Forensic Science International 277 (2017) 30–40

Contents lists available at ScienceDirect

Forensic Science International

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

Empirical test of the performance of an acoustic-phonetic approach to forensic voice comparison under conditions similar to those of a real case

Ewald Enzinger[a,1], Geoffrey Stewart Morrison[a,b,c,*]

[a] School of Electrical Engineering & Telecommunications, University of New South Wales, Sydney, New South Wales, Australia
[b] Independent Forensic Consultant, Vancouver, British Columbia, Canada
[c] Department of Linguistics, University of Alberta, Edmonton, Alberta, Canada

ARTICLE INFO

Article history:
Received 12 November 2016
Received in revised form 18 March 2017
Accepted 10 May 2017
Available online 17 May 2017

Keywords:
Forensic voice comparison
Acoustic-phonetic
Spectrographic
Validity
Reliability
Admissibility

ABSTRACT

In a 2012 case in New South Wales, Australia, the identity of a speaker on several audio recordings was in question. Forensic voice comparison testimony was presented based on an auditory-acoustic-phonetic-spectrographic analysis. No empirical demonstration of the validity and reliability of the analytical methodology was presented. Unlike the admissibility standards in some other jurisdictions (e.g., US Federal Rule of Evidence 702 and the *Daubert* criteria, or England & Wales Criminal Practice Directions 19A), Australia's Unified Evidence Acts do not require demonstration of the validity and reliability of analytical methods and their implementation before testimony based upon them is presented in court. The present paper reports on empirical tests of the performance of an acoustic-phonetic-statistical forensic voice comparison system which exploited the same features as were the focus of the auditory-acoustic-phonetic-spectrographic analysis in the case, i.e., second-formant (F2) trajectories in /o/ tokens and mean fundamental frequency (f0). The tests were conducted under conditions similar to those in the case. The performance of the acoustic-phonetic-statistical system was very poor compared to that of an automatic system.

1. Introduction

In a New South Wales, Australia, case that went to trial in 2012, the defendant was accused of lodging fraudulent tax returns via the Australian Tax Office's automated telephone system. The system verbally asked the caller questions using a synthesized or pre-recorded voice, and used automatic speech recognition to interpret the caller's spoken responses. The system also recorded the outgoing and incoming audio. A suspect was questioned in a police interview room, and that interview was recorded. The suspect was charged and put on trial.

The prosecution instructed a forensic practitioner who performed a forensic voice comparison, produced a written report, and testified in court. The practitioner's analysis was based on a combination of auditory, acoustic-phonetic, and spectrographic approaches (details provided in Section 3 below). The practitioner did not provide an empirical demonstration of the validity and reliability of her approach and its implementation. The defense instructed another forensic practitioner, the second author of the present paper, who provided a written critique of the first practitioner's report and testified in court, but did not analyze the actual audio recordings. During *voir dire* the defense attempted to have the first practitioner's testimony excluded, but it was ruled admissible. Before the jury, the defense argued that the practitioner's testimony should be given no weight since the validity and reliability of her approach and its implementation had not been demonstrated.

In the research study reported in the present paper we empirically test the performance of an acoustic-phonetic-statistical forensic voice comparison system which exploits the same types of acoustic properties that the first forensic practitioner focused on, i.e., second-formant (F2) trajectories in /o/ tokens and mean fundamental frequency (f0). We compare the performance of the acoustic-phonetic-statistical system with that of a standard automatic system, a Gaussian mixture model – universal background model (GMM-UBM) which used mel frequency

* Corresponding author at: Forensic Speech Science Laboratory, Centre for Forensic Linguistics, Aston University, Birmingham, England, United Kingdom.
E-mail address: geoff-morrison@forensic-evaluation.net (G.S. Morrison).
[1] Now at Eduworks, Corvallis, OR, USA.

http://dx.doi.org/10.1016/j.forsciint.2017.05.007
0379-0738/

cepstral coefficients (MFCCs) to measure acoustic properties of the speech. We empirically test both systems under conditions similar to those in the case. The relevant population, speaking styles, and recording conditions of the recordings of speakers of known and questioned identity vary from case to case to the extent that the results of testing a system under the conditions of one case may provide little information as to the performance of that same system under the conditions of another case. We have therefore argued that the validity and reliability of a forensic voice comparison system should be tested on a case-by-case basis [1,2]. When we perform a forensic voice comparison for presentation in court, we make all possible enquiries regarding the recording conditions, and go to all practical lengths to obtain data which are representative of the relevant population and which reflect the speaking styles and the recording conditions in the case. For the current research activity, however, we do not go to the same lengths. Instead, we simulate conditions which are broadly similar to those in the case, and rather than collect new data which would more closely reflect the conditions of the case, we make the best use we can of speaker recordings from an existing database. The tests of validity and reliability are therefore conducted under conditions which are forensically realistic and similar to those in the case, but not exactly the same.

We proceed by first discussing legal admissibility (Section 2). We then describe and critique the testimony provided by the practitioner (Section 3). We then describe the acoustic-phonetic-statistical and automatic systems, the methodology for testing, and the test results (Sections 4–5). We end with discussion and conclusion (Section 6).

2. Admissibility

The aural-spectrographic approach to forensic voice comparison has been in use since the 1960s, but has been highly controversial. For reviews, see [1–9]. From the beginning, a major objection from the scientific community was that the validity and reliability of the approach had not been empirically demonstrated under casework conditions [10,11]. Worldwide, however, the approach is still very popular. A recent INTERPOL survey of law enforcement agencies found it to be the second most popular approach, after the auditory-acoustic-phonetic approach [12].

In the United States, until the 1990s, testimony based on the aural-spectrographic approach was admitted by about 60% of courts and rejected by about 40% [13]. Following the publication of a National Research Council report [3] in 1979, the FBI continued to use the aural-spectrographic approach for investigative purposes (until 2011), but, as a matter of policy, no longer presented court testimony based on this approach. The number of cases in which testimony based on the aural-spectrographic approach was presented in court by others gradually declined. In *Angleton*[1] in 2003 following an admissibility hearing under Federal Rule of Evidence (FRE) 702 and *Daubert*,[2] the aural-spectrographic approach was ruled inadmissible. *Daubert* explained that "The subject of an expert's testimony must be 'scientific . . . knowledge.' The adjective 'scientific' implies a grounding in the methods and procedures of science. Similarly, the word 'knowledge' connotes more than subjective belief or unsupported speculation." It also stated that "Ordinarily, a key question to be answered in determining whether a theory or technique is scientific knowledge that will assist the trier of fact will be whether it can be (and has been) tested." Key criteria for admissibility under FRE 702 (amended in 2000 in light of *Daubert*) include that "(b) the testimony is based on sufficient facts or data; (c) the testimony is the product of reliable principles and methods; and (d) the expert has reliably applied the principles and methods to the facts of the case."[3] The court in *Angleton* found that "The potential rate of error of the aural spectrographic method is unknown and may vary considerably, depending on the conditions of the particular application." "The evidence and testimony show that there is great dispute among researchers and the few practitioners in the field over the accuracy and reliability of voice spectrographic analysis to determine the identity of recorded speakers. . . . The post-*Daubert* case law casts doubt on the reliability and admissibility of voice spectrograph analysis." "[The practitioner's] testimony is unreliable under Rule 702. He is applying a technique that, in general, lacks the reliability necessary for admission under Rule 702. . . . [His] testimony does not meet the standards necessary for admission. It is properly excluded as unhelpful and confusing to the jury." Based on published rulings, testimony based on the aural-spectrographic approach does not appear to have survived a *Daubert* challenge since then. For a more thorough review of admissibility of forensic voice comparison under FRE 702 and *Daubert* (and under *Frye*[4]) see [2].

Admissibility of expert testimony under Australia's Uniform Evidence Acts (UEA)[5] requires that an expert witness have "specialized knowledge based on his or her training, study or experience", but does not require any demonstration of the validity and reliability of their analytical approach and its implementation. Predating the introduction of the New South Wales UEA, the aural-spectrographic approach was ruled admissible in *Gilmore* in 1977.[6] The decision in *Gilmore* was based in substantial part on the fact that in the early to mid 1970s the spectrographic method had been ruled admissible by a number of courts in the US. Notwithstanding US courts' subsequent rejection of the aural-spectrographic approach, the stated reason for its admission in the 2012 New South Wales case was that it had been ruled admissible 35 years earlier in *Gilmore*.

3. Auditory-acoustic-phonetic-spectrographic forensic voice comparison

The practitioner's approach to forensic voice comparison in the 2012 case was based on a combination of auditory, acoustic-phonetic, and spectrographic analyses, which focused on the features outlined below.

There were a large number of /o/ tokens in the recordings of the speaker of questioned identity (hereafter the *questioned-speaker recording*) because many of the automated telephone system's questions resulted in responses which were the word "no". The practitioner cited research literature [14] describing an ongoing sound change in Australian English in which an innovative pronunciation of /o/, i.e., something approaching [oi], is produced by a small proportion of speakers, mainly females under age 30. The practitioner stated that she heard this variant of /o/ in both the known-speaker recording (the recording of the police interview with the defendant) and questioned-speaker recording. The

[1] *United States v Robert N. Angleton*, 269 F.Supp. 2nd 892 (S.D. Tex. 2003).

[2] *William Daubert et al. v Merrell Dow Pharmaceuticals Inc.*, 509 US 579 (1993). In 2014 in England & Wales guidelines were introduced including admissibility criteria that are similar to FRE 702 – *Daubert*. The current version appears in section 19A of *Criminal Practice Directions* [2015] EWCA Crim 1567 Consolidated with Amendment No. 2 [2016] EWCA Crim 1714.

[3] *Daubert* explains that "In a case involving scientific evidence, *evidentiary reliability* will be based upon *scientific validity*." Emphasis in original.

[4] *Frye v. United States*, 293 F. 1013 (D.C.Cir.1923).

[5] *Evidence Act 1995* (Commonwealth of Australia), *Evidence Act 2011* (Australian Capital Territory), *Evidence Act 1995* (New South Wales), *Evidence Act 2001* (Tasmania), *Evidence Act 2008* (Victoria).

[6] *R v Gilmore* [1977,2 NSWLR 935].