i.e., a priori the probability that the speaker of questioned origin 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 mismatch in the distance from the speakers to the microphone. Data from undisputed utterances produced by speaker *A* that were used for speaker model training were mostly *far*, while those of speaker *B* were *near*, and the questioned utterance was *near*.

Our purpose here is to illustrate how a forensic voice comparison may be conducted under the conditions of this particular case; however, nothing we say should be taken as an explicit or implicit comment about the strength of evidence in the actual case. For this illustration, we used recordings from a research database. We did not use the recording from the actual case. We picked recordings of a pair of speakers from the research database to stand in place of the speakers on the actual casework recording, then processed these recordings to reflect the recording conditions of the case.[3]

We describe how we calculated a likelihood ratio using data from a single pair of speakers, and how we assessed the validity and reliability of the system we used to make this calculation. An initial baseline analysis is conducted without applying any compensation for the mismatch in recordings conditions (distance to microphone) between the training data from the two speakers. We then use additional pairs of speakers to investigate the effect of this mismatch and to test the effectiveness of three compensation strategies:

- adjustment for bias in the likelihood ratio output of the system by shifting log likelihood ratios using an offset estimated from likelihood-ratio values calculated in matched and mismatched conditions,[4]
- mapping feature vectors in the *far* condition to more closely resemble the distribution of those in the *near* condition, and
- transforming features using canonical linear discriminant functions (CLDF), discarding dimensions that are believed to mostly capture variability due to mismatched distances while retaining those believed to mostly capture speaker-specific information.

We then select the most promising of these methods and recalculate the likelihood ratio for the recording of the first pair of speakers.

Copies of the data and the MATLAB (MathWorks Inc., 2013) scripts used to perform the calculations in this paper are available from http://ewaldenzinger.entn.at/nearfar/.

2. Methodology

2.1. Database

Recordings of pairs of male speakers were taken from a database of Australian English voice recordings designed and collected for the purpose of conducting forensic research and casework (Morrison et al., 2015). See Morrison et al. (2012) for details of the data collection protocol. The recordings used were of telephone conversations between pairs of speakers. Each speaker sat in a separate sound booth (IAC 250 Series Mini Sound Shelter) and talked to the other speaker over a telephone.[5] High-quality

[1] Both speakers spoke to each other and to an emergency call center operator at the other end of the telephone connection. Speaker *A*'s comments directed at the operator may have been purely in response to what speaker *B* said; speaker *A*, being relatively far from the telephone, may not have been able to hear what the operator was saying.

[2] There was no confusion between the voice of the female operator versus speakers *A* and *B*, who were both male.

[3] Only telephone transmission and near-far difference were included in this processing, background noise was not added. In the original case the speakers were outside in a quiet environment. Portions of the recordings with any transient background noise or both speakers speaking at once were excluded from analysis. No steady background noise was detected, but any steady background noise would equally have affected both speakers on the original recording.

[4] Calculations were made on training pairs of speakers and applied to test pairs of speakers. Training and test data did not overlap.

[5] This was an intercom system with a handset at each end resembling that of a traditional landline telephone. Note that the microphones used to make the recordings were not attached to this system.