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*], 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

[*] Now at Eduworks, Corvallis, OR, USA.

[**] Corresponding author. Now at Forensic Speech Science Laboratory, Centre for Forensic Linguistics, Aston University, Birmingham, England, United Kingdom. E-mail address: geoff-morrison@forensic-evaluation.net

Enzinger, E., Morrison, G.S. (2017). Empirical test of the performance of an acoustic-phonetic approach to forensic voice comparison under conditions similar to those of a real case. *Forensic Science International*, 277, 30–40.
http://dx.doi.org/10.1016/j.forsciint.2017.05.007

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

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.

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

Highlights

- Aural-spectrographic forensic voice comparison testimony was presented in court.
- The analysis focused on F2 trajectories and mean f0 in /o/ tokens.
- No demonstration of validity and reliability was provided.
- We empirically tested an acoustic-phonetic analysis based on the same features.
- Under conditions similar to those of the case, the performance was very poor.

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 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

[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.

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 [13] 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

[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]

recording of the police interview with the defendant) and questioned-speaker recording. The practitioner made spectrograms of /o/ tokens from both the known- and questioned-speaker recordings. Fig. 1 shows spectrograms of some /o/ tokens taken from a recording of a speaker in our database, the style of the figure reflects that of figures included in the practitioner's report (no axis labels or scales were provided). The practitioner stated that "As the known sample is spontaneous conversational speech, a spectrogram of a sample of 'no' utterances is far more variable than the lodgment calls, but a trained eye can readily discern that many utterances show the rising F2 associated with this pronunciation." The practitioner's report did not include any spectrograms of /o/ tokens from the known-speaker recording, and did not include any quantitative measurements of second formant trajectories. The practitioner did not provide results of any empirical tests to demonstrate her ability, based on subjective aural and visual judgement, to class speakers as a member of the group with this innovative pronunciation, or to identify individual speakers either in general or within this group.

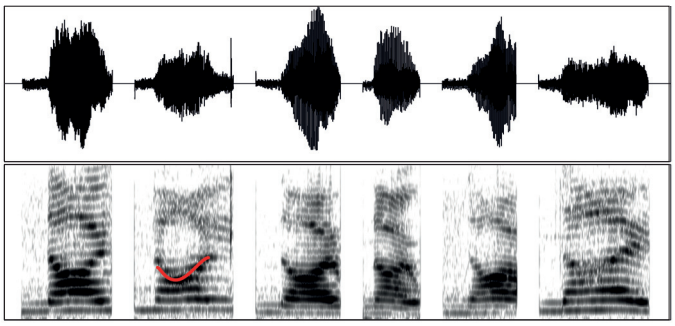

Fig. 1. Example of the style of plot presented by the practitioner to show second formant trajectories.

The practitioner reported that the fundamental frequency of the questioned-speaker recordings had "a highly consistent pitch range, at around 210Hz, slightly above average for female speakers" and "All callers show an average pitch around 210-220Hz, which is normal to slightly high for a female". No documentation or data were provided to support the statement regarding the average f0 for female speakers, and the relevant population was not refined beyond "female speakers". The practitioner provided a figure similar in style to Fig. 2 showing spectrograms of /o/ tokens from the questioned-

speaker recordings and f0 tracks (our plot shows /o/ tokens taken from a speaker in our database). No axis labels or scales were provided, but we presume the frequency scale for the spectrogram was ten times that for the f0 tracks. The horizontal dashed line indicated 210 Hz. The practitioner stated that "The voice in the known sample is of mid to high pitch, averaging somewhat over 200Hz", but provided no plots or exact measurements of the f0 of this speaker. The practitioner also stated that "The pitch is similar in both samples (though it is more variable in the known sample than the questioned sample, this is readily explained by differences in the style of language)", but also admitted that f0 "is not of itself a reliable identifier". If we accept her statements regarding average f0 in the relevant population, then the probability of finding the f0 values observed for the known-speaker and the questioned-speaker recordings would be quite high if these were two speakers selected at random from the relevant population. Thus we would surmise that the evidence was about equally likely given either a same-speaker or a different-speaker hypothesis.

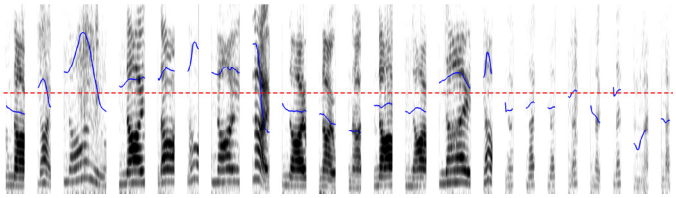

Fig. 2. Example of the style of plot presented by the practitioner to show fundamental frequency.

The practitioner explicitly eschewed the use of likelihood ratios in favor of what she called "a plain language conclusion". Her conclusion was:

> There is ... much to support a hypothesis that [the defendant] is the speaker in the nine lodgment calls. Specifically there are substantial similarities in auditory voice quality, pitch and the pronunciation of a particular vowel heard frequently in both samples. It is important to emphasise that the analysis presented here does not of itself 'prove' that the known and questioned samples originated from the same speaker. The present report must be evaluated by the court in light of other evidence which is rightly not available

> to me. As a guide, I would suggest that this speech evidence can be seen as offering substantial support to any other evidence available to the court that indicates the lodgment calls were made by [the defendant]. If there is no such other evidence available to the court, the speech evidence can be seen as suggesting the lodgment calls were made by [the defendant]. (Emphasis in original)

Part of the logic of the likelihood ratio framework is that a forensic practitioner should evaluate only the strength of the particular piece of evidence they are asked to examine, and that this evaluation should be independent of any other evidence in the case. Hence, the resulting statement of the strength of evidence should not be dependent on any other evidence in the case [7],[9],[15]–[18]. In contrast to this logic, the practitioner's conclusion as to the strength of the evidence was apparently dependent on the strength of other evidence presented in the trial, and changed depending on the strength of the other evidence.

The practitioner's conclusion as to the strength of evidence appears to us to be based entirely and directly on her subjective judgement, and she presented no results of tests that would indicate how good she was in making such judgments under conditions reflecting those of the case (or under any conditions). There was also no evidence that she took any steps to reduce the potential for cognitive bias [19].

Under conditions involving telephone transmission, the performance of acoustic-phonetic-statistical systems based on formant measurements has been found to be poor [20],[21]. Under conditions involving telephone transmission, the performance of acoustic-phonetic-statistical systems based on f0 measurements has also been found to be poor [21], and even with studio-quality recordings performance can be poor [22],[23]. To shed light on whether the practitioner's approach could potentially have had a reasonable degree of validity and reliability under the conditions of the 2012 case, below we empirically test the performance of an acoustic-phonetic-statistical system under conditions similar to those in the 2012 case, and based on the same features as the practitioner focused on, i.e., second-formant (F2) trajectories in /o/ tokens and mean fundamental frequency (f0).

4. Methodology

We work in a paradigm which includes the calculation of likelihood ratios on the basis of relevant data, quantitative measurements, and statistical models, and empirical testing of system performance under conditions reflecting those of the case under investigation. Many specialists in the field of forensic inference and statistics consider the likelihood ratio framework to be the logically correct framework for the evaluation of forensic evidence [24]–[33]. Procedures based on relevant data, quantitative measurements and statistical models are transparent and replicable, and are resistant to cognitive bias if the output of the statistical model is directly reported as the strength of evidence [34]. Empirical testing of validity and reliability is the only way to demonstrate how well a forensic analysis system actually works [1],[15],[35]–[37]. For examples of forensic voice comparison conducted within this paradigm, see [38]–[41]. The analysis below is intended to constitute another example.

4.1. Definition of hypotheses

Based on the circumstances of the case, as described above, we evaluate the strength of the evidence as a likelihood ratio which answers the following two-part question:

- What is the probability of the evidence given the prosecution hypothesis, i.e., what is the probability of getting the measured acoustic properties of the voice on the questioned-speaker recording if the speaker on that recording were the defendant?

versus

- What is the probability of the evidence given the defense hypothesis, i.e., what is the probability of getting the measured acoustic properties of the voice on the questioned-speaker recording if the speaker on that recording were not the defendant but some other speaker selected at random from the relevant population?

These hypotheses are mutually exclusive and, within the circumstances of the case, exhaustive. In this case we would not expect any reasonable disagreement with the observation that the speaker on the questioned-speaker recording was an adult female Australian English speaker. This is a fact that would

be obvious to the trier of fact. We therefore restricted the relevant population to adult female Australian English speakers. We assume here that the speaker was arrested and charged on the basis of information not related to the properties of her voice, and hence do not further restrict the population.

Note that we have not restricted the population to female Australian English speakers who produce the innovative variant of /o/ mentioned by the practitioner. Any strength of evidence which may be associated with whether a speaker produces this variant is not overtly assessed, but it may potentially be accounted for as part of the general acoustic analysis, which includes an analysis of formant trajectories.

4.2. Samples of the relevant population

In order to train and test the forensic voice comparison systems, we need recordings from a sample of speakers representative of the relevant population. These recordings also have to reflect the speaking styles and recording conditions of the questioned-speaker and known-speaker recordings in the case. Recordings of 136 adult female Australian English speakers were taken from a database of voice recordings [42]. The database was designed and collected specifically for the purpose of conducting forensic research and casework (see [43] for details of the data collection protocol). The database included speakers from multiple geographical areas in Australia, who spoke with a range of broad to general Australian accents. High-quality audio recordings of the speakers were made at 44.1 kHz sampling frequency 16 bit quantization. Most speakers were recorded in multiple non-contemporaneous sessions (separated by one to two weeks). In each session they produced recordings in multiple different speaking styles. Recordings of an information-exchange task conducted over the telephone were used as the starting point for simulating questioned-speaker-condition recordings, and recordings from a simulated police interview task were used as the starting point for simulating known-speaker-condition recordings.

4.3. Questioned-speaker conditions

In the case proper, there were actually multiple recordings in which the identity of the speaker was in question, but for simplicity we focus on one representative questioned-speaker recording. The questioned-speaker recording was of a telephone call – for the current research we assume that this was

a mobile telephone (Global System for Mobile communication, GSM) to landline connection. The recording provided for analysis was single-channel PCM with a sampling frequency of 8 kHz and 16 bit quantization.

Sections during which the questioned speaker was speaking were manually located and extracted from the questioned-speaker recording. Sections of background noise when no one was speaking were also extracted. This resulted in 198 sections corresponding to the questioned-speaker utterances with a total duration of 76 s, and 356 sections of background noise with a total duration of 275 s.

We manually located and marked the beginning and end of all "no" tokens in the questioned-speaker recording. There were 15 "no" tokens.

4.4. Simulation of questioned-speaker conditions

The procedures we use for recording condition simulation are broadly similar to those previously used in [39]. The description below is somewhat abridged, see [39] for additional details.

High-quality recordings from the information exchange task in the database were used as the starting point for simulating questioned-speaker-condition recordings. Of the tasks in our database, this task had the most similar speaking style to the speaking style of the questioned-speaker recording in the case. If we were actually to perform a forensic analysis for submission to the court in this case, we would prefer to collect new data which reflected conditions even more similar to those of the case. We would have speakers make telephone calls and interact with an automated system. There may be substantial differences in the speaking style used for interacting with a machine compared to interacting with another human. In addition, if we wanted to specifically analyze "no" tokens we would design the task so as to elicit lots of "no" tokens.

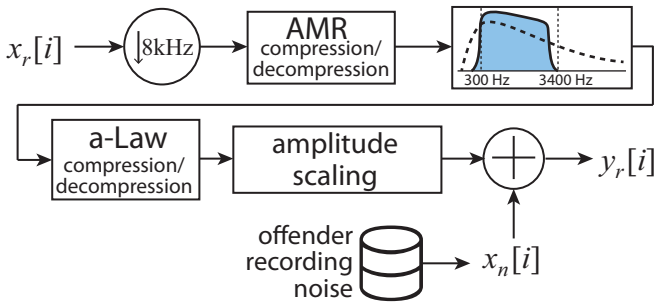

Fig. 3. Procedure for simulating the questioned-speaker recording condition.

We simulated the recording condition using the following procedure (see also Fig. 3):

1. The speech signals were downsampled to a sampling frequency of 8 kHz.

2. Transmission over a mobile-telephone network was simulated. We first simulated the handset characteristic with a linear-phase finite impulse response (FIR) filter [44]. To simulate the effects of the adaptive multi-rate narrow-band (AMR-NB) speech codec [45] used in mobile-telephone transmission, we use the reference fixed-point software implementation [46]. The AMR-NB codec specifies eight modes operating at bit rates between 4.75 and 12.20 kbit/s. Telephone networks can adapt transmission parameters such as the bit rate allocated to speech every two speech frames (40 ms), depending on the quality of the wireless transmission channel [47]. When a call is established, a maximum of four AMR codec modes are selected as the active codec set, beginning with the second-lowest codec mode. In half-rate operation, modes in the active codec set are selected from the five lower-bitrate modes (4.75–7.40 kbit/s), in full-rate operation from all eight modes. Neither the sequence of AMR codec modes used in the mobile-telephone transmission of the actual questioned-speaker recording nor general statistics about the usage of modes were available. For the purpose of simulating the mobile-telephone transmission we selected four modes as the active codec set in full-rate operation. These modes are denoted *A*, *B*, *C*, and *D*, corresponding to bit rates of 5.15, 6.70, 7.95, and 12.20 kbit/s. These modes cover low, mid, and high bit rates. Using a Markov chain (graphically represented in Fig. 4) we randomly generated a sequence of AMR codec modes with predefined transition

probabilities. The transition probabilities were set to enable simulation of dynamic bit rate adaptation over relatively short time spans. The generated mode sequence is then used by the software implementation, operating in discontinuous transmission (DTX) mode, to encode and subsequently decode the speech signal.[7]

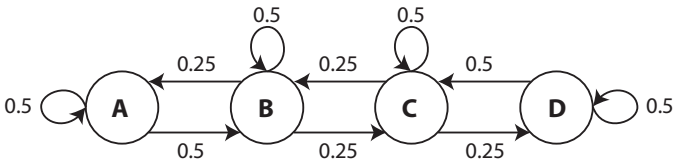

Fig. 4. Markov chain used to generate a sequence of AMR codec modes with predefined transition probabilities.

3. A filter simulating the landline-telephone bandpass characteristic (approximately 300–3400 Hz [51]) was applied using an implementation provided by the "potsband" function of the Voicebox toolbox [52] in MATLAB.

4. The signal was compressed and subsequently decompressed using the a-law algorithm [53] used in landline telephone systems in Australia. The implementations were those in the "lin2pcma" and the "pcma2lin" functions of the Voicebox toolbox.

5. The amplitude of the resulting signal was then adjusted to match the amplitude of the questioned-speaker recording.

[7] Prior published studies on the effect of the AMR codec on acoustic measurements in forensic voice comparison did not vary the bit rate over the course of one recording [48],[49] or used randomly generated mode sequences using a uniform distribution [50].

6. Sections of background noise taken from the original questioned-speaker recording were concatenated in random order (with an onset and offset ramp to avoid transients) and added to each questioned-speaker-condition recording.

An audio recording illustrating the results of adding each step of questioned-speaker-condition simulation is provided at: https://entn.at/implcase/

4.5. Known-speaker conditions

The known-speaker recording was obtained during a police interview at a police station. For the current research we assume that the recording system was a standard model used by New South Wales Police: DHC DVD TripleDeck Interview Recorder – Australian Police Model by David Horn Communications Ltd, Luton, UK. This system records a stereo audio signal at a sampling frequency of 48 kHz 20 bit quantization per channel and saves the audio in a compressed format using the MPEG-1 layer 2 (MP2) standard [54] at 256 kbit/s bitrate and 16 bit output precision. The microphones are Knowles Acoustics MB6052ASC-1 electret condenser microphones.

For the known-speaker recording from the actual case, sections during which the known-speaker was speaking were manually located and extracted, as were sections containing background noise when nobody was speaking. This resulted in 1281 sections corresponding to the known speaker's utterances with a total duration of 28 min, and 293 sections of background noise with a total duration of 573 s.

We manually located and marked the beginning and end of all "no" tokens in the known-speaker recording. There were 107 "no" tokens.

4.6. Simulation of known-speaker conditions

The procedures used for recording condition simulation are broadly similar to those previously used in [39]. The description below is somewhat abridged, see [39] for additional details.

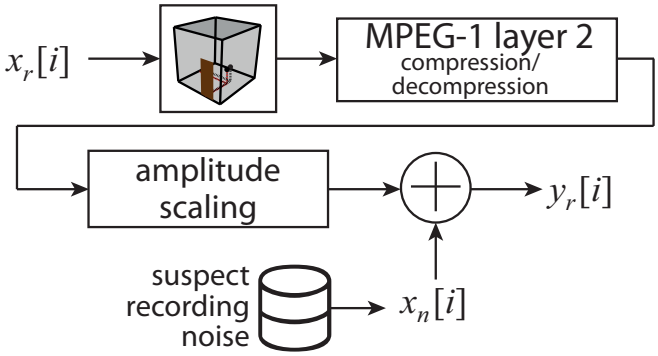

Fig. 5. Procedure for simulating the known-speaker recording condition.

High-quality recordings from the interview task in the database were used as the starting point for simulating known-speaker-condition recordings. We simulated the recording condition using the following procedures (see also Fig. 5):

1. The speech signals from the database were convolved with an estimate of the impulse response of the room in which the known-speaker recording had been made. The police interview had been conducted in a reverberant room; however, the specific characteristics of the room were not known. We estimated the reverberation time (RT60)[8] from the speech sections of the known-speaker recording using the estimation algorithm described in [55] as implemented in [56].[9] Then, using an implementation of the image source model [58] we simulated the room impulse response. We

[8] RT60 is the time taken for sound intensity in a room to drop by 60 dB after a continuous sound is abruptly switched off.

[9] Blind reverberation time estimation is a continuously developing research area. The procedures described here were used to simulate realistic recording conditions that are representative of those on the suspect recording in the case. The effect of the procedure used on the validity and reliability of forensic voice comparison relative to actually recording in the relevant reverberant environment is an open research question which is not addressed in the present paper. If we were preparing a report to submit to the court and had access to the interview room, we would prefer to actually measure the impulse response of the room [57].

specified room dimensions and microphone and speaker positions as the same as we used in a previous case (see [39]). The resulting impulse response had the frequency by reverberation time properties shown in Fig. 6.

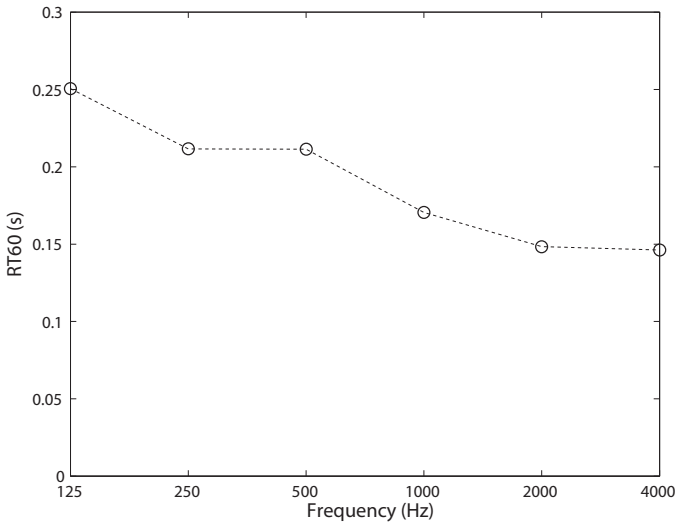

Fig. 6. Frequency by reverberation-time properties of the simulated room impulse response (RT60).

2. The reverberation-degraded recordings were then encoded and decoded at 256 kbit/s using a software implementation [59],[60] of the MPEG-1 layer 2 standard.

3. The same procedures for amplitude adjustment and addition of background noise as were applied to the questioned-speaker-condition recordings were applied to the known-speaker-condition recordings, matching signal amplitude to the original known-speaker recording and adding noise taken from the original known-speaker recording.

An audio recording illustrating the results of adding each step of known-speaker-condition simulation is provided at: https://entn.at/implcase/

4.7. Acoustic-phonetic-statistical forensic voice comparison system

4.7.1. Acoustic analysis

The practitioner's analysis focused on second formant (F2) trajectories and mean fundamental frequency (f0) in realizations of /o/ in tokens of the word "no".

Tokens of /o/ in the word "no" were manually located in the known- and questioned-speaker-condition database recordings, and the start and end points of those tokens marked using SOUNDLABELLER [61].

There were between 3 and 72 tokens (median 36) in each known-speaker-condition recording, and between 0 and 13 tokens (median 3) in each questioned-speaker-condition recording (7510 tokens total across all known-speaker-condition recordings, and 515 tokens total across all questioned-speaker condition recordings). The number of "no" tokens in the database recordings was lower than in the known- and questioned-speaker recordings in the case, which had 107 and 15 "no" tokens respectively. As previously mentioned, if performing casework employing this acoustic-phonetic-statistical approach we would collect data specifically for this case including speaking tasks specifically designed to elicit a large number of "no" responses. For the present research, we proceed using the tokens available in the database recordings.

We made human-supervised measurements of the mean f0 and of the F2 trajectory in each /o/ token. The measurements were made using FORMANTMEASURER [62]. See [20] for details of the measurement system and procedures. We fitted discrete cosine transforms (DCTs) to the F2 trajectories (for an overview of this approach, see [63]). DCTs were fitted to trajectories with frequency in hertz, and time on an equalized-duration timescale (the measured points for each token were linearly interpolated to a set of 126 points, see [64]). The zeroth through third DCT coefficients were used in subsequent analyses. This resulted in a feature vector [f0-mean, F2-DCT0, F2-DCT1, F2-DCT2, F2-DCT3] describing the properties of each /o/ token.

4.7.2. Division of data into training and test sets

83 of the 136 speakers from the database either had only one recording session each or did not have any "no" tokens in their questioned-speaker-condition recording. Since same-speaker comparisons for these speakers could not be constructed, data from these speakers could not be used for testing. The

known-speaker-condition recordings of these speakers were therefore used as training data. The remaining 53 speakers had at least two recordings and some /o/ tokens in their questioned-speaker-condition recording. Known-speaker-condition and questioned-speaker-condition recordings from these speakers were used as test data.

4.7.3. Likelihood ratio calculation – MVKD

Likelihood ratios were calculated using Aitken & Lucy's multivariate kernel density formula (MVKD) [65],[66], as implemented in [67]. The likelihood ratio output of the MVKD formula was treated as a score which had to be converted to a likelihood ratio. Score to likelihood ratio conversion was performed using logistic regression [68]–[71]. A regularization coefficient of 0.001 was used for robust training of the logistic regression model [72]. Training was performed using data from the test set, but with cross validation to avoid training and testing on the same data. Cross validation was leave out all data from the relevant one speaker for a same-speaker comparison, and leave out all data from the relevant two speakers for a different-speaker comparison.

4.8. Automatic forensic voice comparison system

4.8.1. Acoustic analysis

Mel frequency cepstral coefficients (MFCCs [73]) were extracted every 10 ms over the entire speech-active portion of each recording using a 20 ms wide hamming window. Vectors consisting of the 1st through 14th MFCC coefficient values and their deltas [74] were used for subsequent modelling. Care was taken such that measurements did not extend into portions of the recording during which the speaker of interest was not speaking, see [39] for details.

The mismatch-compensation techniques, feature warping [75] and probabilistic feature mapping [76],[77], were applied to the MFCC+delta vectors.

4.8.2. Division of data into training and test sets

The same division of 83 speakers in the training set and 53 speakers in the test set as had been used for the acoustic-phonetic-statistical system was used for the automatic systems. The training set

consisted only of known-speaker-condition data, and the test set of both known-speaker-condition and questioned-speaker-condition data.

4.8.3. Likelihood ratio calculation – GMM-UBM

Details of this system and procedures are identical to those described in [39]. A background model was trained from data of speakers in the background set. To avoid the background model being biased towards speakers with longer recordings, the number of MFCC+delta vectors used from each speaker in the training set was 11,618 (the number available from the shortest recording in the training set). 11,618 vectors were randomly selected from each speaker.

From each known-speaker-condition recording in the test set 12,367 vectors were used (this was the number available from the shortest known-speaker-condition recording in the test set). 12,367 vectors were randomly selected from each recording.

3,522 MFCC+delta vectors were extracted from the questioned-speaker recording. 3,522 contiguous vectors were selected from the beginning of each questioned-speaker-condition recording in the test set.

Scores were calculated using a Gaussian mixture model universal background model (GMM-UBM [78]) with 512 Gaussian components. Scores were then converted to likelihood ratios using logistic regression with cross-validated training on the test set.

4.9. Fused system

A system combining the output of the acoustic-phonetic-statistical system and the automatic system was created. Logistic regression was used to fuse the scores from the two systems [79],[80]. The same cross-validation procedure as was used in training the score to likelihood ratio mapping for individual systems was also used for the fused system.

4.10. System testing

For all speakers in the test set, each speaker's Session 1 questioned-speaker-condition recording was compared with their own Session 2 known-speaker-condition recording, and with their Session 3 known-speaker-condition recording if one was available. These were same-speaker comparisons. Each

speaker's Session 1 questioned-speaker-condition recording was also compared with every other speaker's Session 1, 2, and 3 known-speaker-condition recordings, as available. These were different-speaker comparisons. Likelihood ratios were calculated for 86 same-speaker comparisons and 4628 different-speaker comparisons.

The log likelihood-ratio cost (mean procedure, C_{llr}-mean) was calculated as a measure of validity of the system, and the 95% credible interval (95% CI, parametric procedure) as a measure of reliability. The procedures were as described in [81]. C_{llr}-pooled was also calculated as a single measure conflating the effects of both validity and reliability [68],[82]. Results are presented graphically using Tippett plots (see [9],[83]) and detection error tradeoff plots (DET [84]). DET plots were obtained using the Receiver Operator Characteristic Convex Hull method [85].

5. Results

The results of testing the validity and reliability of the acoustic-phonetic-statistical system, automatic system, and fused system are graphically represented in Figs. 7 through 9. Fig. 7 plots C_{llr} and 95% CI values. Tippett plots are provided in Fig. 8, and DET plots in Fig. 9.

The tests of the acoustic-phonetic-statistical system resulted in a C_{llr}-mean value of 0.842 and a 95% CI of ±0.451 orders of magnitude. The C_{llr}-pooled value was 0.834. For the automatic system, the corresponding values were C_{llr}-mean = 0.332, 95% CI = ±0.606, and C_{llr}-pooled = 0.401. For the fused system they were C_{llr}-mean = 0.307, 95% CI = ±0.757, and C_{llr}-pooled = 0.360.

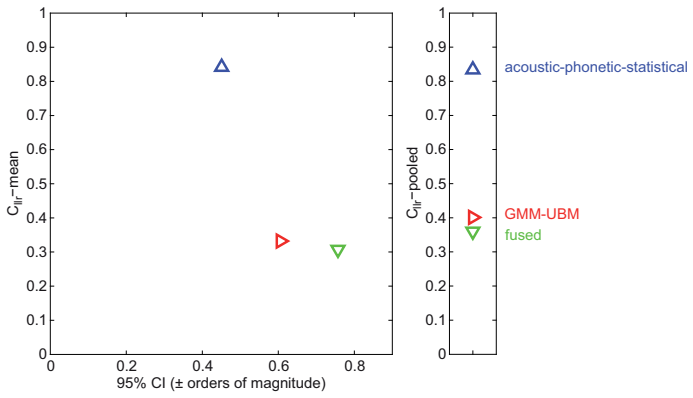

Fig. 7. Measures of validity and reliability of the tested forensic voice comparison systems.

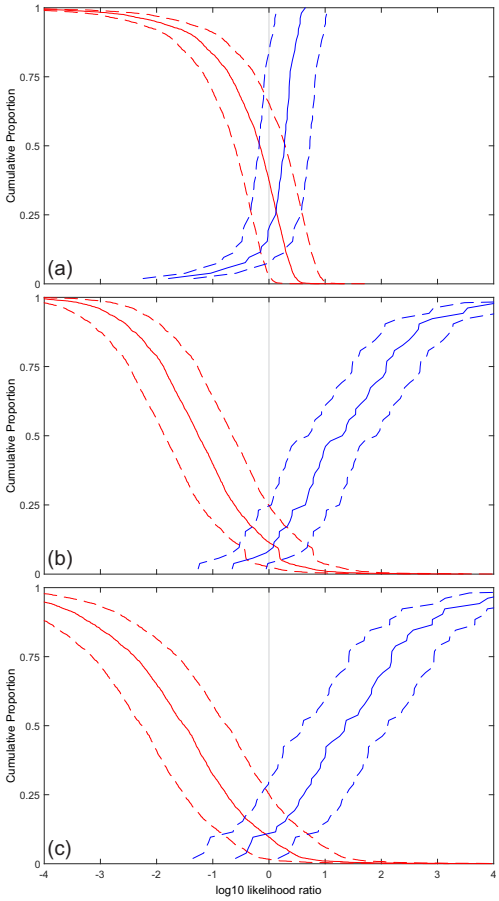

Fig. 8. Tippett plots of the results of system tests. (a) acoustic-phonetic-statistical system, (b) GMM-UBM system, (c) fusion of both systems. Solid lines increasing to the left represent the cumulative proportion of different-speaker comparisons in the test set with mean $\log_{10}(\mathrm{LR})$ values equal to or greater than the value indicated on the *x*-axis. Solid lines increasing to the right represent the cumulative proportion of same-speaker comparisons in the test set with mean $\log_{10}(\mathrm{LR})$ values equal to or less than the value indicated on the *x*-axis. Dashed lines to the left and right of the solid lines represent the 95% credible intervals.

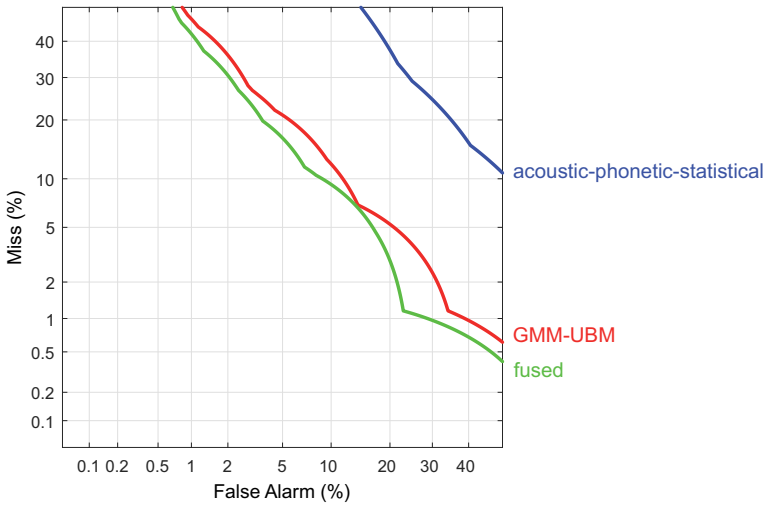

Fig. 9. Detection error tradeoff (DET) plot comparing the performance of the tested forensic voice comparison systems.

6. Discussion & Conclusion

We tested the performance of an acoustic-phonetic-statistical system and an automatic system under forensically realistic conditions similar to those of a real case. The acoustic-phonetic-statistical system used second formant trajectory and mean fundamental frequency measurements from tokens of /o/ in the word "no". These features were the same features that a forensic practitioner had focused on in performing an auditory-acoustic-phonetic-spectrographic analysis for the case. The performance of this system was very poor. Its C_{llr}-pooled value was 0.834. A system which always responded with a likelihood ratio of 1, and therefore gave no information to assist the trier of fact in making their decision, would result in a C_{llr}-pooled value of 1. The performance of the automatic system was vastly superior to that of the acoustic-phonetic-statistical system. The fusion of the two systems gave some improvement in performance over the automatic system alone (a 10% improvement in C_{llr}-pooled).

Despite its poor performance, the acoustic-phonetic-statistical system was a very expensive system to implement since it required substantial human time and expertise to locate and mark the beginning and end of each of 8,025 /o/ tokens and to conduct human-supervised measurements of their F2 trajectories and mean f0. The use of this system also presented other problems: Our database actually included recordings from 242 female Australian English speakers, but we only used recordings of 136

speakers. This was because of the expense of locating and measuring the /o/ tokens. In contrast, the automatic system could make use of all these recordings at negligible additional cost. An automatic system trained on the original recordings from 83 speakers plus recordings from an additional 106 speakers, and tested on the same test set as before, had a C_{llr}-pooled value of 0.363, a 9.5% improvement compared to the original automatic system. About the same level of improvement as resulted from fusing the original automatic system with the acoustic-phonetic-statistical system, but at much less cost in skilled human labor.

In the research reported in the present paper, we did not use as many "no" tokens for training and testing as were available in the original known- and questioned-speaker recording. This was because the recordings in our database did not contain as many "no" tokens. As previously stated, if we were actually preparing a report for presentation in court, and were using this acoustic-phonetic-statistical approach, we would collect new data in which there would be guaranteed to be a large number of "no" tokens. Using a larger number of tokens may have resulted in better performance from the acoustic-phonetic-statistical system, but at substantial additional cost in additional data collection and additional skilled human labor. Taking both performance and cost into consideration, we think it would be more efficient simply to use the automatic system trained on the larger amount of already available data.

Given the poor performance of the acoustic-phonetic-statistical system under the tested conditions, if testimony based on this system were tendered in a case with similar conditions, we would not expect it to be ruled admissible if the jurisdiction in which it were tendered required demonstration of a sufficient degree of validity and reliability. We ourselves would not even attempt to tender testimony based on such a poorly performing system. We might consider the performance of the automatic system to be reasonable, but we would expect an opposing side to argue that its degree of validity and reliability is insufficient. This is a matter which should be debated before the judge at an admissibility hearing.

It should be noted that we did not test the forensic practitioner and her implementation of her auditory-acoustic-phonetic-spectrographic approach. The poor performance of our acoustic-phonetic-statistical system based on the same features (F2 trajectories and mean f0 in tokens of /o/) may suggest that her performance would be poor, but ultimately we do not know what it would be. That, however, is

exactly the problem: We do not know what her performance would be, and neither would a judge or jury. This should surely have rendered her testimony inadmissible had it been tendered in a jurisdiction which required demonstration of a sufficient degree of validity and reliability.

The present study only tested one set of conditions based on those of one forensic case. We do not know how well the systems we tested here would perform under other realistic casework conditions. Future research could potentially compare the performance of acoustic-phonetic-statistical and automatic systems under conditions reflecting those of other real forensic cases. Based on our current level of knowledge, however, we expect that the performance of the acoustic-phonetic-statistical system would likely be so poor that it is not worth expending the resources on testing it further. We believe that it would be better to invest resources in testing other systems which *a priori* would be expected to have higher levels of performance.

Statistical systems could incorporate other quantitative measurements of acoustic or other properties of the voices on the recordings under comparison, and could make measurements on tokens of multiple phonemes. Based on the results in [21], however, we do not believe that adding additional phonemes to an acoustic-phonetic system would meaningfully improve performance. If some other type of measurement were used, the validity and reliability of a system making use of those measurements would have to be empirically tested under conditions reflecting those of the case [1],[2]. How well a particular feature works under the conditions of a case is a matter which requires empirical demonstration. The level of performance obtained should also be weighed against the cost of implementing the system [20].

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. The second author would like to thank the Isaac Newton Institute for Mathematical Sciences for its hospitality during the program Probability and Statistics in Forensic Science which was supported by the Engineering and Physical Sciences Research Council (EPSRC) grant number EP/K032208/1. Unless explicitly stated otherwise, all opinions expressed are those of the authors and do not necessarily represent the policies or opinions of any of the organizations with which the authors have been or are currently affiliated.

References

[1] Morrison GS. Distinguishing between forensic science and forensic pseudoscience: testing of validity and reliability, and approaches to forensic voice comparison. Sci. Justice 2014; 54: 245–256. http://dx.doi.org/10.1016/j.scijus.2013.07.004

[2] Morrison GS, Thompson WC. Assessing the admissibility of a new generation of forensic voice comparison testimony. Columbia Sci. and Tech. Law Rev. 2017; 18.

[3] National Research Council. On the theory and practice of voice identification. Washington DC: National Academies Press, 1979.

[4] Gruber JS, Poza FT. Voicegram identification evidence. American Jurisprudence Trials 1995; 54(1): §1–§133.

[5] Meuwly D. Le mythe de l'empreinte vocale I. Revue Internationale de Criminologie et Police Technique 2003; 56: 219–236.

[6] Meuwly D. Le mythe de l'empreinte vocale II. Revue Internationale de Criminologie et Police Technique 2003; 56: 361–374.

[7] Rose P. Forensic Speaker Identification. London, UK: Taylor & Francis, 2002.

[8] Solan LM, Tiersma PM. Hearing voices: speaker identification in court. Hastings Law Journal 2003; 54: 373–435.

[9] Morrison GS. Forensic voice comparison. In: Freckelton I, Selby H (Eds.), Expert Evidence. Sydney, Australia: Thomson Reuters, 2010.

[10] Bolt RA, Cooper FS, David Jr. EE, Denes PB, Pickett JM, Stevens KN. Speaker identification by speech spectrograms: a scientists' view of its reliability for legal purposes. J. Acoust. Soc. Amer. 1970; 47: 597–612. http://dx.doi.org/10.1121/1.1911935

[11] Bolt RA, Cooper FS, David Jr. EE, Denes PB, Pickett JM, Stevens KN. Speaker identification by speech spectrograms: some further observations. J. Acoust. Soc. Amer. 1973; 54: 531–534. http://dx.doi.org/10.1121/1.1913613

[12] Morrison GS, Sahito FH, Jardine G, Djokic D, Clavet S, Berghs S, Goemans Dorny C. INTERPOL survey of the use of speaker identification by law enforcement agencies. Forensic Sci. Intl. 2016; 263: 92–100. http://dx.doi.org/10.1016/j.forsciint.2016.03.044

[13] Faigman DL, Blumenthal JA, Cheng EK, Mnookin JL, Murphy EE, Sanders J. Talker identification: I. Legal Issues. In Faigman DL, Saks MJ, Sanders J, Cheng EK (eds.) Modern scientific evidence: The law and science of expert testimony (2015/2016 Ed.). Thomson West. 2015, Vol. 5, Ch. 36, §37.1–37.3.

[14] Cox F. The Acoustic Characteristics of /hVd/ Vowels in the Speech of Some Australian Teenagers. Aus. J. Ling. 2006; 26(2): 147–179. http://dx.doi.org/10.1080/07268600600885494

[15] Evett IW. Interpretation: a personal odyssey. In: Aitken GGG, Stoney DA (eds.), The use of statistics in forensic science. Chichester, UK: Ellis Horwood, 1991; 9–22.

[16] Robertson B, Vignaux GA. Interpreting Evidence. Chichester, UK: Wiley and Sons, 1995.

[17] Buckleton JS. A framework for interpreting evidence. In: Buckleton JS, Triggs CM, Walsh SJ (eds.), Forensic DNA evidence interpretation. Boca Raton, FL: CRC, 2005; 27–63.

[18] Balding DJ, Steele CD. Weight-of-evidence for Forensic DNA Profiles (2nd ed.). Chichester, UK: Wiley and Sons, 2015. http://dx.doi.org/10.1002/9781118814512

[19] Found B. Deciphering the human condition: the rise of cognitive forensics. Aus. J. Forensic Sci. 2015; 47(4): 386–401. http://dx.doi.org/0.1080/00450618.2014.965204

[20] Zhang C, Morrison GS, Enzinger E, Ochoa F. Effects of telephone transmission on the performance of formant-trajectory-based forensic voice comparison – female voices. Speech Comm. 2013; 55: 796–813, http://dx.doi.org/10.1016/j.specom.2013.01.011

[21] Zhang C, Enzinger E. Fusion of multiple formant-trajectory- and fundamental-frequency-based forensic-voice-comparison systems: Chinese /ei1/, /ai2/, and /iau1/. In: Proceedings of the 21st International Congress on Acoustics 2013; Jun 2–7, Montréal, Québec, Canada. 060044. http://dx.doi.org/10.1121/1.4798793

[22] Li J, Rose P. Likelihood ratio-based forensic voice comparison with f-pattern and tonal f0 from the Cantonese /oy/ diphthong. Proc. 14th Australasian Intl. Conf. on Speech Sci. and Tech. 2012; 201–204.

[23] Wang CY, Rose P. Likelihood ratio-based forensic voice comparison with Cantonese /i/ F-pattern and tonal F0. Proc. 14th Australasian Intl. Conf. on Speech Sci. and Tech. 2012; 209–212.

[24] Evett IW, Aitken CGG, Berger CEH, Buckleton JS, Champod C, Curran JM, Dawid AP, Gill P, González-Rodríguez J, Jackson G, Kloosterman A, Lovelock T, Lucy D, Margot P, McKenna L, Meuwly D, Neumann C, NicDaéid N, Nordgaard A, Puch-Solis R, Rasmusson B, Redmayne M, Roberts P, Robertson B, Roux C, Sjerps MJ, Taroni F, Tjin-A-Tsoi T, Vignaux GA, Willis SM, Zadora G. Expressing evaluative opinions: a position statement. Sci. Justice 2011; 51: 1–2. http://dx.doi.org/10.1016/j.scijus.2011.01.002

[25] Berger CEH, Buckleton JS, Champod C, Evett IW, Jackson G. Evidence evaluation: a response to the Court of Appeal judgment in R v T. Sci. Justice 2011; 51: 43–49. http://dx.doi.org/10.1016/j.scijus.2011.03.005.

[26] Robertson B, Vignaux GA, Berger CEH. Extending the confusion about Bayes. Mod. Law Rev. 2011; 74: 444–455.

[27] Redmayne M, Roberts P, Aitken CGG, Jackson G. Forensic science evidence in question. Crim. Law Rev. 2011; 5: 347–356.

[28] Morrison GS. The likelihood-ratio framework and forensic evidence in court: a response to R v T. Intl. J. Evidence and Proof 2012; 16: 1–29. http://dx.doi.org/10.1350/ijep.2012.16.1.390

[29] Thompson WC. Discussion paper: Hard cases make bad law—reactions to R v T. Law, Prob. Risk 2012; 11(4):347–359.

[30] Curran JM. Is forensic science the last bastion of resistance against statistics? Sci. Justice 2013; 53: 251–252. http://dx.doi.org/10.1016/j.scijus.2013.07.001

[31] Willis SM, McKenna L, McDermott S, O'Donnell G, Barrett A, Rasmusson B, Höglund T, Nordgaard A, Berger CEH, Sjerps MJ, Lucena Molina JJ, Zadora G, Aitken CGG, Lovelock

T, Lunt L, Champod C, Biedermann A, Hicks TN, Taroni F. ENFSI Guideline for Evaluative Reporting in Forensic Science, European Network of Forensic Science Institutes, 2015. (URL: http://enfsi.eu/sites/default/files/documents/external_publications/m1_guideline.pdf)

[32] Drygajlo A, Jessen M, Gfroerer S, Wagner I, Vermeulen J, Niemi T. Methodological guidelines for best practice in forensic semiautomatic and automatic speaker recognition. European Network of Forensic Science Institutes, 2015. (URL: http://www.enfsi.eu/sites/default/files/documents/guidelines_fasr_and_fsasr_0.pdf)

[33] Morrison GS, Kaye DH, Balding DJ, Taylor D, Dawid P, Aitken CGG, Gittelson S, Zadora G, Robertson B, Willis SM, Pope S, Neil M, Martire KA, Hepler A, Gill RD, Jamieson A, de Zoete J, Ostrum RB, & Caliebe A. A comment on the PCAST report: Skip the "match"/"non-match" stage Forensic Sci. Intl. 2017; 272: e7–e9. http://dx.doi.org/10.1016/j.forsciint.2016.10.018

[34] Morrison GS, Stoel RD. Forensic strength of evidence statements should preferably be likelihood ratios calculated using relevant data, quantitative measurements, and statistical models — a response to Lennard (2013) Fingerprint identification: how far have we come? Aust. J. Forensic Sci. 2014; 46: 282–292. http://dx.doi.org/10.1080/00450618.2013.833648

[35] National Research Council. Strengthening Forensic Science in the United States: A Path Forward. Washington, DC: National Academies Press, 2009. (URL: http://www.nap.edu/catalog.php?record_id=12589)

[36] Forensic Science Regulator. Codes of practice and conduct for forensic science providers and practitioners in the criminal justice system (Version 3.0), Forensic Science Regulator, Birmingham, UK, 2016. (URL: https://www.gov.uk/government/publications/forensic-science-providers-codes-of-practice-and-conduct-2016)

[37] President's Council of Advisors on Science and Technology, Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods, 2016. (URL: https://obamawhitehouse.archives.gov/administration/eop/ostp/pcast/docsreports/)

[38] Enzinger E, Morrison GS. Mismatched distances from speakers to telephone in a forensic-voice-comparison case. Speech Comm. 2015; 70: 28–41. http://dx.doi.org/10.1016/j.specom.2015.03.001

[39] Enzinger E, Morrison GS, Ochoa F. A demonstration of the application of the new paradigm for the evaluation of forensic evidence under conditions reflecting those of a real forensic-voice-comparison case. Sci. Justice 2016; 56: 42–57. http://dx.doi.org/10.1016/j.scijus.2015.06.005

[40] Zhang C, Morrison GS, Enzinger E, Ochoa FE. Effects of telephone transmission on the performance of formant-trajectory-based forensic voice comparison – female voices. Speech Comm. 2013; 55: 796–813. http://dx.doi.org/10.1016/j.specom.2013.01.011

[41] Zhang C, Morrison GS, Enzinger E. Use of relevant data, quantitative measurements, and statistical models to calculate a likelihood ratio for a Chinese forensic voice comparison case involving two sisters. Forensic Sci. Intl. 2016; 267: 115–124. http://dx.doi.org/10.1016/j.forsciint.2016.08.017

[42] Morrison GS, Zhang C, Enzinger E, Ochoa F, Bleach D, Johnson M, Folkes BK, De Souza S, Cummins N, Chow D. Forensic Database of Voice Recordings of 500+ Australian English speakers, 2015. (URL: http://databases.forensic-voice-comparison.net/)

[43] Morrison GS, Rose P, Zhang C. Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice. Aust. J. Forensic Sci. 2012; 44: 155–167. http://dx.doi.org/10.1080/00450618.2011.630412

[44] International Telecommunication Union. ITU-T Recommendation P.48 (11/88): Specification for an intermediate reference system, 1988. URL: http://www.itu.int/rec/T-REC-P.48-198811-I

[45] 3rd Generation Partnership Project (3GPP). ETSI TS 126 090 Mandatory Speech Codec speech processing functions; Adaptive Multi-Rate (AMR) speech codec; Transcoding functions, 2012. (URL: http://www.etsi.org/deliver/etsi_ts/126000_126099/126090/11.00.00_60/ts_126090v110000p.pdf)

[46] 3rd Generation Partnership Project (3GPP). ETSI TS 126 073 ANSI C code for the Adaptive Multi Rate (AMR) speech codec, 2012. (URL: http://www.etsi.org/deliver/etsi_ts/126000_126099/126073/11.00.00_60/ts_126073v110000p.pdf)

[47] 3rd Generation Partnership Project (3GPP). ETSI TS 45 009 Digital cellular telecommunications system (Phase 2+); Link adaptation. (URL: http://www.etsi.org/deliver/etsi_ts/145000_145099/145009/11.00.00_60/ts_145009v110000p.pdf)

[48] Guillemin BJ, Watson C. Impact of the GSM mobile phone network on the speech signal – some preliminary findings. Int. J. of Speech, Lang. and the Law 2008; 15(2): 193–218. http://dx.doi.org/10.1558/ijsll.v15i2.193

[49] Alzqhoul EAS, Nair BBT, Guillemin BJ. Comparison between speech parameters for forensic voice comparison using mobile phone speech. In: Proceedings of the 15th Australasian International Conference on Speech Science and Technology; 2014 Dec 3–5; Christchurch, New Zealand. 28–31.

[50] Alzqhoul EAS, Nair BBT, Guillemin BJ. Impact of dynamic rate coding aspects of mobile phone networks on forensic voice comparison. Sci Justice 2015; 55(5): 363–74. http://dx.doi.org/10.1016/j.scijus.2015.04.006

[51] International Telecommunication Union. ITU-T Recommendation G.151 (11/88): General performance objectives applicable to all modern international circuits and national extension circuits, 1988. (URL: http://www.itu.int/rec/T-REC-G.151-198811-W/en)

[52] Brookes M. VOICEBOX: Speech Processing Toolbox for MATLAB [computer program]. Department of Electrical & Electronic Engineering, Imperial College, London, UK. 1998. (URL: http://www.ee.ic.ac.uk/hp/staff/drb/voicebox/voicebox.html)

[53] International Telecommunication Union. ITU-T Recommendation G.711 (11/88): Pulse code modulation (PCM) of voice frequencies, 1988. (URL: http://www.itu.int/rec/T-REC-G.711-198811-I/en)

[54] International Organization for Standardization (ISO). ISO/IEC International Standard 11172-3:1993 – Information technology – Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s – Part 3: Audio. 1993.

[55] Löllmann HW, Yilmaz E, Jeub M, Vary P. An improved algorithm for blind reverberation time estimation. In: Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC) 2010; Tel Aviv, Israel.1–4. (URL: http://www.iwaenc.org/proceedings/2010/HTML/Uploads/1076.pdf).

[56] Jeub M. Blind Reverberation Time Estimation, version 1.1. Mathworks File Exchange, File ID: #35740, 2010. (URL: http://www.mathworks.com/matlabcentral/fileexchange/35740-blind-reverberation-time-estimation)

[57] Farina A. Impulse response measurements. In: Proceedings of the 23rd Nordic Sound Symposium, Bolkesjø, Norway. 2007; 27–30. (URL: http://www.angelofarina.it/Public/Papers/238-NordicSound2007.pdf)

[58] Lehmann EA, Johansson AM. Prediction of energy decay in room impulse responses simulated with an image-source model. J. Acoust. Soc. Amer. 2008; 124: 269–77. http://dx.doi.org/10.1121/ 1.2936367

[59] TwoLame, software version 0.3.13. (URL: http://www.twolame.org/)

[60] MPEG Audio Decoder (MAD), software version 0.15.1. (URL: http://www.underbit.com/products/mad/)

[61] Morrison GS. Soundlabeller: Ergonomically designed software for marking and labelling portions of sound files [computer program]. 2010. (URL: http://geoff-rorrison.net/#SndLbl)

[62] Morrison GS, Nearey TM. FormantMeasurer: Software for efficient human-supervised measurement of formant trajectories [computer program]. 2011. (URL: http://geoff-rorrison.net/#FrmMes)

[63] Morrison GS. Vowel Inherent Spectral Change in Forensic Voice Comparison. In: Vowel Inherent Spectral Change. Heidelberg, Germany: Springer-Verlag, 2013, 263–83. http://dx.doi.org/10.1007/978-3-642-14209-3_11

[64] Morrison GS. Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs. J. Acoust. Soc. Am. 2009; 125: 2387–97. http://dx.doi.org/10.1121/1.3081384

[65] Aitken CGG, Lucy D. Evaluation of trace evidence in the form of multivariate data. J. Royal Stat. Soc., Series C (Applied Statistics), 2004; 53: 109–22. http://dx.doi.org/10.1046/j.0035-9254.2003.05271.x

[66] Aitken CGG, Lucy D. Corrigendum: Evaluation of trace evidence in the form of multivariate data. J. Royal Stat. Soc., Series C (Applied Statistics), 2004; 53: 665–6. http://dx.doi.org/10.1111/j.1467-9876.2004.02031.x

[67] Morrison GS. multivar_kernel_LR: Matlab implementation of Aitken & Lucy's (2004) forensic likelihood-ratio software using multi-variate kernel density estimation [computer program]. 2007. (URL: http://geoff-rorrison.net/#MVKD)

[68] Brümmer N, du Preez J. Application-independent evaluation of speaker detection, Computer Speech and Language 2006; 20: 230–275. http://dx.doi.org/10.1016/j.csl.2005.08.001

[69] Ramos-Castro D, González-Rodríguez J, Ortega-Garcia J. Likelihood ratio calibration in a transparent and testable forensic speaker recognition framework. In: Proceedings of IEEE Odyssey 2006: The Speaker and Language Recognition Workshop 2006; Jun 28–30, San Juan, Puerto Rico. 1–8. http://dx.doi.org/10.1109/ODYSSEY.2006.248088.

[70] González-Rodríguez J, Rose P, Ramos D, Toledano DT, Ortega-García J. Emulating DNA: rigorous quantification of evidential weight in transparent and testable forensic speaker recognition. IEEE Trans Audio, Speech, and Language Processing 2007; 15: 2104–2115. http://dx.doi.org/10.1109/TASL.2007.902747.

[71] Morrison GS. Tutorial on logistic-regression calibration and fusion: Converting a score to a likelihood ratio. Aus. J. Forensic. Sci. 2012; 45: 173–97. http://dx.doi.org/10.1080/00450618.2012.733025

[72] Morrison GS. Robust version of train_llr_fusion.m from Niko Brümmer's FoCaL Toolbox [computer program]. 2009. (URL: http://geoff-rorrison.net/#TrainFus)

[73] Davis S, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoustics, Speech Signal Proc. 1980; 28: 357–366. http://dx.doi.org/10.1109/TASSP.1980.1163420

[74] Furui S. Speaker-independent isolated word recognition using dynamic features of speech spectrum. IEEE Trans. Acoust., Speech and Sig. Proc. 1986; 34: 52–9. http://dx.doi.org/10.1109/TASSP.1986.1164788

[75] Pelecanos J, Sridharan S. Feature warping for robust speaker verification. In: Proceedings of Odyssey 2001: The Speaker Recognition Workshop 2001; Jun 18–22, Crete, Greece. 213–18.

[76] Mak MW, Yiu KK, Kung SY. Probabilistic feature-based transformation for speaker verification over telephone networks. Neurocomputing 2007; 71: 137–46. http://dx.doi.org/10.1016/j.neucor.2007.08.003

[77] Reynolds DA. Channel robust speaker verification via feature mapping, in: Proc. IEEE Int. Conf. Acoust., Speech Signal Process. 2003; Apr 6–10, Hong Kong; 53–56. http://dx.doi.org/10.1109/ICASSP.2003.1202292

[78] Reynolds DA, Quatieri TF, Dunn RB. Speaker Verification Using Adapted Gaussian Mixture Models. Digital Signal Process. 2000; 10: 19–41. http://dx.doi.org/10.1006/dspr.1999.0361

[79] Pigeon S, Druyts P, Verlinde P. Applying Logistic Regression to the Fusion of the NIST'99 1-Speaker Submissions. Digital Signal Process. 2000; 10: 237–248. http://dx.doi.org/10.1006/dspr.1999.0358

[80] Brümmer N, Burget L, Černocký J, Glembek O, Grézl F, Karafiát M, van Leeuwen DA, Matějka P, Schwarz P, Strasheim A. Fusion of Heterogeneous Speaker Recognition Systems in the STBU Submission for the NIST Speaker Recognition Evaluation 2006. IEEE Trans. Audio, Speech Lang. Proc. 2007; 15: 2072–2084. http://dx.doi.org/10.1109/TASL.2007.902870

[81] Morrison GS. Measuring the validity and reliability of forensic likelihood-ratio systems. Sci. Justice 2011; 51: 91–98. http://dx.doi.org/10.1016/j.scijus.2011.03.002

[82] van Leeuwen DA, Brümmer N. An Introduction to Application-Independent Evaluation of Speaker Recognition Systems. In: Speaker Classification I. Fundamentals, Features, and

Methods. New York: Springer. 2007(4343); 330–353. http://dx.doi.org/10.1007/978-3-540-74200-5_19

[83] Meuwly D. Reconnaissance de locuteurs en sciences forensiques: l'apport d'une approche automatique. PhD Dissertation, University of Lausanne, Lausanne, Switzerland, 2001.

[84] Martin A, Doddington G, Kamm T, Ordowski M, Przybocki M. The DET Curve in Assessment of Detection Task Performance. In: Proceedings of Eurospeech, 1997, 1895–1898.

[85] Brümmer N. Tools for ROCCH DET Curves. (URL: https://sites.google.com/site/focaltoolkit/rocch)