Forensic Science International 267 (2016) 115–124

Contents lists available at ScienceDirect

Forensic Science International

journal homepage: www.elsevier.com/locate/forsciint

Use of relevant data, quantitative measurements, and statistical models to calculate a likelihood ratio for a Chinese forensic voice comparison case involving two sisters

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

[a] School of Criminal Investigation, Southwest University of Political Science & Law, Chongqing, China
[b] Chongqing Institutes of Higher Education Forensic Technology Research Centre, Chongqing, China
[c] Morrison & Enzinger, Independent Forensic Consultants, Vancouver, British Columbia, Canada
[d] Department of Linguistics, University of Alberta, Edmonton, Alberta, Canada

ARTICLE INFO

Article history:
Received 8 April 2016
Received in revised form 2 August 2016
Accepted 8 August 2016
Available online 16 August 2016

Keywords:
Forensic voice comparison
Paradigm
Likelihood ratio
Validity
Reliability
China

ABSTRACT

Currently, the standard approach to forensic voice comparison in China is the aural-spectrographic approach. Internationally, this approach has been the subject of much criticism. The present paper describes what we believe is the first forensic voice comparison analysis presented to a court in China in which a numeric likelihood ratio was calculated using relevant data, quantitative measurements, and statistical models, and in which the validity and reliability of the analytical procedures were empirically tested under conditions reflecting those of the case under investigation. The hypotheses addressed were whether the female speaker on a recording of a mobile telephone conversation was a particular individual, or whether it was that individual's younger sister. Known speaker recordings of both these individuals were recorded using the same mobile telephone as had been used to record the questioned-speaker recording, and customised software was written to perform the acoustic and statistical analyses.

1. Introduction

The aural-spectrographic approach to forensic voice comparison (Kersta [1], Tosi [2], National Research Council [3]) is the standard approach in the People's Republic of China. Guidelines on forensic voice comparison are issued by the Ministry of Justice and by the Ministry of Public Security. The Ministry of Public Security guidelines are based closely on the International Association for Identification's "voice comparison standard" (see Gruber and Poza [4,§57]). The aural-spectrographic approach dates back to the 1960s, and is still popular worldwide (Morrison et al. [5]). It has, however, been much criticised for being based on subjective judgement and not having been empirically tested under casework conditions (Bolt et al. [6,7], Gruber and Poza [4], Meuwly [8,9], Solan and Tiersma [10], Morrison [11]). In the 2003 *Angleton* case [12], it was ruled inadmissible under the US Federal Court's *Daubert* standard [13].

There is a paradigm shift ongoing in forensic science in general (Saks and Koehler [14]) and in forensic voice comparison in particular (Morrison [15]). The new paradigm involves the use of the likelihood ratio framework as the logically correct framework for the evaluation of forensic evidence; calculation of numeric likelihood ratios based on relevant data, quantitative measurements, and statistical models; and empirical testing of validity and reliability of forensic analysis systems under conditions reflecting those of the case under investigation. The European Network of Forensic Science Institutes recommends the use of the likelihood ratio framework, both in general (*Guideline for Evaluative Reporting in Forensic Science* [16]) and specifically for forensic voice comparison (*Methodological guidelines for best practice in forensic semiautomatic and automatic speaker recognition* [17]). For introductions to the likelihood ratio framework, see: Robertson and Vignaux [18], Balding and Steele [19,ch. 1–3 and 11], Rose [20], Morrison [21], the latter two in the context of forensic voice comparison. Procedures based on relevant data, quantitative measurements, and statistical models are more robust to cognitive bias (Found [22]), are transparent and replicable, and are practically easier to test than procedures based on subjective judgement (Morrison and Stoel [23]). The 2009 US National

* Corresponding author.
E-mail address: geoff-morrison@forensic-evaluation.net (G.S. Morrison).

http://dx.doi.org/10.1016/j.forsciint.2016.08.017
0379-0738/

Research Council report on *Strengthening Forensic Science in the United States* [24] was highly critical of many branches of forensics science and recommended "The development and establishment of quantifiable measures of the reliability and accuracy of forensic analyses" (p. 23). The Forensic Science Regulator of England and Wales' *Codes of Practice and Conduct* [25] has mandated that forensic analysis methods and their implementation be validated prior to use. Morrison [11] reviews calls from the 1960s onwards for the validity and reliability of forensic voice comparison to be tested under casework conditions. The use of this paradigm for forensic voice comparison under the conditions of actual cases has previously been illustrated in Enzinger and Morrison [26], Enzinger et al. [27], and Enzinger [28,ch. 4 and 5].

A number of research papers have been published which calculate likelihood ratios on the basis of quantitative measurements and statistical models applied to recordings of Chinese speakers [29–49], and a number of law enforcement agencies in China have purchased commercial forensic voice comparison systems based on such approaches, but as far as we are aware such approaches have not previously been used in casework and presented as evidence in court in China.

The present paper reports on what we believe is the first Chinese court case in which the strength of evidence was evaluated via a forensic voice comparison analysis which calculated a likelihood ratio using relevant data, quantitative measurements, and statistical models. The first-named author is a forensic practitioner who was engaged by the court. The second- and third-named authors acted as consultants to the first-named author. They assisted the first-named author by developing forensic analysis software tailored to the particular circumstances of this case. Below we first describe the circumstances of the case and the competing hypotheses adopted, Sections 2 and 3. We then describe the methodology used to collect relevant data in the form of recordings of the two known speakers designated in the hypotheses, Section 4.1. We then describe the methodology used to make measurements of acoustic properties of the speech on the recordings of the known speakers and of the speech on the recording of the speaker of questioned identity, Sections 4.2 and 4.3. This is followed by a description of the statistical models used to calculate likelihood ratios using these measurements as input, Section 4.4. We then present the results of this analysis, including results of testing the performance of the system prior to actually calculating the likelihood ratio with respect to the questioned speaker recording, Sections 4.5 and 5.

2. Case conditions

The case reported in the present paper was a civil case. The complainant recorded a telephone call using software installed on her mobile telephone (OPPO Electronics Corp. model R809T smartphone, running Android OS4.2, connected to China Mobile's GSM/TD-SCDMA network). The call lasted approximately 25 min and consisted of a conversation in Mandarin between the complainant and a female interlocutor. The speaker of interest in this case is that interlocutor, whom we will refer to as the speaker of questioned identity, or, for brevity, the *questioned speaker*. The complainant stated that she believed that the questioned speaker was the respondent, whom we will designate *speaker A*. The respondent denied being the questioned speaker. She stated that the questioned speaker was her younger sister, whom we will designate *speaker B*.

3. Hypotheses

We adopted the following competing hypotheses, which are mutually exclusive and exhaustive given the circumstances of the case:

Hypothesis A (H_A). The questioned speaker was *speaker A*.

Hypothesis B (H_B). The questioned speaker was *speaker B*.

We defined the evidence in the case, E, as the measured acoustic properties of the speech of the questioned speaker on the mobile telephone recording. The particular type of acoustic measurements being those described in Section 4.3.

Our task was therefore to calculate a likelihood ratio, $\text{LR} = p(E|H_A)/p(E|H_B)$, which would quantify the probability of obtaining the measured acoustic properties on the questioned speaker recording if it were produced by *speaker A* versus the probability of obtaining the measured acoustic properties on the questioned speaker recording if it were produced by *speaker B*.

4. Methodology

4.1. Data collection

In China, it is common to make recordings of suspects specifically in order to perform forensic voice comparison analyses. It is also relatively common in China that these recordings can be made using the same recording device as was used to record the questioned speaker.

In this case, both *speaker A* and *speaker B* were cooperative. Following instructions provided by the first-named author, and as directed by the judge, an officer of the court recorded 5 separate telephone conversations with each speaker over the course of two days (it was not practically possible to obtain a larger number of recordings spread over a longer time period). The court officer used the same telephone hardware and software and the same telephone network as the complainant had used to make the original recording of the questioned speaker. Each recording lasted approximately 10 min.

4.2. Data preparation

All the recordings were converted from AMR format to raw PCM.[1] The PCM recordings were used for all subsequent operations, this format being compatible with our data preparation and acoustic analysis software.

The first-named author (whose first language is Mandarin) manually marked the beginning and end of each utterance made by the respective speaker of interest in the questioned speaker recording, and in each of the known *speaker A* recordings and the known *speaker B* recordings.[2]

4.3. Acoustic analysis

We extracted acoustic information from the speech signals in the form of mel frequency cepstral coefficients (MFCCs, Davis and Mermelstein [50]). We extracted vectors of 14 MFCCs plus deltas (deltas were calculated using MFCC vectors from two before to two after the current MFCC vector, Furui [51]). These are a standard form of measurement in automatic speaker recognition and other speech processing applications (Kinnunen and Li [52], Hansen and Hasan [53]).

We extracted MFCC vectors every 10 ms using 20 ms wide windows. MFCCs were extracted within the frequency range 300–3300 Hz. Care was taken to avoid taking measurements of anything

[1] The conversion was performed using FFmpeg version 2.7.2 (https://www.ffmpeg.org/).

[2] The task was performed using SoundLabeller software (http://geoff-morrison.net/#SndLbl), working with 30 s long sections at a time and combining the result into a single label file for each recording.