Forensic Science International 341 (2022) 111499

Contents lists available at ScienceDirect

Forensic Science International

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

Speaker identification in courtroom contexts – Part I: Individual listeners compared to forensic voice comparison based on automatic-speaker-recognition technology

Nabanita Basu [a,1], Agnes S. Bali [b,2], Philip Weber [a,3], Claudia Rosas-Aguilar [a,c,4], Gary Edmond [d,5], Kristy A. Martire [b,6], Geoffrey Stewart Morrison [a,e,*,7]

[a] Forensic Data Science Laboratory, Aston University, Birmingham, UK
[b] School of Psychology, University of New South Wales, Sydney, New South Wales, Australia
[c] Instituto de Lingüística y Literatura, Universidad Austral de Chile, Valdivia, Chile
[d] School of Law, Society & Criminology, University of New South Wales, Sydney, New South Wales, Australia
[e] Forensic Evaluation Ltd, Birmingham, UK

ARTICLE INFO

Article history:
Received 4 August 2022
Received in revised form 4 October 2022
Accepted 12 October 2022
Available online 15 October 2022

Keywords:
Admissibility
Forensic voice comparison
Likelihood ratio
Speaker identification
Validation
x-vector

ABSTRACT

Expert testimony is only admissible in common law if it will potentially assist the trier of fact to make a decision that they would not be able to make unaided. The present paper addresses the question of whether speaker identification by an individual lay listener (such as a judge) would be more or less accurate than the output of a forensic-voice-comparison system that is based on state-of-the-art automatic-speaker-recognition technology. Listeners listen to and make probabilistic judgements on pairs of recordings reflecting the conditions of the questioned- and known-speaker recordings in an actual case. Reflecting different courtroom contexts, listeners with different language backgrounds are tested: Some are familiar with the language and accent spoken, some are familiar with the language but less familiar with the accent, and others are less familiar with the language. Also reflecting different courtroom contexts: In one condition listeners make judgements based only on listening, and in another condition listeners make judgements based on both listening to the recordings and considering the likelihood-ratio values output by the forensic-voice-comparison system.

* Corresponding author at: Forensic Data Science Laboratory, Aston University, Birmingham, UK.
E-mail address: geoff-morrison@forensic-evaluation.net (G.S. Morrison).
[1] ORCID: 0000-0003-2234-2995
[2] ORCID: 0000-0002-0166-0989
[3] ORCID: 0000-0002-3121-9625
[4] ORCID: 0000-0002-8544-7965
[5] ORCID: 0000-0003-2609-7499
[6] ORCID: 0000-0002-5324-0732
[7] ORCID: 0000-0001-8608-8207

https://doi.org/10.1016/j.forsciint.2022.111499

1. Introduction

1.1. Research question

The present paper addresses the question of whether speaker identification[8] by a judge listening alone would be more or less accurate than the output of a forensic-voice-comparison system that is based on state-of-the-art automatic-speaker-recognition technology. This question is important because expert testimony is only admissible in common law if it will potentially assist the trier of fact to make a decision that they would not be able to make unaided. If the trier of fact's speaker identification were equally accurate or more accurate than a forensic-voice-comparison system, then testimony based on the output of that system would not be admissible.

The introduction to the present paper also considers the question of whether speaker identification by a jury listening as a collaborative group would be more or less accurate than the output of a forensic-voice-comparison system that is based on state-of-the-art automatic-speaker-recognition technology, but the present paper does not include empirical research addressing this question. The present paper describes experiments using individual listeners. A future paper will describe experiments using groups of twelve collaborating listeners.

In the remaining sections of the introduction of the present paper:

- we discuss the legal context related to forensic voice comparison conducted by experts and to speaker identification performed by triers of fact (§1.2);
- we describe prior research on speaker identification by lay listeners (§1.3);
- and we outline the empirical research we have conducted in order to address the research question (§1.4).

1.2. Legal context

Identifying a speaker by comparing voices is fairly common in modern criminal proceedings. Indeed, the identification of a speaker is often the main, and sometimes the ultimate, issue confronting the trier of fact – whether jurors or judge(s). In most common law (but also inquisitorial or civil law) systems, resolving the identity of a speaker is left to the trier of fact. Sometimes the trier of fact will be responsible for resolving the question of identity based on their own listening to speech recordings – often comparing lawfully intercepted recordings with recorded police interviews, but occasionally comparing intercepted recordings with a defendant's live speech in court. Sometimes, the trier of fact is presented with opinions about a speaker's identity by police officers or interpreters who have no specialized training in forensic voice comparison.[9] Sometimes, the trier of fact is presented with expert testimony from a forensic practitioner who does have specialized training in forensic voice comparison.[10] Even when expert testimony is presented, the trier of fact may still listen to the recordings and attempt to perform their own speaker identification.[11]

The present paper is concerned with whether speaker identification by individual lay listeners (like a judge), or by groups of lay listeners acting collaboratively (such as in the deliberations of a jury or appellate court), would be more or less accurate than testimony presented by an expert witness who has conducted a forensic voice comparison using state-of-the-art automatic-speaker-recognition technology. This is an important question because expert evidence is only admissible if it is capable of rationally assisting the trier of fact. Unless a forensic-voice-comparison system is more accurate than the trier of fact, testimony based on that system should not be admitted (and in some jurisdictions it is considered irrelevant). In recent years, substantial advances have been made in automatic-speaker-recognition technology, leading to improved performance. Advances have also been made in the application of automatic-speaker-recognition technology as part of forensic-voice-comparison systems. Given these advances, some conventional legal assumptions may require revision.

Historically, courts in common-law jurisdictions have not concerned themselves with the actual performance of either speaker identification by lay listeners or forensic voice comparison by forensic practitioners. In most cases little empirical data, such as the results of tests of human listeners' abilities or the results of validations of forensic-voice-comparison systems, have been introduced to inform admissibility determinations, or cross-examination and judicial directions. Even with the gradual rise of reliability standards for expert opinion evidence following *Daubert*,[12] with respect to forensic voice comparison and speaker identification, most courts have promoted approaches based on what might be characterised as "common sense" and the experience of judges. We present some examples:

In *Bulejcik v The Queen*,[13] the High Court of Australia characterised speaker identification as routine; such that where samples of recorded speech are available, jurors should be entitled to make their own comparisons:

> Recognition of a speaker by the sound of the speaker's voice is a commonplace of human experience. … A person who is not familiar with the voice of a putative speaker may be able ... to recognise the speaker's voice by comparison with an established example of that voice ... There was no reason why, subject to a satisfactory warning, the jury should not have had regard to the sound of the appellant's voice in determining whether the appellant's voice had been recorded on Exhibit D.

Expert testimony based on forensic voice comparison was not proffered in this case.

[8] In the well-established terminology of the research literature on speaker identification and speaker recognition by human listeners, "speaker identification" refers to a situation where a listener who is unfamiliar with the speaker or speakers compares a voice they hear on one occasion (e.g., while a crime is being committed) with a voice that they hear on another occasion (e.g., during a voice lineup) and, based on listening, attempts to determine whether the same speaker was speaking on both occasions. "Speaker identification" also refers to a situation where a listener who is unfamiliar with the speaker or speakers listens to two (or more) voice recordings and, based on listening, attempts to determine whether the same speaker is speaking on both recordings. The latter is the focus of the present paper. "Speaker identification" also refers to a situation in which one voice is recorded (e.g., a recording of a crime being committed) and the other is live (e.g., a defendant speaking in court). "Speaker identification" contrasts with "speaker recognition", which refers to the situation where a listener hears a voice (live or recorded) and states that they recognize the voice as that of a person who is familiar to them (and usually names that person). The present paper reports on speaker-identification research, not on speaker-recognition research.

[9] For a criticism of this practice, see Edmond et al. [1]. Most indirect witnesses called in Australia are translators or police officers involved in investigations relying on telephone intercepts and other covert recordings, e.g.: *Kheir v The Queen* (2014) 43 VR 308; *Tran v The Queen* and *Chang v The Queen* [2016] VSCA 79; *Nguyen v The Queen* [2017] NSWCCA 4; *R v Phan* [2017] SASCFC 70; *Davey v Tasmania* [2020] TASCCA 12. Ready admission of such experience-based testimony – characterised as lay opinion and as "ad hoc expert" opinion – has made recourse to expert forensic-voice-comparison testimony relatively uncommon. In Australia, and elsewhere, liberal admissibility practice relies on cross-examination and judicial directions (and the increasingly remote possibility of rebuttal expert testimony) to identify and effectively convey problems to jurors (and appellate judges) in the context of accusatorial proceedings.

[10] For reviews of the admissibility of forensic voice comparison in US jurisdictions (under both *Daubert* and *Frye*) and in UK jurisdictions (England & Wales and Northern Ireland), see Morrison & Thompson [2] and Morrison [3] respectively. Briefer reviews of admissibility in Australia and in Canada are included in Morrison & Enzinger [4].

[11] For criticism of this practice, see Edmond [5].

[12] *William Daubert* et al. *v Merrell Dow Pharmaceuticals Inc.*, 509 US 579 (1993)

[13] *Bulejcik v The Queen* (1996) 185 CLR 375

In *R v Flynn*,[14] the England & Wales Court of Appeal stated:

> The appellant submits that the judge misdirected the jury by instructing them that they should not attempt to compare the voices heard on the covert recording with the voices of the appellants which they had heard when they gave evidence in the trial. Apart from the decision in *R. v Chenia* [2003] 2 Cr. App. R. 6 (p.83) there is no decision which supports the direction given by the judge. On the contrary, there are passages in other authorities, ..., which suggest that the jury should be permitted to make such a comparison providing the judge directs the jury to listen to the tapes guided by the evidence of the voice recognition witnesses, whether expert or lay listeners.

In this case, practitioners of the auditory-acoustic-phonetic approach to forensic voice comparison had stated that the quality of the recordings was too poor for them to be able to conduct forensic voice comparisons. The poor quality of the recordings was a factor in the Court of Appeal ruling that speaker-recognition/speaker-identification testimony by police officers should not have been admitted at trial. Given such poor-quality recordings, it is curious that the Court of Appeal thought that it was appropriate for the jury to attempt to perform their own speaker identifications.

In *United States v Arce-Lopez*,[15] the defendant sought to have expert testimony based on forensic voice comparison admitted. The court found that:

> the jury is "perfectly well-equipped" to listen to the witnesses testify in court, compare their voices to the voice in the audio recordings, and draw conclusions about whose voice is in the audio recordings. ... Accordingly, this is "not an area in which expert testimony would be helpful to the jury." See *United States v. Salimonu*, 182 F.3d 63, 74 (1st Cir.1999) ... the Court finds that this expert testimony will not "help the trier of fact to understand the evidence or to determine a fact in issue," Fed.R.Evid. 702(a)

The published ruling stated that the proffered forensic-voice-comparison testimony was based on "biometric analysis", but it is unclear from the ruling what approach to forensic voice comparison was actually used or whether any validation results were provided.

More recently, in *R v Dunstan*[16] (an appeal hearing in Ontario, Canada), Morrison appeared as an expert witness and presented the likelihood-ratio output of a forensic-voice-comparison system that was based on automatic-speaker-recognition technology. Morrison's report included the results of an empirical validation of the forensic-voice-comparison system under conditions reflecting those of the case under investigation. Although admissibility *per se* was not an issue in this hearing, during cross-examination, Morrison was asked why the judge could not simply listen to the recordings and make a decision. The cross-examining lawyer relied upon a more-than-a-decade-old study to suggest that the performance of automatic-speaker-recognition systems was not better than human listeners.

In the next subsection, we review prior published research comparing the performance of lay listeners with that of automatic-speaker-recognition systems, and in the remainder of the paper we report new empirical research comparing the performance of individual lay listeners with that of a forensic-voice-comparison system which is based on state-of-the-art automatic-speaker-recognition technology.

1.3. Prior research on speaker identification by lay listeners compared to automatic-speaker-recognition systems

There are a number of published studies that have directly compared speaker identification by lay listeners with the output of automatic-speaker-recognition systems. Many of these studies, however, are outdated: Over the last two decades, there have been substantial advances in automatic-speaker-recognition technology (GMM-UBM-based systems have been replaced by i-vector-based systems, which in turn have been replaced by x-vector-based systems), and each new generation of technology has resulted in substantial improvements in performance.[17] Also, the conditions of the voice recordings used in these studies have seldom reflected the sorts of relatively poor-quality recordings conditions or the sorts of mismatched conditions between questioned-speaker and known-speaker recordings that are commonly encountered in forensic casework (the studies were not necessarily intended to address questions of forensic interest). Also, the studies have usually had each listener listen independently, have then applied a function (a simple function such as mean or mode, or a more complex function such as a calibration model) to the pooled responses from all the listeners, and then compared the output of that function with the output of an automatic-speaker-recognition system. This does not reflect the situation where a judge alone listens to the questioned- and known-speaker recordings, nor the situation where a group of people constituting a jury listen and collaboratively come to a decision. In addition, equal-error rate (EER) has often been used to compare listeners' pooled responses with the output of automatic-speaker-recognition systems. EER obscures potentially poorly calibrated responses: To calculate EER, the classification threshold is shifted to the point where the miss rate equals the false-alarm rate, whereas a system with a pre-determined classification threshold may be biased and produce a higher miss rate than false-alarm rate or *vice versa*. Finally, the use of a categorical-decision framework in these studies is suboptimal for assessing the performance of a forensic-voice-comparison system that outputs a likelihood ratio or of a human listener who expresses degree of confidence in their speaker identification decision – treating a likelihood ratio of 2 the same as a likelihood ratio of 2000, or treating a listener's "maybe" the same as their "very sure", ignores the fact that, in a legal-decision-making context, different likelihood-ratio values or different degrees of confidence would be expected to have different magnitudes of impact on downstream decision making (especially, for listeners, if the decision maker is the listener).

Human Assisted Speaker Recognition (HASR) evaluations were run by the National Institute of Standards and Technology (NIST) in 2010 and 2012 (Greenberg et al. [8]). The HASR evaluations were not intended to reflect forensic casework conditions. Whereas automatic-speaker-recognition systems are routinely tested on tens or hundreds of thousands of test pairs, most participants in the HASR 2010 evaluation only provided responses to a set of 15 test pairs, and not to a larger set of 150 test pairs that was also available. The HASR 2010 recordings were high quality, but the different-speaker test pairs were selected to be challenging: Multiple earlier automatic-speaker-recognition systems had made errors on these pairs, and, in pilot tests, listeners judged them difficult to distinguish (see Greenberg et al. [8] for details). HASR 2012 test pairs were also selected to be challenging.

On the HASR test sets, automatic-speaker-recognition systems outperformed systems based on pooled responses from groups of lay listeners [9–12]. In Ramos et al. [10], after it was calibrated, a system based on pooled listener responses achieved a log-likelihood-ratio-cost (C_{llr}) value of 1, i.e., on average the human-listener system provided no useful information (see §2.7.2 below for an explanation

[14] *R v Flynn and St John* [2008] EWCA Crim 970
[15] *United States v Arce-Lopez* 979 F Supp 2d 228 (D Puerto Rico 2013)
[16] *R v Dunstan* [2018] ONSC 4153

[17] For an overview of the different generations of automatic-speaker-recognition technology that have been used for forensic voice comparison, see Morrison et al. [6]. Morrison & Enzinger [7] compares the results of validations of forensic-voice-comparison systems based on different generations of automatic-speaker-recognition technology.

of the C_{llr} metric). In Matějka et al. [13], for each trial, in addition to being able to listen to the pair of recordings, listeners were provided with the score output by an automatic-speaker-recognition system in response to that pair of recordings. The listeners were familiar with this automatic-speaker-recognition system, and they could take its output into consideration while making their judgement. For only one of the ten listeners was the classification-error rate (CER)[18] better than that of the stand-alone automatic-speaker-recognition system.

In contrast to the performance of lay listeners, forensic practitioners employing auditory-acoustic-phonetic methods had the same CER as an automatic-speaker-recognition system (Schwartz et al. [14]), or a better EER than an automatic-speaker-recognition system (Saeidi & van Leeuwen [15]).

Kahn et al. [9] noted that there was high inter-listener variability for lay listeners: Individual listeners' CERs ranged from 34 % to 56 %. Miss rates ranged from 13 % to 90 %, and false-alarm rates ranged from 10 % to 97 %. Some listeners were biased toward giving same-speaker responses (resulting in fewer misses but more false alarms), and others were biased toward giving different-speaker responses (resulting in fewer false alarms but more misses). Similarly, individual lay listeners' EERs in Ramos et al. [10] ranged from 22 % to 60 %. Large inter-listener variability has often been observed in the broader research literature on speaker identification and speaker recognition by lay listeners.[19]

Similar results have been obtained in studies using sets of voice recordings other than the HASR sets. Not all of the other sets were deliberately selected to be challenging. With few exceptions, automatic-speaker-recognition systems outperformed lay listeners [19–25]. The exceptions primarily occurred in earlier studies using older automatic-speaker-recognition technology (e.g., GMM-UBM systems as opposed to i-vector systems). Even then, in the oldest study (Schmidt-Nielsen & Crystal [19]), although the detailed results needed to make all the necessary comparisons were not presented, it appears that most individual listeners' EERs (as opposed to EERs based on pooled responses) would have been worse than the EERs of the automatic-speaker-recognition systems tested. Park et al. [26] noted that i-vector-based systems outperformed lay listeners for voice recordings of longer durations, but that lay listeners outperformed i-vector-based systems for voice recordings of shorter durations, e.g., less than 10 s. Short questioned-speaker recordings are common in forensic casework. In contrast, in van Dijk et al. [24], although listeners could listen for longer, their average listening time was ~18 s, and the EER for their pooled responses was 27 %, but, using 20 s from each recording (close the human listeners' average listening time), an i-vector-based system's EER was only 7 %. Using only 5 s from each recording, the i-vector-based system's EER was 23 %, i.e., even using short recordings it performed better than the pooled responses of listeners who listened not only for longer but listened for as long as they wanted.

The current state of the art in automatic speaker recognition is based on deep-neural-network embeddings (DNN embeddings) called x-vectors [6,26–31]. Compared to i-vector-based systems, newer x-vector-based systems have been found to have better performance, especially on mismatched conditions and on short voice recordings. A recently-published study, Hughes et al. [32], appears to be the only study so far that has compared speaker identification by lay listeners with an x-vector-based forensic-voice-comparison system.[20] That study elicited (as a number between 0 and 100) listeners' judgements as to: typicality of the questioned-speaker recording, similarity of the questioned- and known-speaker recordings, and posterior probability for same-speaker. Several functions were applied to pooled-listener responses. Some of these functions included divisions of similarity responses by typicality responses, others used the posterior-probability responses, and all included cross-validated calibration using logistic-regression. The x-vector-based system outperformed the human listeners: The best C_{llr} for a function applied to pooled-listener responses was 0.69, but the C_{llr} value for the x-vector-based system was 0.26. Individual listeners' EERs ranged from 13 % to 67 %, but the EER for the x-vector-based system was 4 %. The paper reported that there was no correlation between the listeners' EERs and their self-reported familiarity with the accent spoken by the speakers (the listeners had reported familiarity as a number between 0 and 100). The language and accent was "Standard Southern British English", and the listeners were all from the UK.

1.4. The current research

In the empirical research reported in the present paper, we conduct a series of experiments in which lay listeners are asked to make same-speaker/different-speaker judgements on pairs of recordings that reflect the conditions of an actual forensic case. The pairs of recordings are a subset of those from the *forensic_eval_01* dataset [33], which has previously been used to perform benchmark validations of multiple forensic-voice-comparison systems [34–40]. The language and accent spoken on these recordings is Australian English.

Individual listeners provide probabilistic judgements in response to pairs of recordings consisting of one questioned-speaker-condition recording and one known-speaker-condition recording. The individual-listener experiments are intended to reflect a context where an individual judge listens and makes a judgement.[21]

We compare the individual-listener responses with the likelihood-ratio values output by the E^3 Forensic Speech Science System (E^3FS^3) in response to the same pairs of recordings. E^3FS^3 is a forensic-voice-comparison system that is based on state-of-the-art automatic-speaker-recognition technology [6,40,41].[22]

Reflecting different courtroom scenarios,[23] we conduct experiments with:

[18] To calculate CER, the pre-determined classification threshold of the system is used and the miss rate and the false-alarm rate obtained. CER is the weighted mean of the miss rate and the false-alarm rate. Weighting may be equal for the miss rate and the false-alarm rate, or may be according to the number of same-source and different-source inputs respectively.

[19] Recent reviews of the broader literature appear in Sherrin [16] and Morrison et al. [17], and a recent study that found large between-listener variability in speaker recognition in a legally relevant context is Rosas et al. [18].

[20] Hughes et al. [32] was published after we began our data collection.

[21] In a future paper, we will present the results of experiments in which groups of twelve listeners collaboratively make judgements. The group-of-listeners experiments are intended to reflect the situation where a group of jury members listen and collaboratively make a judgement.

[22] More information about E3FS3 is available from http://forensic-voice-comparison.net/E3FS3/

[23] In *Bulejcik v The Queen* (1996) 185 CLR 375 (already discussed in §1.2) an Australian jury listened to the defendant who had a foreign accent (he was from Yugoslavia [*sic*]) and a recording on which the speaker had a foreign accent, and in *Nguyen v The Queen* [2017] NSWCCA 4 an Australian jury listened to recordings on which the speaker or speakers had a Vietnamese accent (in addition, they heard testimony from an "ad hoc expert"). These are similar to our less-familiar-accent condition. In *Li v The Queen* (2003) 139 A Crim R 281 an Australian jury listened to recordings in Cantonese (in addition, they heard testimony from "ad hoc experts" and from a forensic-voice-comparison practitioner using an experience-based auditory approach – no validation was conducted). In *Tran v The Queen* and *Chang v The Queen* [2016] VSCA 79 and in *R v Phan* [2017] SASCFC 70 Australian juries listened to recordings in Vietnamese (in addition, they heard lay testimony from interpreters). These are similar to our less-familiar-language condition. In *Korgbara v The Queen* (2007) 71 NSWLR 187 an Australian jury listened to recordings that were in Igbo and to the appellant (who could speak Igbo) speaking English in court. We do not test a mismatched familiar-language versus unfamiliar-language condition. Edmond et al. [1] provides commentary on several of these cases.

1. listeners who are familiar with the language and accent spoken on the recordings
2. listeners who are familiar with the language but less familiar with the accent
3. listeners who are less familiar with the language

The three different language backgrounds are:

1. Australian-English listeners
2. North-American-English listeners
3. Spanish-language listeners

In the broader research literature on speaker identification by lay listeners, listeners have been found to perform more poorly when the speakers spoke with accents that were less familiar for the listeners and even more poorly when the speakers spoke languages that were less familiar for the listeners. Recent reviews of the broader literature, including review of the effect of language and accent familiarity, appear in Sherrin [16] and Morrison et al. [17], and a recent review focusing on the effect of language familiarity appears in Perrachione [42].

In addition to the experiments outlined above, in order to assess the effect of participants receiving expert testimony and being able to listen to voice recordings, we conduct an additional experiment. In that experiment, we provide participants with information about the forensic-voice-comparison system, including validation results, and for each recording pair we provided participants with the likelihood-ratio value output by the forensic-voice-comparison system in response to that pair.

2. Methodology

2.1. Ethical approval

Ethical approval for this research was obtained from both the University of New South Wales Human Research Ethics Advisory Panel C: Psychology, and from the Aston Institute for Forensic Linguistics Research Ethics Subcommittee.

2.2. Stimuli

2.2.1. Source

Stimuli were taken from the recordings in the *forensic_eval_01* dataset [33].[24] The *forensic_eval_01* recordings reflect the conditions of the questioned-speaker recording and the known-speaker recording from an actual forensic case. The speakers on the recordings are adult male speakers of Australian English. The questioned-speaker condition reflects a landline-telephone call, with background babble noise, saved using lossy compression. The known-speaker condition reflects an interview recorded in a reverberant room, with background ventilation-system noise. Prior to publication, the recordings were manually diarized, i.e., interlocutor speech, transient noises, and long periods during which the speaker of interest was not speaking were removed. Including remaining short pauses between utterances, the questioned-speaker condition recordings were 46 s long, and the known-speaker-condition recordings were 126 s long.

The *forensic_eval_01* dataset includes a training set and a validation set. Each speaker was recorded on multiple occasions separated from each other by approximately a week or more. The validation set consists of a total of 223 recordings from 61 speakers, 61 questioned-speaker-condition recordings (which always came from the first available recording session) and 162 known-speaker-condition recordings, allowing for the construction of 111 same-speaker pairs of recordings and 9720 different-speaker pairs of recordings (from 3660 pairs of speakers). The *forensic_eval_01* validation protocol in [33] requires a forensic-voice-comparison system to provide a likelihood-ratio value in response to each of these 9831 pairs of recordings.

2.2.2. Subset selection

It is not reasonable to ask human listeners to respond to thousands of pairs of stimuli. For the present research, we therefore selected a subset of 61 pairs of recordings from the *forensic_eval_01* validation set. We initially considered using 122 pairs of stimuli, but pilot tests indicated that this number took too long and was too fatiguing for listeners, so we reduced the number to 61. To shorten the time participants would potentially take to complete each comparison trial, we also reduced the duration of each of the recordings to approximately 15 s (listeners were, however, able to listen to each recording multiple times).

Each speaker in the validation set had one questioned-speaker-condition recording. From each questioned-speaker-condition recording we randomly selected an ~15 s long section of speech. Each speaker had at least two known-speaker-condition recordings. We randomly selected one known-speaker-condition recording from each speaker, and from that recording randomly selected an ~15 s long section of speech. 15 s intervals within each recording were randomly selected from a uniform distribution, with the condition that they did not extend beyond the beginning or end of the recording. A researcher then manually extracted sections of speech that began and ended near the beginning and end of the randomly selected intervals, but made cuts at natural pauses rather than in the middle of words.

2.2.3. Construction of pairs of stimuli

For the individual-listener experiment, we selected:

- 31 same-speaker pairs
  - For each speaker, the same-speaker pair was constructed as the ~15 s long section from that speaker's questioned-speaker-condition recording plus the ~15 s long section from their randomly selected known-speaker-condition recording.
  - A constraint was imposed so that the questioned- and known-speaker-condition recordings did not come from the same recording session.
- 30 different-speaker pairs
  - For each speaker, a different-speaker pair was constructed as the ~15 s long section from that speaker's questioned-speaker-condition recording plus the ~15 s long section from a randomly selected different speaker's randomly selected known-speaker-condition recording.
  - A constraint was imposed so that if a pair consisted of a questioned-speaker-condition recording of speaker *A* and a known-speaker-condition recording of speaker *B*, another pair could not consist of a questioned-speaker-condition recording of speaker *B* and a known-speaker-condition recording of speaker *A*.

If recordings of a speaker were used to construct a same-speaker pair, recordings of that speaker were not also used to construct different-speaker pairs.

A copy of the stimuli used to conduct the experiments is available from https://forensic-voice-comparison.net/speaker-recognition-by-humans/

[24] The database is available from https://forensic-voice-comparison.net/databases/#forensic_eval_01

2.3. Participants (listeners)

Participants were recruited using an online recruitment platform, Prolific.[25] The experiment was advertised as taking up to 2 h to complete, and participants who completed the experiment were paid GBP 21 (the amount recommended by Prolific for 2 h of participant time).

There were three sets of individual listeners defined by language background:

1. Australian-English listeners
   These listeners were familiar with both the accent and language spoken by the speakers on the stimulus recordings.
2. North-American-English listeners
   These listeners were familiar with the language spoken by the speakers on the stimulus recordings but less familiar with the accent.[26]
3. Spanish-language listeners
   These listeners were less familiar with both the language and the accent spoken by the speakers on the stimulus recordings.[27]

The target number of listeners to recruit for each language background was 60.

To be eligible, each participant had to self report that they:

1. were 18 years of age or older
2. were a fluent speaker of English (for language backgrounds 1 and 2) or Spanish (for language-background 3)
3. were currently a resident of one of:
   3.1 Australia
   3.2 United States or Canada
   3.3 Spain, Mexico, or Chile[28]
4. had lived for at least 4 years in their current country of residence, or were a citizen of, one of:
   4.1 Australia
   4.2 United States or Canada
   4.3 Spain, Mexico, Chile, or another predominantly Spanish-speaking country
5. had completed at least an undergraduate degree
6. did not have a diagnosed hearing loss

The sub-criterion for eligibility criteria 3 and 4 had to correspond with the language background. Criteria 2–4 did not require a participant to be a first-language and first-accent speaker of the particular language and accent background, but did require them to be familiar with that language and accent background.

The education criterion was included because the individual-listener experiments are intended to inform us about how an individual judge might perform with respect to speaker identification. Recruiting judges *per se* we considered to be impractical. Judges would be expected to have a relatively high level of education. We therefore recruited participants who had completed at least an undergraduate degree.[29]

Potential participants were directed from Prolific to bespoke experiment software that we developed. Participants accessed the experiment software using a web browser.

Potential participants were first asked questions to determine whether they were eligible. If they were eligible, they were provided with a copy of the informed-consent information. If a participant gave consent, they were asked several demographic questions.

We asked participants their age. They could enter a number or “prefer not to say”. We asked participants their gender. They could enter “male”, “female”, “other”, or “prefer not to say”.

We asked participants what their first language was and what other languages they spoke fluently.

We also asked participants how familiar they were with English in general, and with Australian English in particular. To answer the first question, participants could choose from:

- Extremely familiar: English is my first language, or I currently live in a predominantly English-speaking part of the world and have been here for more than 4 years
- Very familiar: I have lived in a predominantly English-speaking part of the world
- Somewhat familiar: For example, I frequently watch English-language TV programmes, have multiple English-speaking friends, and/or I have visited a predominantly English-speaking part of the world
- Not familiar

To answer the second question, participants could choose from:

- Extremely familiar: Australian English is my first accent and language, or I currently live in Australia and have been here for more than 4 years
- Very familiar: I have lived in Australia
- Somewhat familiar: For example, I frequently watch Australian TV programmes, have multiple Australian friends, and/or I have visited Australia
- Not familiar

We also asked participants:

- In general, how good do you think you are at identifying speakers, i.e., if you hear two voice recordings, how good do you think you are at correctly deciding whether they are recordings of the same speaker or of two different speakers?
- How good do you think you are at identifying adult male Australian-English speakers, i.e., if you hear two voice recordings, how good do you think you are at correctly deciding whether they are recordings of the same adult male Australian-English speaker or of two different adult male Australian-English speakers?
- If you heard a large number of pairs of recordings of adult male Australian-English speakers, what percentage of the pairs do you think you would get “right”, i.e., if they were recordings of the same speaker you would say they were recordings of the same speaker and if they were recordings of different speakers you would say they were recordings of different speakers? Count saying “can’t decide” as incorrect.

To answer each of the first two questions, participants chose a value on a five-point Likert scale which had labels: “very poor”,

[25] https://prolific.co/ Since recruitment and payment of participants was handled by Prolific, we did not have access to participants’ personal identifying information.

[26] We chose North-American-English listeners rather than European-English listeners because there are greater cultural links between Australia and the British Isles than between Australia and the US & Canada. By recruiting North-American-English listeners, we were therefore less likely to recruit listeners who happened to be familiar with Australian English.

[27] The online recruitment platform, Prolific, is entirely in English, so the Spanish-language participants had some degree of familiarity with English.

[28] These particular Spanish-speaking countries were chosen because they happened to be the only Spanish-speaking countries from which Prolific recruits participants.

[29] Requiring an even higher level of education would have made the recruitment pools available on Prolific smaller, and impractically small for the Australian-English pool, which was already by far the smallest pool. Also, unlike in the US & Canada where a professional law degree is a graduate degree, in Australia a professional law degree is an undergraduate degree.

"poor", "neutral", "good", "very good". To answer the third question, participants typed a number between 0 and 100 in a box. The first two questions (but not the third) were repeated at the very end of the experiment after the participants had responded to all the stimulus pairs.

Information about the experiment, including informed-consent text, demographic questions, instruction text, and the text on the experiment screens, was provided in either English or Spanish, depending on the language background of the participant.

2.4. Experiment procedures

A demonstration of the bespoke software used to run the individual-listener experiment is available at https://forensic-voice-comparison.net/speaker-recognition-by-humans/. The software was designed to run on any modern web browser running on any modern operating system on any device, but participants were advised that the software display was optimized for larger screens, e.g., desktops, laptops, and tablets, rather than smartphones, and it was strongly recommended to participants that they not run the experiment on a smartphone.

After completing eligibility questions, providing informed consent, and answering demographic questions, each participant was presented with written instructions explaining the task,[30] plus a sound check to make sure they could hear audio playing on their device. They were instructed to perform the experiment in a quiet place, and were asked whether they were listening using headphones or loudspeakers. They were then presented with a warmup trial. The warmup trial was a different-speaker trial that was identical in form to the experiment trials. Participants were not told that this was a warmup. Their responses to this trial were not included in subsequent analysis. Each participant was then presented with the 61 experiment trials in random order, a different random order for each participant. Randomly mixed with the experiment trials were 4 attention-check trials. We describe the experiment trials and the attention-check trials in the paragraphs below. Each trial screen included a counter showing the number of trials completed out of the total (66 including warmup and attention-check trials). A participant could take a break whenever they wanted. If they closed their browser, they could later resume using the link originally provided by Prolific. On resuming after having closed the browser, a participant had to repeat the sound check, after which the experiment resumed where they had left off. The experiment could not be resumed if more than 7 days had passed since the participant first started the experiment.[31]

A screenshot of an experiment trial is shown in Fig. 1. The participant was presented with two sets of audio-playback controls, one labelled "questioned-speaker recording" and the other labelled "known-speaker recording". Using each set of controls, the participant could start and stop playing the recording, and navigate to any point between the beginning and end of a recording. Only one recording would play at a time.

The participant was also presented with two response boxes. The first response box was embedded in the following sentence:

- I think the properties of the voices on the recordings are _____ times **more likely if they are both recordings of the same adult male Australian-English speaker** than if they are recordings of two different adult male Australian-English speakers.

The second response box was embedded in the following sentence:

- I think the properties of the voices on the recordings are _____ times **more likely if they are recordings of two different adult male Australian-English speakers** than if they are both recordings of the same adult male Australian-English speaker.

Participants were instructed to enter a number that was 1 or greater in one of the boxes. Participants were instructed that if they thought the properties of the voices on the recordings were a little more likely if they were recordings of the same speaker than if they were recordings of different speakers they should enter a number in the first box that is a little larger than 1, and if they thought the properties of the voices on the recordings were a lot more likely if they were recordings of the same speaker than if they were recordings of different speakers they should enter a number in the first box that is a lot larger than 1; and *mutatis mutandis* for the second box if they thought the properties of the voices on the recordings were more likely if they were recordings of different speakers than if they were recordings of the same speaker. The instructions (deliberately) did not suggest any particular numbers to use. Participants were instructed that if they thought the properties of the voices on the recordings were exactly equally likely irrespective of whether they were recordings of the same speaker or recordings of different speakers, they should enter 1 in either one of the boxes.[32]

The software checked that the participant had listened to at least 5 s of each recording, and that they had entered a number 1 or greater in one, but only one, of the boxes. If these criteria were met, when the participant pressed the "next" button, they moved to the next trial. If not all criteria were met, the participant received a message indicating the criterion or criteria which had not been met. Once a participant had moved to the next trial, they could not return to an earlier trial.

In addition to saving the responses entered into the response boxes, for each recording, the software saved the total listening time.

The screen for an attention-check trial looked the same as the screen for an experiment trial, but instead of hearing a pair of questioned-speaker-condition and known-speaker-condition recordings, the participant heard a recording (the same recording on both players) that told them to enter a particular number in one of the boxes.[33] For the English-language listeners, the instructions were spoken in English by a synthetic voice with an Australian accent, and for the Spanish-language listeners they were spoken in Spanish by a synthetic voice with a European-Spanish accent.

After the last pair of recordings, the questions about how good the participant thought they were at speaker identification were repeated, and the participant was presented with a "submit" button. The participant could withdraw from the study at any point before pressing the "submit" button by simply not proceeding with the

[30] The participant could also access the instructions whenever they wanted during the experiment.

[31] Prolific's display of the link timed out after 24 h, and, if participants completed the experiment more then 24 h after they began, Prolific issued a warning. Using Prolific's messaging service, we sent each participant their link, and informed them that they could ignore the warning issued by Prolific.

[32] The intent was to elicit subjectively assigned likelihood-ratio values. The logically correct output for a forensic-evaluation system (including a forensic-voice-comparison system) is a likelihood ratio. In order to compare like with like, we therefore had to attempt to elicit likelihood-ratio values from listeners. It may be that some (or many) listeners did not fully understand the implied request to provide a ratio of likelihoods, and they may instead have provided numbers that represented their "certainty" as to whether the recordings were of the same speaker or of different speakers, but this still provided an unconstrained number (theoretically, on a logarithmic scale, between minus infinity and plus infinity, rather then being constrained to a range such as 0–1 or 0–100) that was a subjectively assigned quantification of the listener's assessment of the strength of the evidence.

[33] For the attention-check trials, the software did not check whether the participant had listened to at least 5 s of each recording.

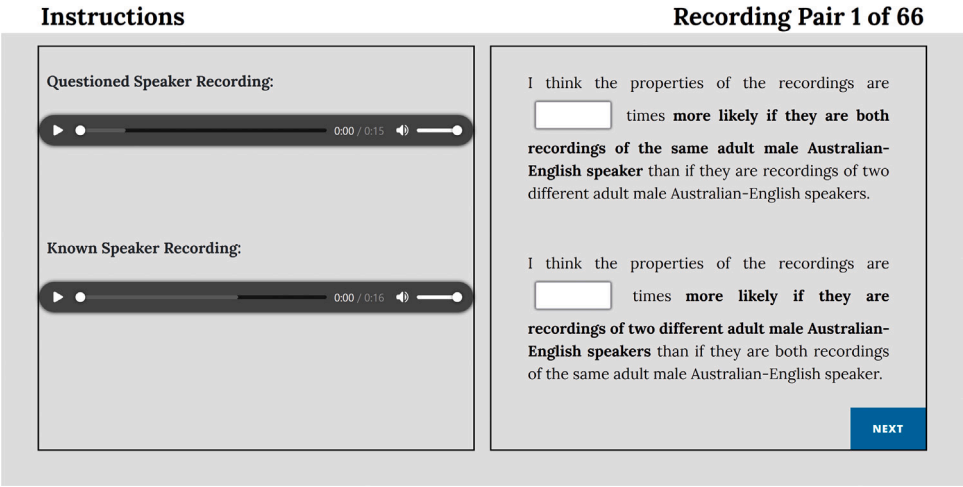

Fig. 1. Screenshot of an experiment trial in the individual-listener experiment.

experiment. If they did not press the "submit" button within 7 days of starting the experiment, their temporarily saved responses were deleted. If the participant pressed the "submit" button within 7 days of starting the experiment, their responses were permanently saved. Since the responses were submitted anonymously, once the "submit" button was pressed, the participant could no longer withdraw their responses.

After each participant submitted their responses, a researcher checked their responses to the attention-check trials and authorized payment if at least two of the four were answered correctly.

2.5. Forensic-voice-comparison system

E^3FS^3 is a forensic-voice-comparison system which is based on state-of-the-art automatic-speaker-recognition technology. It extracts x-vectors using a Residual Network (ResNet). Backend models include linear discriminant analysis (LDA) for mismatch compensation and dimension reduction, probabilistic linear discriminant analysis (PLDA) to calculate uncalibrated likelihood ratios (scores), and logistic regression for calibration.[34] For more detailed descriptions of this system, see Morrison et al. [31] and Weber et al. [41]. For a previous report on the validation of this system, see Weber et al. [40].

Sections of recordings from the *forensic_eval_01* training set were used to train LDA and PLDA. From each recording in the training set, a 15 s long section was randomly selected, and an x-vector was extracted from that section. In addition to the in-domain *forensic_eval_01* data, out-of-domain data from the SRE2018 Test dataset [45] were adapted to the *forensic_eval_01* conditions using the correlation alignment (CORAL) algorithm [46,47], and the in-domain and adapted data were together used to train the LDA and PLDA.

The validation data consisted of the same ~15-long recording-sections as had been used with the human listeners. An x-vector was extracted from each section.

For calibration, all recordings in the *forensic_eval_01* validation set were used. From each recording in the *forensic_eval_01* validation set, three 15 s long non-overlapping sections were randomly selected, and an x-vector was extracted from each section.[35] All possible questioned-speaker-condition versus known-speaker-condition pairs of recording-sections were constructed, excluding same-speaker pairs constructed from different recordings made during the same recording session. Hereinafter, these will be referred to as the calibration data.

Leave-one-speaker-out / leave-two-speakers-out cross validation was employed: In a cross-validation loop in which the score to be calibrated was a same-speaker score, e.g., a recording of speaker *A* compared to another recording of speaker *A* in the validation data, all scores in the calibration data that resulted from comparisons in which one or both members of the pair was a recording of speaker *A* were excluded and the remaining calibration data were used to train the calibration model (leave-one-speaker-out). In a cross-validation loop in which the score to be calibrated was a different-speaker score, e.g., a recording of speaker *A* compared to a recording of speaker *B* in the validation data, all scores in the calibration data that resulted from comparisons in which one or both members of the pair was a recording of speaker *A* or a recording of speaker *B* were excluded and the remaining calibration data were used to train the calibration model (leave-two-speakers-out).

2.6. Experiment in which forensic-voice-comparison results are provided

In order to assess the effect of providing participants with expert testimony on forensic voice comparison and also allowing them to listen to the recordings and perform their own speaker identification, we ran another version of the individual-listener experiment with a new set of North-American-English listeners.[36] In that version, along with the instructions, we provided participants with the information about the forensic-voice-comparison system given in Text Box 1 and in Figure 2.

The text that appeared on the experimental screens had the form of one of the following, as applicable:

- **Output of forensic-voice-comparison system:** The acoustic properties of the questioned-speaker and known-speaker recordings are X times **more likely if they were both produced by the same adult male Australian-English speaker** than if they

[34] A regularized version of logistic regression was used with a regularization weight equivalent to 1 pseudo-speaker relative to the number of speakers used for training the logistic-regression model (see Morrison & Poh [44] for details).

[35] These sections were automatically extracted and were not manually adjusted to not begin or end in the middle of words. Although beginning or ending in the middle of words might be disturbing for human listeners, it is not an issue for the forensic-voice-comparison system.

[36] We used North-American-English listeners because, of the three language backgrounds, they constituted the largest pool of potential participants available on Prolific. Listeners who had participated in the earlier experiment were excluded from participating in this experiment.

Text Box 1
Information that was provided to participants about the forensic-voice-comparison system.

To help you make your decision, for each pair of recordings, we provide the likelihood-ratio output by a forensic-voice-comparison system in response to the same pair of recordings. The output appears to the left of the screen, below the audio players. When deciding the value you think is appropriate to enter into either the first box or the second box, you can take into consideration your own listening of the recordings and you can take into consideration the likelihood-ratio value output by the forensic-voice-comparison system.

The overwhelming majority of experts in forensic inference and statistics agree that the likelihood-ratio framework is the logically correct way for a forensic practitioner to evaluate strength of evidence. Its use is also recommended by key organizations including the American Statistical Association, European Network of Forensic Science Institutes, Forensic Science Regulator for England & Wales, and National Institute of Forensic Science of the Australia New Zealand Policing Advisory Agency.

In the context of forensic voice comparison, a likelihood ratio quantifies:

how much more likely the acoustic properties of the questioned-speaker and known-speaker recordings would be if they were both produced by the same speaker compared to if they were each produced by a different speaker from the relevant population; or

how much more likely the acoustic properties of the questioned-speaker and known-speaker recordings would be if they were each produced by a different speaker from the relevant population compared to if they were both produced by the same speaker.

In this case, the relevant population is adult male speakers of Australian English.

The forensic-voice-comparison system used was the E^3 Forensic Speech Science System (E^3FS^3). This system makes use of state-of-the-art automatic-speaker-recognition technology, which includes the use of deep neural networks. It has been developed by the Forensic Data Science Laboratory at Aston University, in collaboration with the Audio, Digital Intelligence and Speech (AUDIAS) Laboratory at the Autonomous University of Madrid, and in partnership with operational forensic laboratories in several organizations including the FBI, Netherlands Forensic Institute, Swedish National Forensic Center, German Federal Police Office, and Chilean Investigative Police.

The forensic-voice-comparison system has been calibrated and validated under the same conditions as those of the pairs of recordings that you will be asked to make judgments on. Calibration and validation was performed in accordance with the recommendations in the 2021 *Consensus on validation of forensic voice comparison*. To perform the validation, the system was presented with a large number of pairs of recordings that were same-speaker pairs and a large number of pairs of recordings that were different-speaker pairs (999 same-speaker pairs and 87,480 different-speaker pairs), and it gave a likelihood-ratio output in response to each pair. Each pair consisted of one recording in questioned-speaker condition and one recording in known-speaker condition. None of the pairs were the same as those that you will be asked to make judgments on.

Given a same-speaker pair, a good output would be a large likelihood-ratio value in favor of the same-speaker hypothesis, a less good output would be a smaller likelihood-ratio value in favor of the same-speaker hypothesis, a worse output would be a small likelihood-ratio value in favor of the different-speaker hypothesis, and a bad output would be a large likelihood-ratio value in favor of the different-speaker hypothesis.

Given a different-speaker pair, a good output would be a large likelihood-ratio value in favor of the different-speaker hypothesis, a less good output would be a smaller likelihood-ratio value in favor of the different-speaker hypothesis, a worse output would be a small likelihood-ratio value in favor of the same-speaker hypothesis, and a bad output would be a large likelihood-ratio value in favor of the same-speaker hypothesis.

The image below shows the validation results in a Tippett plot. The blue curve rising to the right shows the proportion of same-speaker pairs that had likelihood-ratio values equal to or less than the value on the *x* axis. The red curve rising to the left shows the proportion of different-speaker pairs that had likelihood-ratio values equal to or greater than the value on the *x* axis. The better the performance of the forensic-voice-comparison system the greater the separation between the same-speaker and different-speaker curves: the further to the right the same-speaker curve will be and the further to the left the different-speaker curve will be. The *x* axis of the Tippett plot shows values greater than 1 and values less than 1.

A value greater than 1 favors the same-speaker hypothesis, e.g., a likelihood ratio of 100 means that the acoustic properties of the questioned-speaker and known-speaker recordings are 100 times more likely if they were both produced by the same speaker than if they were each produced by a different speaker from the relevant population.

A value less than 1 favors the different-speaker hypothesis, e.g., a likelihood ratio of 1/100 means that the acoustic properties of the questioned-speaker and known-speaker recordings are 100 times more likely if they were each produced by a different speaker from the relevant population than if they were both produced by the same speaker.

were produced by two different adult male Australian-English speakers.

- **Output of forensic-voice-comparison system:** The acoustic properties of the questioned-speaker and known-speaker recordings are X times **more likely if they were produced by two different adult male Australian-English speakers** than if they were both produced by the same adult male Australian-English speaker.

If the likelihood-ratio value X was greater than 10, it was rounded to the nearest integer. If it was less than 10, it was rounded to 1 decimal place. No value happened to be rounded to 1.0.

For each attention-check trial, an arbitrary number was given as the likelihood-ratio value output by the forensic-voice-comparison system. This number was different from the number that the recording told participants to enter into one of the boxes. The recording told the participants to ignore the number that was given as the output by the forensic-voice-comparison system.

2.7. Metrics for analysis of response data

2.7.1. Introduction

For each response by an individual listener: if a number was entered into the first box, it was treated as a likelihood-ratio value;

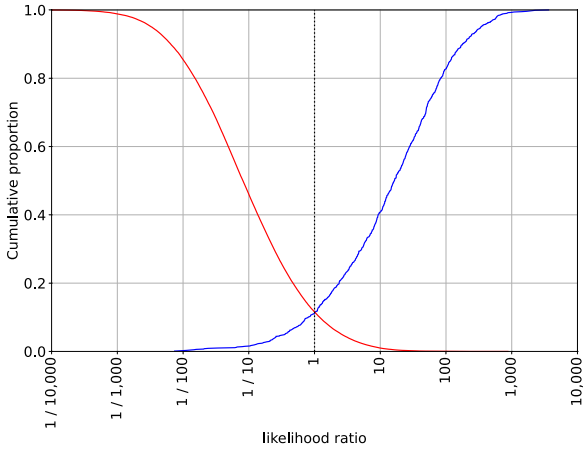

Fig. 2. Tippett plot presented to participants as part of the instructions in the experimental condition in which, in addition to listening to each pair of recordings, participants were provided with the likelihood-ratio output by a forensic-voice-comparison system in response to the same pair of recordings.

and if a number was entered into the second box, one divided by that number was treated as a likelihood-ratio value.

Three different performance metrics were calculated:[37]

- C_{llr} (§2.7.2) is a standard metric of the performance of forensic-evaluation systems. It measures the accuracy of systems that output likelihood ratios.
- D_{llr} (§2.7.3) is a metric of the scale of a listener's log-likelihood-ratio values relative to the log-likelihood-ratio values output by the forensic-voice-comparison system.
- B_{llr} (§2.7.4) is a metric of the shift of a listener's log-likelihood-ratio values relative to the log-likelihood-ratio values output by the forensic-voice-comparison system.

2.7.2. C_{llr}

For each listener, and for the forensic-voice-comparison system, the responses to the stimulus pairs were used to calculate a C_{llr} value [48]. C_{llr} was calculated using Equation (1), in which Λ_s and Λ_d are likelihood-ratio responses corresponding to same-speaker and different-speaker stimulus pairs respectively, and N_s and N_d are the number of same-speaker and different-speaker stimulus pairs respectively.

$$C_{llr} = \frac{1}{2}\left(\frac{1}{N_s}\sum_{i=1}^{N_s} log_2\left(1+\frac{1}{\Lambda_{s_i}}\right) + \frac{1}{N_d}\sum_{j=1}^{N_d} log_2(1+\Lambda_{d_j})\right) \quad (1)$$

C_{llr} is a standard metric of the performance of forensic-evaluation systems. It measures the accuracy of systems that output likelihood ratios. Its use is recommended in the *Consensus on validation of forensic voice comparison* [49]. For a system that always responded with a likelihood ratio of 1 irrespective of the input, the posterior odds would always equal the prior odds, and the system would therefore provide no useful information. Such a system would have a C_{llr} value of 1. If the C_{llr} value is less than 1, the system is providing useful information, and the better the performance of the system the lower the C_{llr} value will be. C_{llr} values cannot be less than or equal to 0. Uncalibrated or miscalibrated systems can have C_{llr} values that are greater than 1.

2.7.3. D_{llr}

In order to compare an individual-listener's responses with the forensic-voice-comparison system's responses, we also calculated a pairwise difference metric, D_{llr}, see Equation (2), in which subscript h represents a human-listener's response and subscript f represents a response by the forensic-voice-comparison system. If the D_{llr} value is greater than 0, the human listener is, on average, better at distinguishing between speakers than is the forensic-voice-comparison system (on average, their likelihood-ratio responses to same-speaker pairs and their likelihood-ratio responses to different-speaker pairs are further apart), and if the D_{llr} value is less than 0, the human listener is, on average, worse at distinguishing between speakers than is the forensic-voice-comparison system (on average, their likelihood-ratio responses to same-speaker pairs and their likelihood-ratio responses to different-speaker pairs are closer together). A D_{llr} of +1 would indicate that, on average, a listener's likelihood-ratio responses to same-speaker pairs and their responses to different-speaker pairs are twice as far apart as those of the forensic-voice-comparison system, a D_{llr} of +2 that they are four times further apart, a D_{llr} of +3 that they are eight times further apart, etc. A D_{llr} of −1 would indicate that, on average, a listener's likelihood-ratio responses to same-speaker pairs and their responses to different-speaker pairs are half as far apart as those of the forensic-voice-comparison system, a D_{llr} of −2 that they are a quarter as far apart, a D_{llr} of −3 that they are an eighth as far apart, etc.

$$D_{llr} = \frac{1}{2}\left(\frac{1}{N_s}\sum_{i=1}^{N_s}\left(log_2(\Lambda_{h,s_i}) - log_2(\Lambda_{f,s_i})\right) + \frac{1}{N_d}\sum_{j=1}^{N_d}\left(log_2(\Lambda_{f,d_j}) - log_2(\Lambda_{h,d_j})\right)\right) \quad (2)$$

2.7.4. B_{llr}

In order to compare an individual-listener's responses with the forensic-voice-comparison system's responses, we also calculated a pairwise relative-bias metric, B_{llr}. B_{llr} is calculated using Equation (3).[38] If the B_{llr} value is greater than 0, then, relative to the forensic-voice-comparison system, the human-listener's responses are biased toward giving larger likelihood-ratio response values (biased in favour of the same-speaker hypothesis), and if the B_{llr} value is less than 0, then, relative to the forensic-voice-comparison system, the human-listener's responses are biased toward giving smaller likelihood-ratio response values (biased in favour of the different-speaker hypothesis). A B_{llr} value of +1 would indicate that, on average, the listener's likelihood-ratio responses are twice as large as those of the forensic-voice-comparison system, a B_{llr} value of +2 that they are four times as large, a B_{llr} value of +3 that they are eight times as large, etc. A B_{llr} value of −1 would indicate that, on average, the listener's likelihood-ratio responses are half as large as those of the forensic-voice-comparison system, a B_{llr} value of −2 that they are a quarter as large, a B_{llr} value of −3 that they are an eighth as large, etc.

$$B_{llr} = \frac{1}{2}\left(\frac{1}{N_s}\sum_{i=1}^{N_s}\left(log_2(\Lambda_{h,s_i}) - log_2(\Lambda_{f,s_i})\right) + \frac{1}{N_d}\sum_{j=1}^{N_d}\left(log_2(\Lambda_{h,d_j}) - log_2(\Lambda_{f,d_j})\right)\right) \quad (3)$$

[37] D_{llr} and B_{llr} are named by analogy with C_{llr}. All three have a base-two logarithmic scale, but they do not have the same range: C_{llr} values are greater than 0, with 1 being a reference value, whereas D_{llr} and B_{llr} values are less than or greater than 0, with 0 being a reference value. D_{llr} and B_{llr} are not costs measured in bits.

[38] Note that Equation (2) and Equation (3) are not the same. The second part of Equation (2) contains $log_2(\Lambda_{f,d_j}) - log_2(\Lambda_{h,d_j})$, whereas the equivalent part of Equation (3) is reversed, i.e., $log_2(\Lambda_{h,d_j}) - log_2(\Lambda_{f,d_j})$.

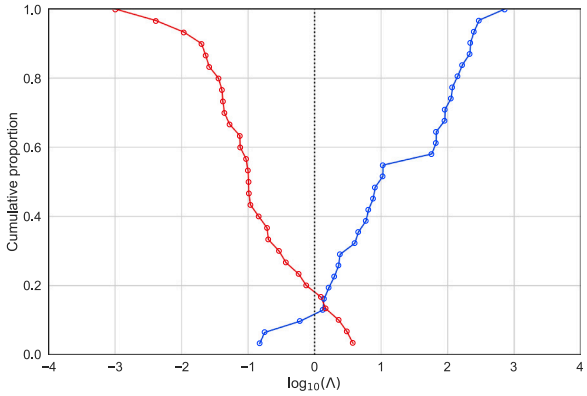

Fig. 3. Tippett plot of validation results from the forensic-voice-comparison system.

3. Results and discussion

3.1. Forensic-voice-comparison system

When previously validated on the full set of full-length *forensic_eval_01* validation recordings, the C_{llr} value for E^3FS^3 was 0.21 [40]. Given the smaller number of shorter validation recordings used in the current research, the C_{llr} value was 0.42. Poorer performance on shorter recordings is what would be expected, see examples in Weber et al. [40].

A Tippett plot of the validation results from E^3FS^3 is provided in Fig. 3. For an explanation of how to interpret Tippett plots, see Appendix C.1 of the *Consensus on validation of forensic voice comparison* [49] and the references cited therein. Likelihood-ratio values resulting from same-speaker comparisons ranged up to approximately 750, and likelihood-ratio values resulting from different-speaker comparisons ranged down to approximately 1/1000.

3.2. Individual listeners with different language backgrounds

3.2.1. Demographics

We excluded from analysis the submissions from listeners who did not answer all of the attention-check trials correctly,[39] and the submissions from listeners who, despite indicating that they were eligible at the informed-consent stage, gave answers to demographic questions about language and accent familiarity which indicated that they did not satisfy eligibility criterion 4.[40] After the removal of these submissions, there were:

- 53 submissions from Australian-English listeners
  - for reported ages, minimum, lower quartile, median, upper quartile, and maximum were 21, 25, 30, 37, and 68 years respectively
  - 29 identified as females and 24 as males
  - 49 identified as first-language English speakers
- 61 submissions from North-American-English listeners
  - for reported ages, minimum, lower quartile, median, upper quartile, and maximum were 22, 27, 32, 39, and 72 years respectively
  - 23 identified as females and 34 as males
  - 53 identified as first-language English speakers
  - 43 stated that they were “somewhat familiar” and 18 that they were “not familiar” with Australian English
- 55 submissions from Spanish-language listeners
  - for reported ages, minimum, lower quartile, median, upper quartile, and maximum were 20, 24, 27, 34, and 72 years respectively
  - 20 identified as females and 35 as males
  - 54 identified as first-language Spanish speakers
  - 17 stated that they were “very familiar” and 39 that they were “somewhat familiar” with English
  - 26 stated that they were “somewhat familiar” and 30 that they were “not familiar” with Australian English

Note that all this information was self reported. We are sceptical about the high proportions of North-American-English participants and Spanish-Language participants who stated that they were somewhat familiar with Australian English. On the Likert scale, “Somewhat familiar” with Australian English was glossed as “For example, I frequently watch Australian TV programmes, have multiple Australian friends, and/or I have visited Australia”.

3.2.2. C_{llr} values

A C_{llr} value was calculated separately for each individual listener’s responses. Fig. 4 shows violin plots of the resulting C_{llr} values grouped by the listeners’ language backgrounds. The horizontal line indicates the C_{llr} value for the forensic-voice-comparison system.

In terms of C_{llr}, there was large inter-listener variability. All of the listeners, however, performed worse than the forensic-voice-comparison system. The lowest C_{llr} from a listener was 0.51, compared to 0.42 for the forensic-voice-comparison system. Just over half of the English-language listeners (30 of the 53 Australian-English listeners, 33 of the 61 North-American-English listeners) and three quarters of the Spanish-language listeners (41 of 55) had $C_{llr} \geq 1$, i.e., they performed worse than a system that provided no useful information.

In terms of C_{llr}, the North-American-English listeners’ median and quartile values were somewhat higher than those of the Australian-English listeners, and 9 of the Australian-English listeners performed better than the best-performing North-American-English listener. The Spanish-language listener’s quartile values were somewhat higher and their median was substantially higher than those of the North-American-English listeners. This suggests that greater language familiarity contributes to better speaker-identification performance.

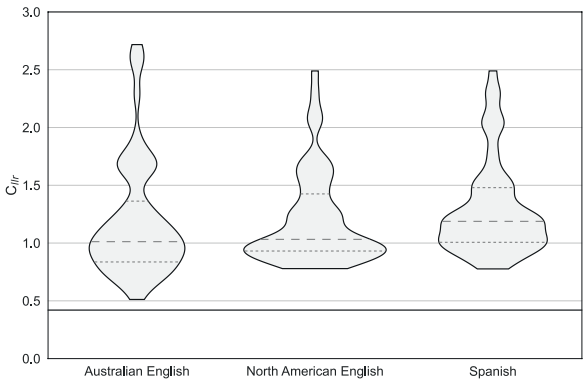

Fig. 4. Violin plots of the C_{llr} values for the responses from individual listeners. The heavy black horizontal line indicates the C_{llr} value for the forensic-voice-comparison system.

[39] We did not exclude submissions for which the failure to answer all the attention-check questions correctly were obviously the result of transposition errors, e.g., entering the correct number in the wrong box or writing “16” for “61”.

[40] To be eligible, Australian-English listeners had to answer “extremely familiar” for both English and Australian English, and North-American-English listeners had to answer “extremely familiar” for English. Although a first-accent Australian-English listener who had lived in the US or Canada for more then 4 years would have satisfied eligibility criterion 4, we excluded from analysis submissions from North-American-English listeners who stated that they were “extremely familiar” with Australian English (which required that they be first-accent Australian-English speakers, or that they be resident in Australia). Although a first-language English speaker who had been resident in Chile, Mexico, or Spain for more then 4 years would have satisfied eligibility criterion 4, we excluded from analysis submissions from Spanish-language listeners who stated that they were first-language English speakers or that they were “extremely familiar” with English (which required that they be first-language English speakers, or that they be resident in a predominantly English-speaking country).

Judges, like other lay listeners, would be expected to exhibit inter-listener variability in speaker-identification performance. A limitation of online recruiting and unsupervised participation in the individual-listener experiment is that many listeners are unlikely to have approached the task as conscientiously as would be expected of a judge listening to questioned- and known-speaker recordings in the context of a legal case. Many listeners in the individual-listener experiment may, therefore, have performed worse than would be expected for judges in the context of a case. We expect, however, that the best-performing listeners in the individual-listener experiment approached the task conscientiously and were intrinsically good at speaker identification. We would not, therefore, expect judges in general to be better at speaker identification than the best-performing listeners in the individual-listener experiment.[41]

3.2.3. D_{llr} values

A D_{llr} value was calculated separately for each individual listener's responses. Fig. 5 shows violin plots of the resulting D_{llr} values grouped by the listeners' language backgrounds.

Apart from a few outliers, across all language backgrounds, all the listeners' D_{llr} values were negative, i.e., compared to the forensic-voice-comparison system, their scaling of log-likelihood-ratio values was narrower: on average, their likelihood-ratio responses to same-speaker pairs and their likelihood-ratio responses to different-speaker pairs were closer to each other than those of the forensic-voice-comparison system. The median scaling for Australian-English listeners responses was about a fifth that of the forensic-voice-comparison system, for North-American-English listeners it was about a sixth, and for Spanish-language listeners it was about a seventh. Within each language background, there was substantial inter-listener variability.

3.2.4. B_{llr} values

A B_{llr} value was calculated separately for each individual listener's responses. Fig. 6 shows violin plots of the resulting B_{llr} values grouped by the listeners' language backgrounds.

The B_{llr} values indicate that, relative to the forensic-voice-comparison system, the listeners were predominantly biased toward giving responses that favoured the different-speaker hypothesis. More than 90% of the listeners (48 of the 53 Australian-English listeners, 57 of the 61 North-American-English listeners, and 50 of the 55 Spanish-language listeners) exhibited relative bias that favoured the different-speaker hypothesis. The median and quartile values across the different language backgrounds were similar. Across language backgrounds, the median relative bias was such that likelihood-ratio values were, on average, a little above half those of the forensic-voice-comparison system. There was, however, substantial inter-listener variability.

The likelihood-ratio output of the forensic-voice-comparison system in response to the validation data may have had a slight absolute bias in favour of the same-speaker hypothesis. This is likely due to sampling variability between the data used to train the calibration model and the data used to validate the system (see [50] for a discussion of this issue). The data used to train the calibration model were deliberately different from those used to validate the system. Calibrating and validating on the same data would result in better

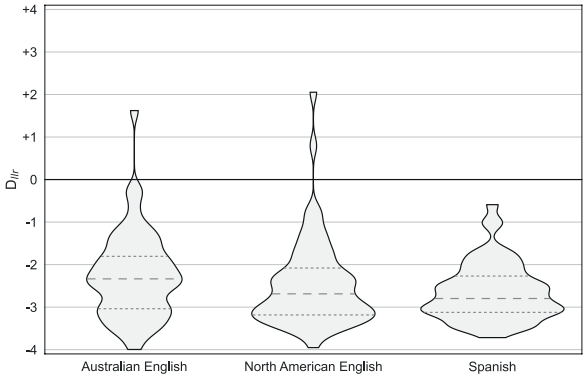

Fig. 5. Violin plots of the D_{llr} values for the comparison of individual listeners' responses with the responses of the forensic-voice-comparison system.

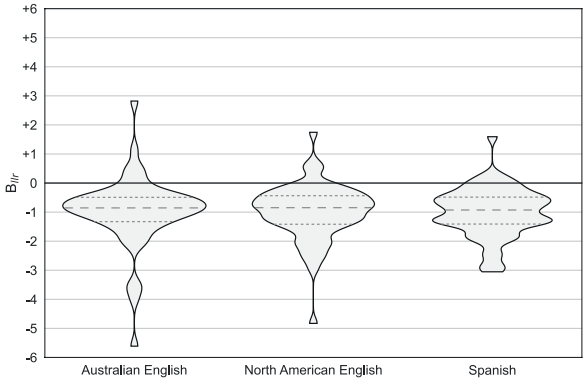

Fig. 6. Violin plots of the B_{llr} values for the comparison of individual listeners' responses with the responses of the forensic-voice-comparison system.

calibrated output for those particular data. What matters, however, is how well the system performs on previously unseen data, such as the questioned-speaker and known-speaker recordings in a case. Even taking into account that, for the particular validation data, the forensic-voice-comparison system may have had a slight absolute bias in favour of the same-speaker hypothesis, the magnitudes of the negative B_{llr} values indicate that the majority of the listeners had absolute biases in favour of the different-speaker hypothesis. These biases could potentially be due to the poor recording conditions and the mismatches in recording conditions between the questioned-speaker-condition and the known-speaker-condition recordings. These would have made the voices on the two recordings in each pair sound more different from one another than had they both been high-quality recordings.

The individual-listener experiment did not include any contextual information that would be expected to bias the listeners. There is concern in the earwitness literature that context could influence a listener to expect to hear a particular individual and bias them toward identifying a speaker whom they hear as that individual, e.g., if a listener is asked to identify a speaker in a showup scenario rather than in a well-designed voice lineup [18,51]. Similar concerns apply when a trier of fact is asked to compare a voice on a recording with the voice of the defendant [5]. Abundant psychology research indicates that judges would not be immune from the potential effects of contextual bias [52]. Given such contexts, different speakers who sound at-least somewhat similar could be incorrectly identified as the same speaker, a situation that would usually favour the prosecution. The observed relative bias in the responses to the individual-listener experiment favoured the different-speaker hypothesis rather than the same-speaker hypothesis. We assume that, since it was observed in a neutral context, this is due to an intrinsic bias. This should still be of concern, however, as bias that usually favours the defence would not be in the interest of victims. It

[41] One of the reviewers suggested that the overall poor performance of the listeners may have been due to a lack of opportunity for training and to listeners not fully understanding the task of assigning a likelihood ratio. The reviewer proposed that this could be addressed, and a fairer comparison obtained, by calibrating listeners' responses. In a courtroom context, however, judges are not trained in speaker identification and their speaker-identification judgements are not calibrated. Lack of training and lack of calibration may be a cause of listeners' poor performance, but it is the untrained uncalibrated performance of individual listeners that is relevant for addressing the research question of whether speaker identification by a judge listening alone would be more or less accurate than the output of a forensic-voice-comparison system that is based on state-of-the-art automatic-speaker-recognition technology.

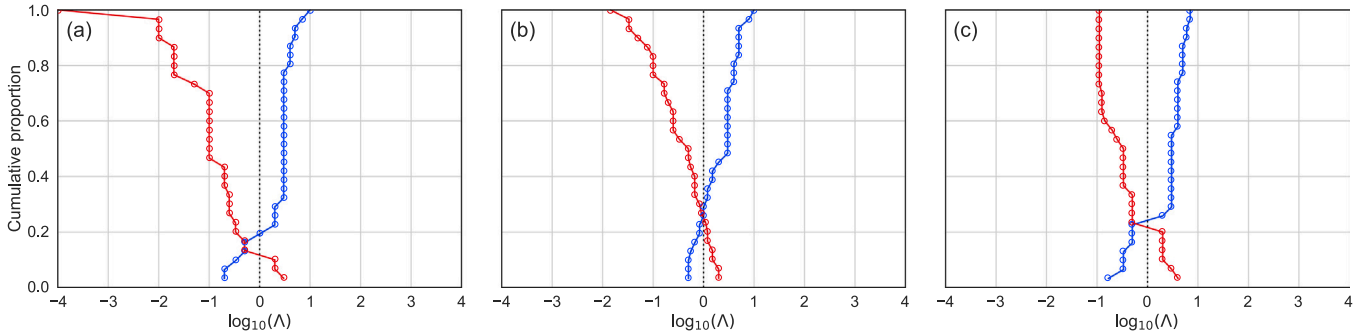

Fig. 7. Tippett plots of the results from the three best-performing listeners, i.e., those with the lowest C_{llr} values.
(a) C_{llr} = 0.51, D_{llr} = −1.3, B_{llr} = −1.5
(b) C_{llr} = 0.64, D_{llr} = −2.2, B_{llr} = −0.7
(c) C_{llr} = 0.65, D_{llr} = −2.1, B_{llr} = −0.6.

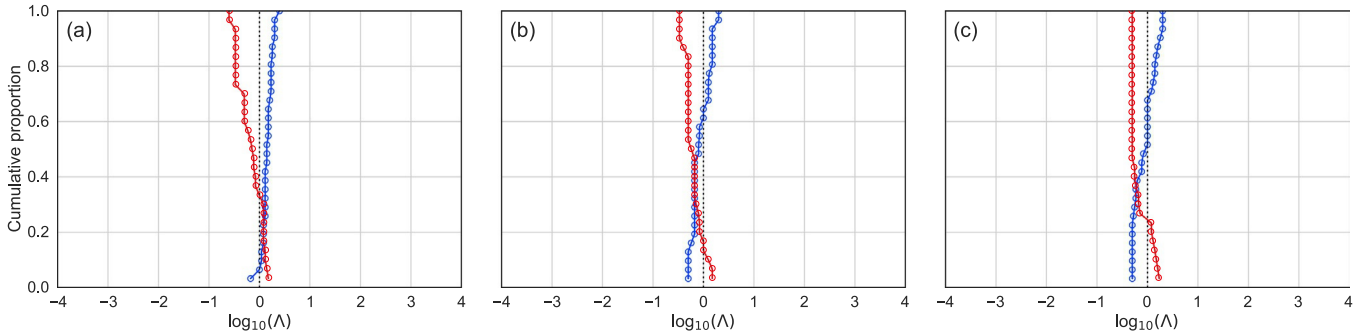

Fig. 8. Example Tippett plots of the results from listeners who used narrow ranges of likelihood-ratio values.
(a) C_{llr} = 0.77, D_{llr} = −2.9, B_{llr} = −0.5
(b) C_{llr} = 0.92, D_{llr} = −3.3, B_{llr} = −0.8
(c) C_{llr} = 0.95, D_{llr} = −3.3, B_{llr} = −0.8.

may also be unwise to assume that this intrinsic bias would counteract the potential effect of a contextual bias in favour of a same-speaker response.

3.2.5. Tippett plots

There was substantial inter-listener variability, but several patterns were discernable in Tippett plots of the listeners' responses. These different patterns may reflect different conscious or unconscious strategies employed by the listeners. In this subsection, we show example Tippett plots of the patterns which we discerned (excluding those that only occurred in a few outliers). In the caption of each figure, we provide the C_{llr}, D_{llr}, and B_{llr} values corresponding to the Tippett plots shown.

Fig. 7 shows Tippett plots of the results from the three best-performing listeners, i.e., those with the lowest C_{llr} values. Compared to the results from the forensic-voice-comparison system (see Fig. 3), the number-one best-performing listener's responses to same-speaker pairs were too low (too close to a log-likelihood ratio of 0 / too close to a likelihood ratio of 1), i.e., they were too conservative. This resulted in both a negative D_{llr} value and a negative B_{llr} value.

Fig. 8 shows example Tippett plots of the results from listeners who used narrow ranges of likelihood-ratio values – the values of their likelihood-ratio responses to same-speaker pairs and their likelihood-ratio responses to different-speaker pairs were too close to each other, i.e., they were too conservative. This resulted in large negative D_{llr} values. Although the same trend can be observed in the responses from the best-performing listeners (Fig. 7), in the examples given in Fig. 8, the pattern is more extreme. It also resulted in higher C_{llr} values.

Fig. 9 shows example Tippett plots of the results from listeners who gave lots of likelihood-ratio-equal-to-one responses. This could be considered an extreme version of the conservative pattern just shown in Fig. 8. It resulted in C_{llr} values close to 1, large negative D_{llr} values, and B_{llr} values close to 0. If these listeners were conscientiously engaged with the task,[42] then this would suggest that, under the conditions tested, they found it difficult to perform speaker identification.

Fig. 10 shows example Tippett plots of the results from listeners who mostly used one response magnitude, i.e., they almost always entered the same number but entered it into the first box or the second box depending on whether they thought the pair of recordings was a same-speaker pair or a different-speaker pair. In panel (a) the listener almost always entered the number 100, and in panel (b) the listener always entered the number 2. A variant of this pattern was mostly using only two or three different numbers, e.g., in panel (c) the listener almost always entered 100 or 200. If these listeners were conscientiously engaged in the task, these results may reflect a strategy whereby they made a categorical decision on same-speaker versus different-speaker, picked a single value to represent that categorical decision, and potentially, if they were more or less certain about their decision, chose other values anchored on that first value. This pattern did not result in consistency in terms of C_{llr}, D_{llr}, or B_{llr} values.

Finally, Fig. 11 shows example Tippett plots of results that were strongly biased toward the different-speaker hypothesis. As discussed in §3.2.4, this was a common pattern. It resulted in C_{llr} values greater than 1, and large negative B_{llr} values. Results similar to panel (a) were particularly common.

[42] All of the responses from one listener were ones, but we suspect that this listener may not have been conscientiously engaged with the task.

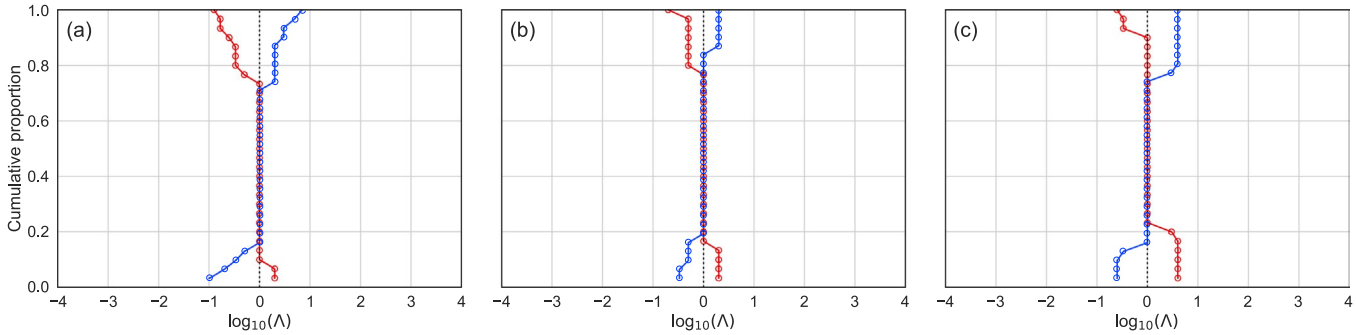

Fig. 9. Example Tippett plots of the results from listeners who gave lots of likelihood-ratio-equal-to-one responses.
(a) C_{llr} = 0.95, D_{llr} = −3.2, B_{llr} = −0.6
(b) C_{llr} = 1.01, D_{llr} = −3.5, B_{llr} = −0.5
(c) C_{llr} = 1.09, D_{llr} =−3.5, B_{llr} = −0.2.

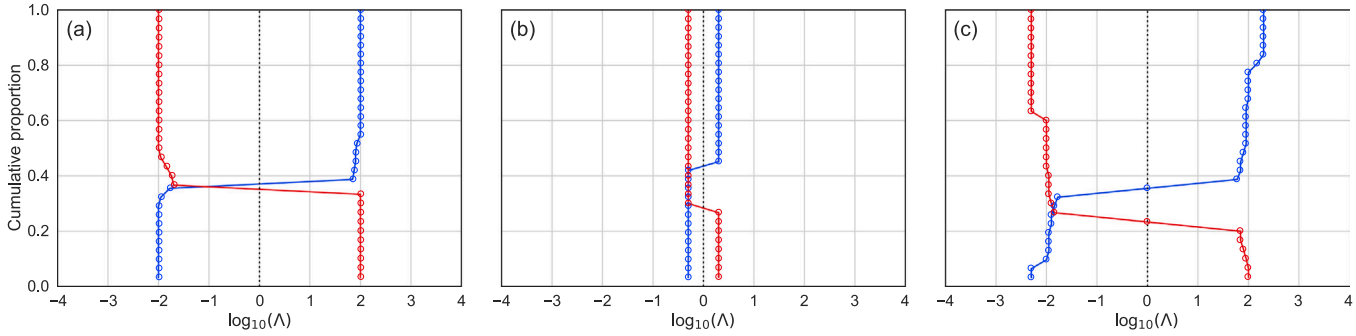

Fig. 10. Example Tippett plots of the results from listeners who mostly used one response magnitude.
(a) C_{llr} = 2.29, D_{llr} = −1.5, B_{llr} = −0.5
(b) C_{llr} = 0.93, D_{llr} = −3.2, B_{llr} = −0.6
(c) C_{llr} = 1.75, D_{llr} = −0.3, B_{llr} = −1.3.

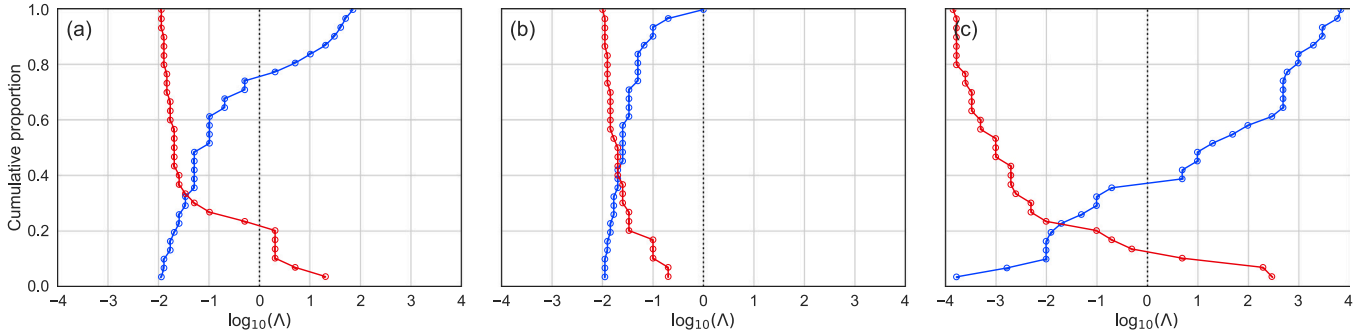

Fig. 11. Example Tippett plots of the results from listeners whose responses were strongly biased toward the different-speaker hypothesis.
(a) C_{llr} = 1.90, D_{llr} = −2.5, B_{llr} = −3.5
(b) C_{llr} = 2.57, D_{llr} = −3.3, B_{llr} = −5.6
(c) C_{llr} = 1.43, D_{llr} = +2.1, B_{llr} = −3.0.

3.2.6. Listeners' beliefs about their own speaker-identification abilities

Both before and after the experiment, each participant was asked to indicate on a 5-point Likert scale how good they thought they were at identifying speakers in general and how good they thought they were at identifying adult male Australian-English speakers in particular. The levels on the scale were: 1. "very poor", 2. "poor", 3. "neutral", 4. "good", 5. "very good". Fig. 12, Fig. 13, and Fig. 14 show the responses from Australian-English, North-American-English, and Spanish-language listeners respectively. In each figure, the left panels show the responses for speakers in general, and the right panels show the responses for adult male Australian-English speakers in particular. In each figure, the top panels show the Likert-scale responses from before the experiment, the middle panels show the Likert-scale responses from after the experiment, and the bottom panels show the pairwise differences between the listeners' Likert-scale responses from before and after the experiment.

For Australian-English listeners, as would be expected, their responses were similar for speakers in general and for adult male Australian-English speakers in particular. For both types of speakers, approximately half the listeners indicated that they thought they were worse at speaker identification after the experiment than they thought they were before the experiment (26 of 53 listeners for speakers in general and 29 of 53 listeners for adult male Australian-English speakers in particular). For each type of speaker, only 2

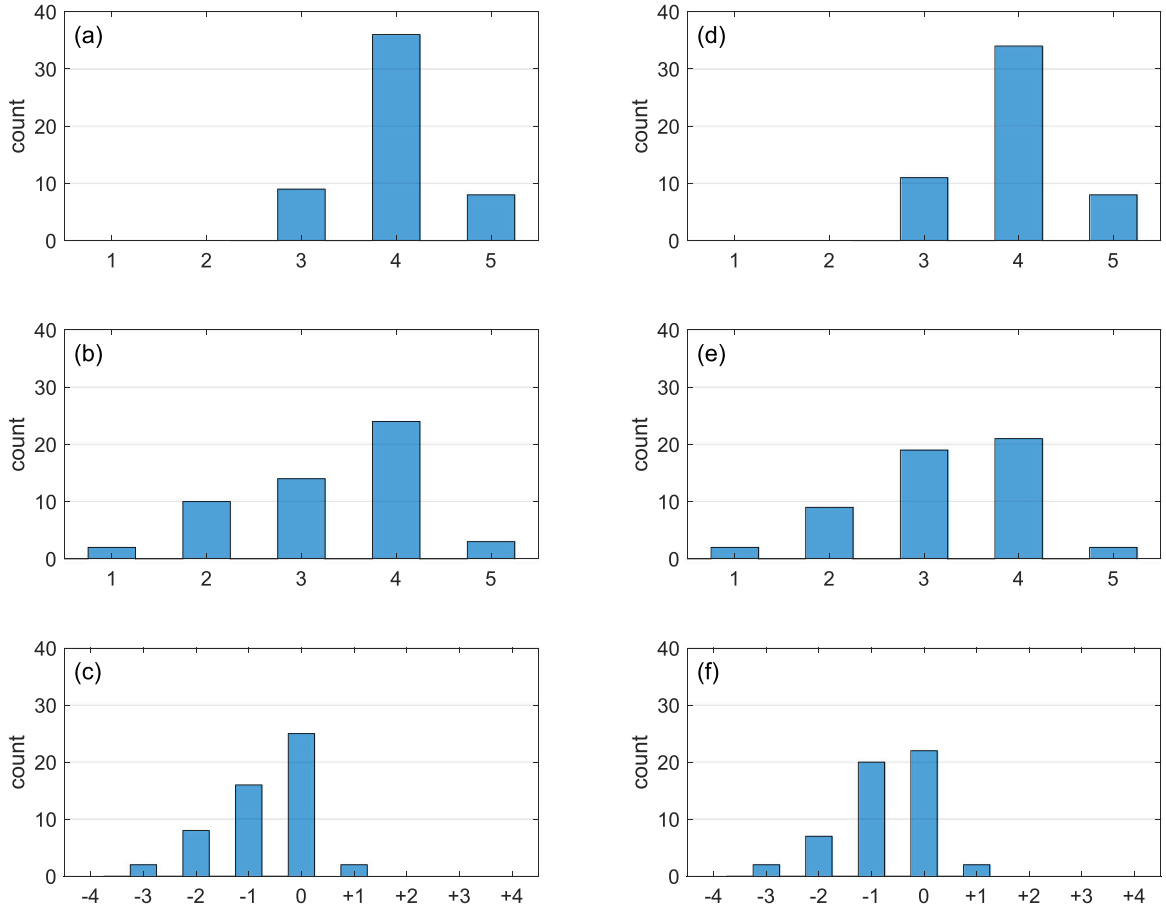

Fig. 12. Individual Australian-English listeners' responses with respect to how good they thought they were at identifying speakers in general, left panels [(a), (b), and (c)], and at identifying adult male Australian-English speakers in particular, right panels [(d), (e), and (f)]. Top panels [(a) and (d)] show Likert-scale responses before the experiment. Middle panels [(b) and (e)] show Likert-scale responses after the experiment. Bottom panels [(c) and (f)] show the pairwise differences between the listeners' Likert-scale responses from before and after the experiment.

listeners indicated that they thought they were better after the experiment than before. This suggests that, even without feedback on the correctness of their answers, the experience of performing the task made many of the Australian-English listeners think that they had initially overestimated their speaker-identification abilities.

For North-American-English listeners, their initial responses indicated that they were less confident in their ability to identify adult male Australian-English speakers in particular than in their ability to identify speakers in general. This suggests that the listeners were aware of the potential impact of accent familiarity on speaker identification. For speakers in general, approximately half the listeners (32 of 61 listeners) indicated that they thought they were worse at speaker identification after the experiment than they thought they were before the experiment. For adult male Australian-English speakers in particular, this was the case for approximately two-thirds of listeners (39 of 61 listeners). For both types of speakers, only a few listeners indicated that they thought they were better after the experiment than before (7 of 61 listeners for speakers in general and 9 of 61 listeners for adult male Australian-English speakers in particular). This suggests that, even without feedback on the correctness of their answers, the experience of performing the task made many of the North-American-English listeners think that they had initially overestimated their speaker-identification abilities.

For Spanish-language listeners, their initial responses indicated that they were less confident in their ability to identify adult male Australian-English speakers in particular than in their ability to identify speakers in general. Also, they were initially less confident in their ability to identify adult male Australian-English speakers than were the North-American-English listeners. This suggests that the Spanish-language listeners were aware of the potential impact of language familiarity on speaker identification. For speakers in general, approximately one-third of the listeners (19 of 55 listeners) indicated that they thought they were worse at speaker identification after the experiment than they thought they were before the experiment. For adult male Australian-English speakers in particular, this was the case for more than two-fifths of listeners (23 of 55 listeners). For both types of speakers, only a few listeners indicated that they thought they were better after the experiment than before (2 of 55 listeners for speakers in general and 6 of 55 listeners for adult male Australian-English speakers in particular). This suggests that, even without feedback on the correctness of their answers, the experience of performing the task made many of the Spanish-language listeners think that they had initially overestimated their speaker-identification abilities for adult male Australian-English speakers. This experience, however, made fewer of the Spanish-language listeners think that they had overestimated their speaker-identification abilities for speakers in general. A similar differential was observed for North-American-English listeners. This suggests that these listeners attributed experiencing more difficulty than expected with the task as being due in part to language unfamiliarity or accent unfamiliarity, with language unfamiliarity resulting in less diminishment to their confidence in their speaker-identification abilities in general than accent unfamiliarity.

Before the experiment, each individual listener was asked to respond to the question:

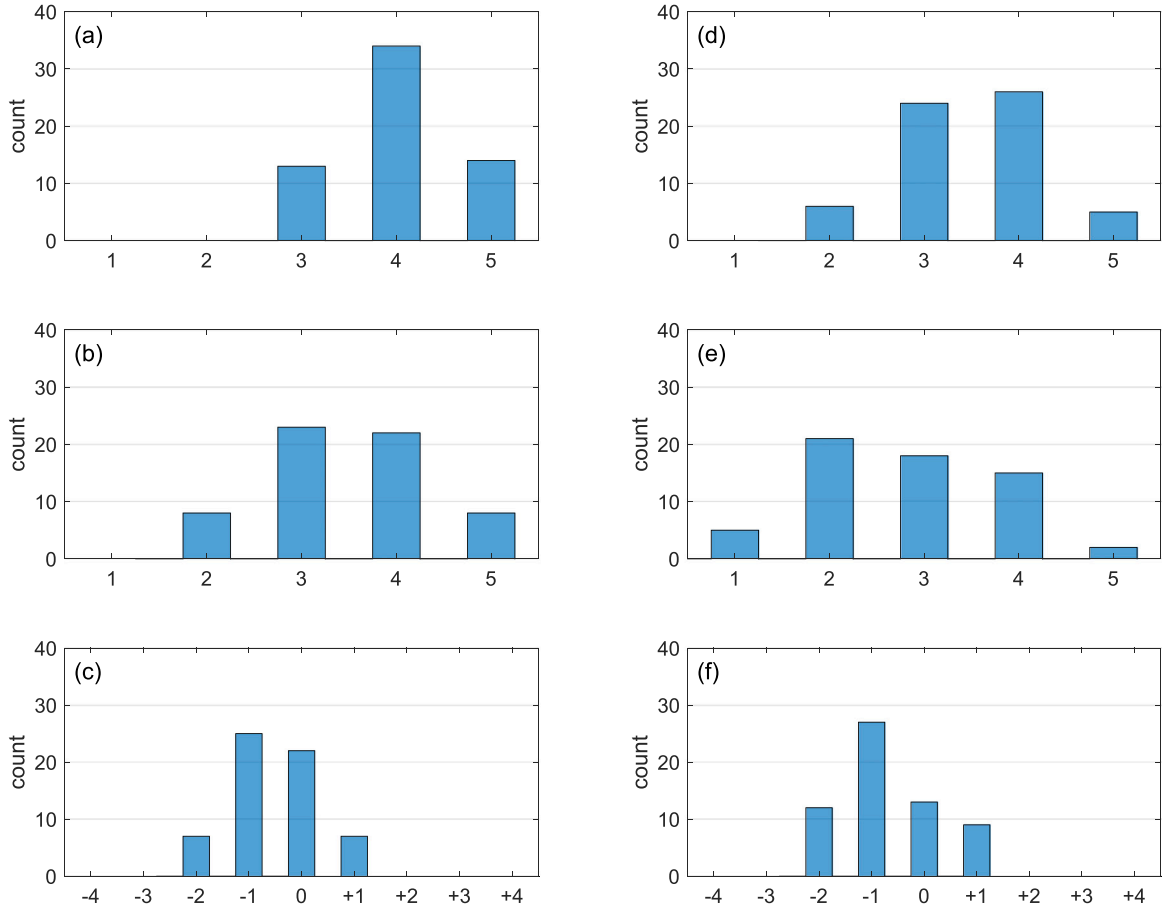

Fig. 13. Individual North-American-English listeners' responses with respect to how good they thought they were at identifying speakers in general left panels [(a), (b), and (c)], and at identifying adult male Australian-English speakers in particular, right panels [(d), (e), and (f)]. Top panels [(a) and (d)] show Likert-scale responses before the experiment. Middle panels [(b) and (e)] show Likert-scale responses after the experiment. Bottom panels [(c) and (f)] show the pairwise differences between the listeners' Likert-scale responses from before and after the experiment.

- If you heard a large number of pairs of recordings of adult male Australian-English speakers, what percentage of the pairs do you think you would get "right", i.e., if they were recordings of the same speaker you would say they were recordings of the same speaker and if they were recordings of different speakers you would say they were recordings of different speakers? Count saying "can't decide" as incorrect.

For each listener, we ignored the magnitudes of their responses and calculated their actual correct-classification rate as the proportion of recording-pairs for which they entered a value greater than one into the correct box (the first box if the recording pair was a same-speaker pair, and the second box if the recording pair was a different-speaker pair). Responses equal to one were counted as errors. This approximates a situation in which listeners had ternary response options: "same speaker", "different speaker", and "don't know".[43] Fig. 15 plots each listener's own initial estimate of their correct-classification rate against their actual correct-classification rate. If a data point is above the diagonal line, the listener overestimated their speaker-identification ability. If a data point is below the diagonal line, the listener underestimated their speaker-identification ability. The vertical distance to the diagonal line indicates the amount by which they overestimated or underestimated their ability. The heavy vertical line represents the correct-classification rate for the forensic-voice-comparison system, 87 %, which was calculated using equal priors and a posterior-odds threshold of 1.[44] If a data point is to the left of the vertical line, the listener's correct-classification rate was worse than that of the forensic-voice-comparison system.

All the individual listeners' correct-classification rates were worse than the correct-classification rate for the forensic-voice-comparison system. In terms of correct-classification rates, some listeners' estimates of their speaker-identification abilities were close to their actual abilities, but others substantially under- or overestimated their abilities.[45] There was substantial inter-listener variation.

Most Australian-English listeners overestimated their speaker-identification abilities, some by large amounts. None

[43] Actual behaviour given those response options could differ.

[44] We do this only for the purpose of being able to make a comparison with the responses given by participants to a question that could be asked without requiring a lot of explanation. We would not present correct-classification rates (or classification-error rates) in the context of a legal case.

[45] The question asked before the experiment did not specify what the recording conditions would be or that there would be a mismatch in recording conditions, so some of the overconfidence could have been due to participants' assuming high-quality recording conditions, and they might have actually performed better on high-quality recordings. One listener (who took part in a slightly modified version of the experiment not otherwise reported in the present paper) sent us a comment stating: "I am 100 % sure I overestimated my ability to discern the voices SOLELY because I did not know the conditions of the sound recordings yet. Had I had a sample of what they would sound like FIRST, then I would have estimated 15 % or lower. were this in a USA court of law and I were a juror (I have sat on two juries in my lifetime.) I would immediately throw out this evidence and 100 % discard it." It turned out, however, that this listener did not substantially overestimate their correct-classification rate: their estimated correct-classification rate was 80 % and their actual correct-classification rate was 72 % (their C_{llr} was 0.84).

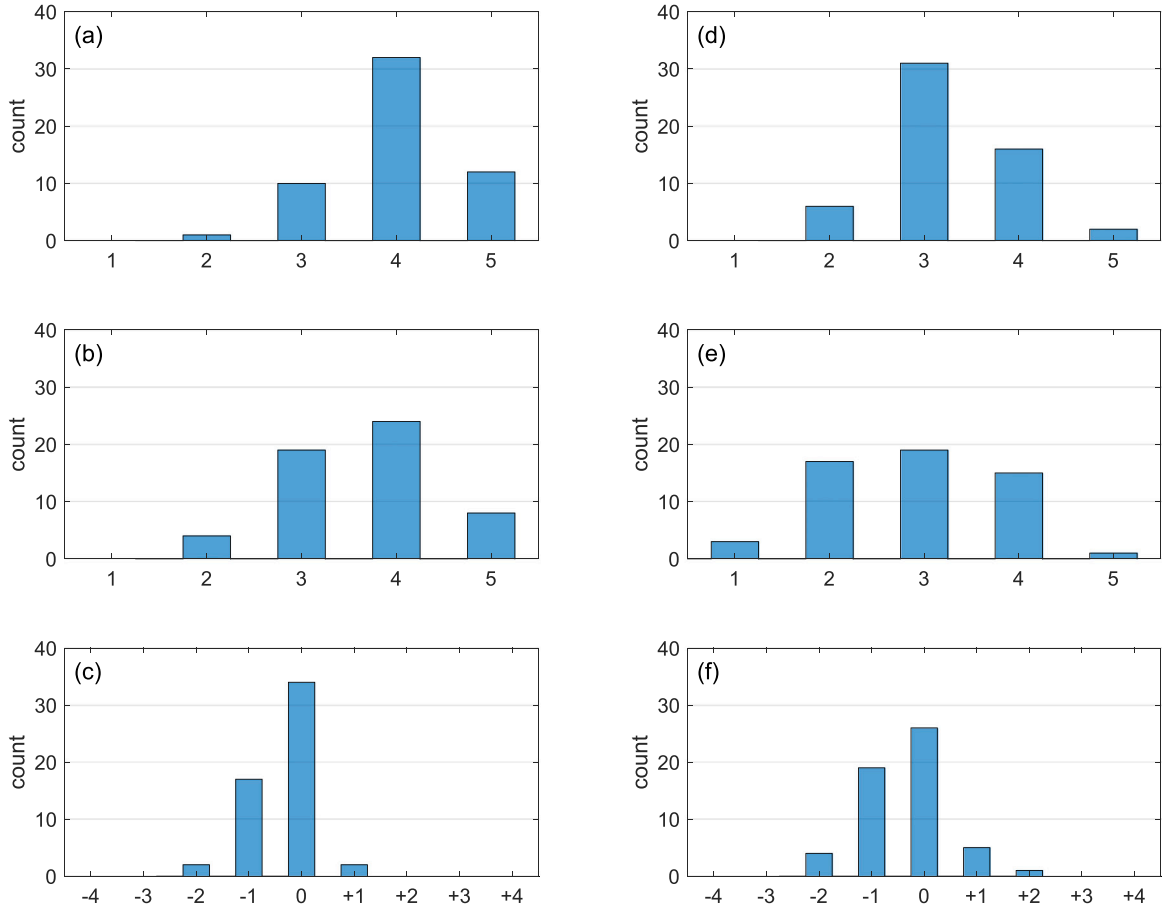

Fig. 14. Individual Spanish-language listeners' responses with respect to how good they thought they were at identifying speakers in general left panels [(a), (b), and (c)], and at identifying adult male Australian-English speakers in particular, right panels [(d), (e), and (f)]. Top panels [(a) and (d)] show Likert-scale responses before the experiment. Middle panels [(b) and (e)] show Likert-scale responses after the experiment. Bottom panels [(c) and (f)] show the pairwise differences between the listeners' Likert-scale responses from before and after the experiment.

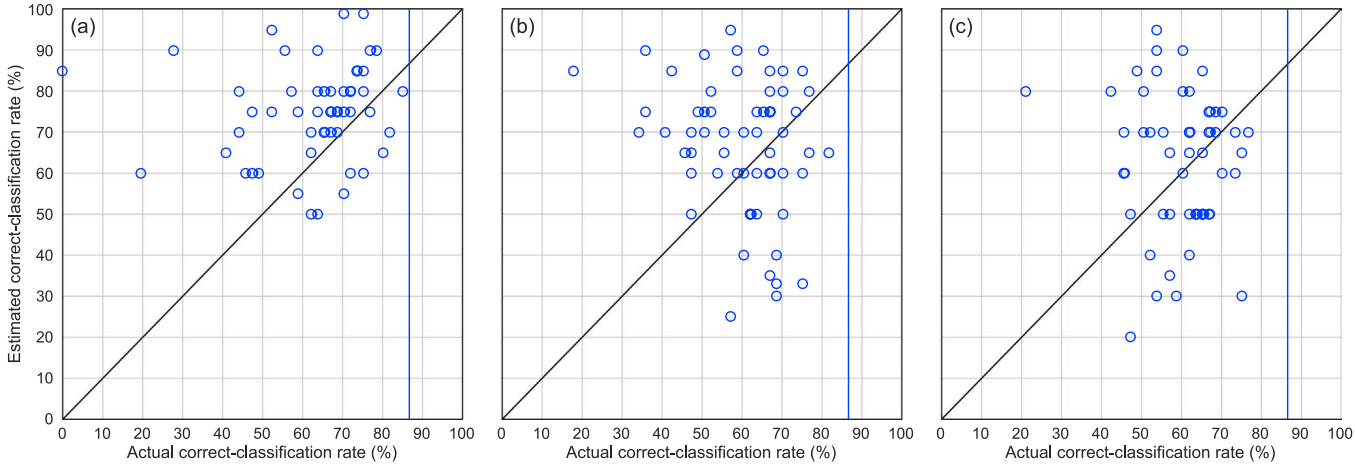

Fig. 15. Plot of each listener's own initial estimate of their correct-classification rate against their actual correct classification rate. (a) Australian-English listeners. (b) North-American-English listeners. (c) Spanish-language listeners.

underestimated their abilities by a large amount. Overconfidence due to familiarity with the accent of the speakers is a potential explanation for this pattern of results. In contrast some North-American-English listeners and some Spanish-language listeners overestimated their speaker-identification abilities by large amounts and some underestimated their speaker-identification abilities by large amounts. The apparent underconfidence of the latter listeners suggests that they were aware of the potential impact of accent or language familiarity on speaker identification.

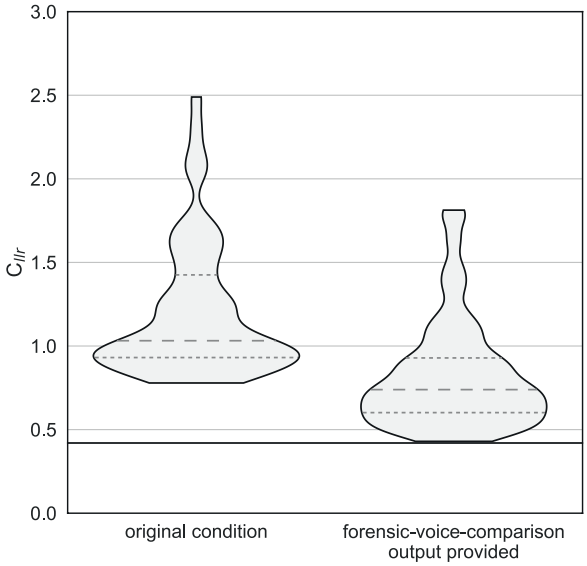

Fig. 16. Violin plots of the C_{llr} values for the responses from individual North-American-English listeners in the original condition, and in the condition in which they were provided with the likelihood-ratio values output by the forensic-voice-comparison system. The heavy black horizontal line indicates the C_{llr} value for the forensic-voice-comparison system.

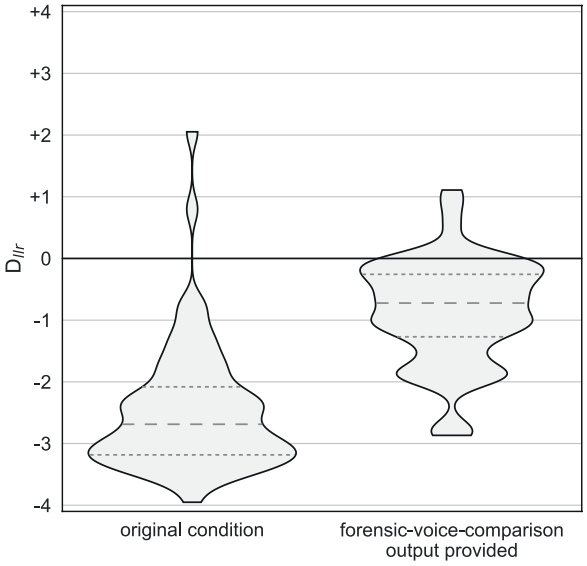

Fig. 17. Violin plots of the D_{llr} values for the comparison of individual North-American-English listeners' responses with the responses of the forensic-voice-comparison system in the original condition, and in the condition in which they were provided with the likelihood-ratio values output by the forensic-voice-comparison system.

3.3. Experiment in which forensic-voice-comparison results were provided

3.3.1. Demographics

After the removal of any submissions that fell under the same exclusion criteria as given in §3.2.1, there were:

- 57 submissions from North-American-English listeners
  - for reported ages, minimum, lower quartile, median, upper quartile, and maximum were 23, 29, 35, 41, and 66 years respectively
  - 32 identified as females and 23 as males
  - 55 identified as first-language English speakers
  - 3 stated that they were "very familiar", 34 that they were "somewhat familiar", and 20 that they were "not familiar" with Australian English

3.3.2. Performance metrics

Fig. 16, Fig. 17, and Fig. 18 provide violin plots of the C_{llr}, D_{llr}, and B_{llr} values resulting from individual North-American-English listeners' responses in the original condition and (other) North-American-English listeners' responses in the condition in which they were provided with the likelihood-ratio values output by the forensic-voice-comparison system.

The performance of the participants who (in addition to being able to listen to the recordings) were provided with the likelihood-ratio output of the forensic-voice-comparison system was better than the performance of the participants who only listened to the recordings. The distribution of their C_{llr} values was substantially lower. In addition to having better C_{llr} values, D_{llr} values and B_{llr} values were also better.

In terms of C_{llr}, no participant outperformed the stand-alone forensic-voice-comparison system. The best performance was from participants who always responded with values that were close to the likelihood-ratio values output by the forensic-voice-comparison system. No participant entered the exact likelihood-ratio values output by the forensic-voice-comparison system. The lowest C_{llr} value for a participant's responses was 0.43 (the C_{llr} value for the stand-alone forensic-voice-comparison system was 0.42). In terms of

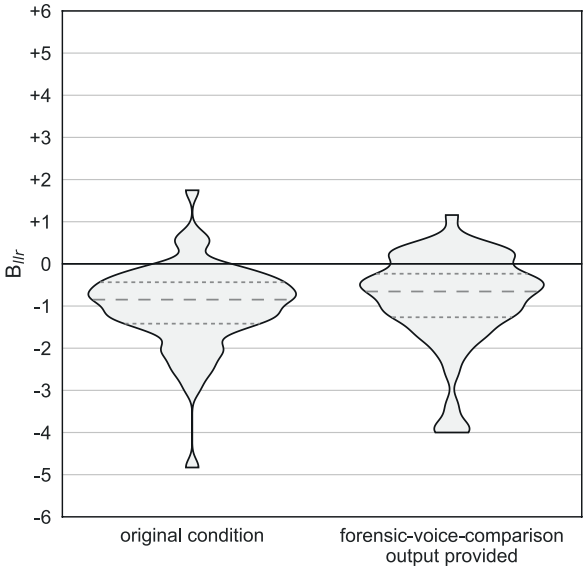

Fig. 18. Violin plots of the B_{llr} values for the comparison of individual North-American-English listeners' responses with the responses of the forensic-voice-comparison system in the original condition, and in the condition in which they were provided with the likelihood-ratio values output by the forensic-voice-comparison system.

correct-classification rate, one participant equalled the performance of the forensic-voice-comparison system at 87 % and one exceeded it at 89 %. All others performed worse than the forensic-voice-comparison system.

The latter results replicate the pattern observed in Matějka et al. [13] (described in §1.3) in which participants could both listen to recordings and consider the output of a automatic-speaker-recognition system. In that study, only 1 of 10 participants outperformed the stand-alone automatic-speaker-recognition system. The full results also replicate the pattern observed in other domains in which an algorithm alone outperforms humans who can adjust the output of the algorithm using their own subjective judgment

who in turn outperform humans who are not exposed to the algorithm and rely only on their own subjective judgment, e.g., Dietvorst et al. [53].

4. General discussion and conclusion

Expert testimony is only admissible in common law if it will potentially assist the trier of fact to make a decision that they would not be able to make unaided. If the trier of fact's speaker identification were equally accurate or more accurate than a forensic-voice-comparison system, then testimony based on the output of the forensic-voice-comparison system would not be admissible.

We tested the accuracy of speaker identification by individual lay listeners. This was intended to be informative with respect to a context in which a judge attempts to identify a speaker. The pairs of recordings that we used for testing reflected the conditions of the questioned-speaker and known-speaker recordings in an actual case. The accuracy of individual listeners' responses was compared with the accuracy of likelihood-ratio values output by E^3FS^3, a forensic-voice-comparison system that is based on state-of-the-art automatic-speaker-recognition technology. There was large inter-listener variation, but all listeners performed worse than the forensic-voice-comparison system. The lowest C_{llr} for a listener's responses was 0.51, which was substantially worse than the C_{llr} of 0.42 for likelihood-ratio values output by the forensic-voice-comparison system. In addition, more than half of the listeners' responses resulted in $C_{llr} \geq 1$, i.e., they performed worse than a system that provided no useful information.

Based on these results, at least under the particular case conditions tested, we infer that the forensic-voice-comparison system would satisfy the admissibility criterion of being more accurate than speaker identification performed by a judge. Also taking into consideration the results of previous research (which was summarized in §1.3), we think it is reasonable to extrapolate this inference to other recording conditions.

Given that forensic voice comparison based on state-of-the-art automatic-speaker-recognition technology outperforms speaker identification by individual listeners, we argue that judges should not attempt to perform speaker identification and should instead rely on expert testimony that is based on a validated forensic-voice-comparison system. We tested a condition in which individual listeners could attempt speaker identification based on listening to the pairs of recordings and could also consider the likelihood-ratio values output by the forensic-voice-comparison system in response to the same pairs of recordings. In terms of C_{llr}, no participant outperformed the stand-alone forensic-voice-comparison system (in terms of correct classification rate only 1 of 55 participants outperformed the stand-alone forensic-voice-comparison system). We therefore argue that judges should rely exclusively on expert testimony that is based on a validated forensic-voice-comparison system – they should not attempt to supplement it by performing their own speaker identification as this will almost certainly lead to a less accurate result.

Based on the results of the present research (and the results of past research), we also infer that forensic voice comparison based on state-of-the-art automatic-speaker-recognition technology would be more accurate than speaker identification performed by a police officer or an interpreter. [46] We therefore argue that, when both questioned-speaker and known-speaker recordings are available or obtainable, a trier of fact (judge or jury) should not be presented with lay speaker-identification testimony or "*ad hoc* expert" speaker-identification testimony, and should instead be presented with expert testimony based on a validated forensic-voice-comparison system.

Edmond [5] provides additional arguments for why judges and juries should not attempt to perform their own speaker identification and why they should not be presented with and should not consider lay or "*ad hoc* expert" speaker-identification testimony. Judges and juries are invited to perform their own speaker identification in the suggestive context of the accusatorial trial. In many cases, they also hear from police officers and interpreters who do not use validated methods and do not manage their own exposure to task-irrelevant information. Other evidence in the case contaminates the trier of fact's speaker-identification judgement and contaminates the speaker-identification testimony of lay and "*ad hoc* expert" witnesses, but the speaker-identification judgement and testimony are then treated as independent support for the evidence that contaminated them, and the speaker-identification judgement by the trier of fact is treated as independent support for the speaker-identification testimony that contaminated it.

The experiments conducted for the present study were decontextualized in that they were not embedded in case contexts that could potentially bias the listeners. Listeners' responses were biased in favour of the different-speaker hypothesis. This may have been because of the poor recording conditions, including the mismatch in conditions between the questioned-speaker-condition and known-speaker-condition recordings. These would have made the members of each pair of recordings sound more different from one another than had they both been high-quality recordings. In a future paper, we plan to report on experiments in which we provide contextual information that could potentially debias or differently bias the results, and on experiments in which we present high-quality versions of the recordings.

We tested individual listeners with different language and accent backgrounds: listeners who were familiar with both the language and accent spoken by the speakers (Australian-English listeners), listeners who were familiar with the language but less familiar with the accent (North-American-English listeners), and listeners who were less familiar with the language (Spanish-language listeners). The results were in accord with expectations based on previous research: the Australian-English listeners performed better than the North-American-English listeners, who in turn performed better than the Spanish-language listeners. Based on these results, speaker identification by judges who are unfamiliar with the language or accent spoken should be of even greater concern than when they are familiar with the language and accent spoken. We have, however, already argued that judges (and juries) should not attempt to perform speaker identification, even when the language and accent spoken are familiar to them, and that they should instead rely on expert testimony that is based on a validated forensic-voice-comparison system. For forensic voice comparison, the language and accent spoken is part of the specification of the relevant population adopted for the case. Assuming that data representative of the relevant population are available or obtainable, a forensic-voice-comparison system can be trained, calibrated, and validated for

[46] We remind the reader of the definitions provided in footnote 1. "Speaker identification" by humans refers to a situation in which the listener hears the voice of an unfamiliar speaker of questioned identity and hears the voice of an unfamiliar speaker of known identity, and makes a judgement as to whether the two voices belong to the same speaker or to different speakers. This is not the same as "speaker recognition" by humans, which refers to a situation where a listener hears the voice a speaker and makes a judgement as to whether it is the voice of a speaker who is familiar to them or not, and if the listener states that the voice is that of a speaker who is familiar to them they usually also state the name of the speaker. The research presented in the present paper relates to "speaker identification", not to "speaker recognition". The present discussion also relates to "speaker identification", not to "speaker recognition".

speakers speaking the language and accent of interest in the case, see the *Consensus on validation of forensic voice comparison* [49].

Previous research has suggested that listeners overestimate their own speaker-identification abilities (and overestimate other listeners' speaker-identification abilities). Both before and after the experiment, we asked listeners to indicate how good they thought they were at speaker identification in general and speaker identification of adult male Australian English speakers (the type of speakers in the experiment). About half the listeners indicated that they thought their ability to identify speakers was worse after the experiment than they thought it was before the experiment, and few indicated that they thought it was better. Even without feedback on the correctness of their responses, the experience of taking part in the experiment appears to have made the former listeners realize that the task is more difficult than they initially believed it to be. For Australian-English listeners, the magnitude of this effect was about the same when they were asked about speakers in general and about adult male Australian English speakers in particular, but for North-American-English listeners and Spanish-language listeners, the magnitude of the effect was less for speakers in general than for adult male Australian English speakers. This suggests that the latter listeners tended to attribute the difficulty of the task as due, at least in part, to the unfamiliar accent or to the unfamiliar language, and thus they remained relatively confident about their speaker-identification abilities in general. Even before the experiment, North-American-English listeners and Spanish language listeners tended to indicate that they thought they were worse at identifying adult male Australian English speakers than identifying speakers in general, with the magnitude of the difference being greater for the Spanish-language listeners. This suggests that these listeners were already aware of the difficulty due to accent unfamiliarity and greater difficulty due to language unfamiliarity.

Before the experiment, we asked listeners to estimate what their correct-classification rate would be for identifying adult male Australian English speakers, and we later compared their estimates with their actual correct-classification rates. Some listeners' estimates were close to their actual correct-classification rates, but others substantially overestimated or underestimated. The inter-listener variability (including within-language-background inter-listener variability) was such that listeners' estimated correct-classification rates could not be used as reliable indicators of actual correct-classification rates. Some of the overestimation may have been due to listeners expecting to hear high-quality recordings rather than the poor-quality and mismatched-condition recordings that they did hear, and listeners may actually have performed better on high-quality recordings. The recordings presented to them did, however, reflect the conditions of recordings in an actual case. Until they experience attempting to perform speaker identification with case-condition recordings, listeners may not appreciate the difficulty due to the conditions, and listening to a single pair of recordings, as may occur in the context of a case, might not provide sufficient experience. Australian-English listeners tended to overestimate what their correct-classification rates would be, but North-American-English listeners and Spanish language listeners both overestimated and underestimated. The general warning that listeners often substantially overestimate their speaker-identification abilities should therefore be modified to a warning that listeners often substantially overestimate their speaker-identification abilities when listening to speakers of a language and accent with which they are familiar, but could substantially overestimate or substantially underestimate their speaker-identification abilities when listening to speakers of a language or accent with which they are less familiar or not familiar. In general, listeners' estimates of their own accuracy should not be taken as indicative of their actual accuracy.

In conclusion:

- Is forensic voice comparison based on state-of-the-art automatic-speaker-recognition technology more accurate than speaker identification by individual lay listeners?
  - Yes.
- Can individual lay listeners outperform forensic voice comparison based on state-of-the-art automatic-speaker-recognition technology by considering the likelihood-ratio output of the forensic-voice-comparison system and also performing their own speaker identification?
  - No.
- Is the accuracy of individual lay listeners' speaker identification worse when the speech is in an unfamiliar accent and even worse when it is in an unfamiliar language?
  - Yes.
- Are individual lay listeners' estimates of their speaker-identification accuracy good indicators of their actual accuracy?
  - No.
- Should judges attempt to perform their own speaker identifications?
  - No. They should rely on expert testimony based on validated forensic-voice-comparison systems.
- Should judges attempt to perform their own speaker identifications in addition to considering likelihood-ratios output by validated forensic-voice-comparison systems?
  - No. They should rely exclusively on expert testimony based on validated forensic-voice-comparison systems.
- Should judges rely on speaker-identification performed by lay or "*ad hoc* expert" listeners?
  - No. They should rely exclusively on expert testimony based on validated forensic-voice-comparison systems.

The experiments reported in the present paper were conducted with individual lay listeners. They were intended to be informative of a context in which an individual judge attempts to perform speaker identification. In a future paper, we will report on experiments conducted with groups of twelve listeners acting collaboratively. Those experiments are intended to be informative of a context in which a group of jury members collaboratively attempt to perform speaker identification.

CRediT authorship contribution statement

Nabanita Basu: Methodology, Software, Writing - review & editing. **Agnes S. Bali:** Methodology, Investigation, Writing - review & editing. **Philip Weber:** Formal analysis, Writing - review & editing. **Claudia Rosas-Aguilar:** Resources, Writing - review & editing. **Gary Edmond:** Conceptualization, Writing - original draft, Writing - review & editing. **Kristy A. Martire:** Conceptualization, Methodology, Supervision, Project administration, Writing - review & editing. **Geoffrey Stewart Morrison:** Conceptualization, Methodology, Supervision, Project administration, Funding acquisition, Writing - original draft, Writing - review & editing.

Disclaimer

All opinions expressed in the present paper are those of the authors, and, unless explicitly stated otherwise, should not be construed as representing the policies or positions of any organizations with which the authors are associated.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Dr Morrison is Director and Forensic Consultant for Forensic Evaluation Ltd. Dr Weber has worked as a contractor for Forensic Evaluation Ltd. Forensic Evaluation Ltd charges clients fees to perform forensic-voice-comparison evaluations, and to submit reports and testify in court regarding forensic voice comparison, and regarding speaker recognition and speaker identification by laypersons.

Acknowledgements

This research was supported by Research England's Expanding Excellence in England Fund as part of funding for the Aston Institute for Forensic Linguistics 2019–2023.

References

[1] G. Edmond, K.A. Martire, M. San Roque, Unsound law: Issues with ('expert') voice comparison evidence, Melb. Univ. Law Rev. 35 (2011) 52–112 https://law.unimelb.edu.au/mulr/issues/previous-issues/volume35.
[2] G.S. Morrison, W.C. Thompson, Assessing the admissibility of a new generation of forensic voice comparison testimony, Columbia Sci. Technol. Law Rev. 18 (2017) 326–434, https://doi.org/10.7916/stlr.v18i2
[3] G.S. Morrison, Admissibility of forensic voice comparison testimony in England and Wales, Crim. Law Rev. 2018 (issue 1) (2018) 20–33 Preprint at http://geoff-morrison.net/#Admissibility_EW_2018.
[4] G.S. Morrison, E. Enzinger, Introduction to forensic voice comparison, in: W.F. Katz, P.F. Assmann (Eds.), The Routledge Handbook of Phonetics, Taylor & Francis, Abingdon, UK, 2019, pp. 599–634, https://doi.org/10.4324/9780429056253-22
[5] G. Edmond, Against jury comparisons, Aust. Law J. 96 (2022) 315–346.
[6] G.S. Morrison, E. Enzinger, D. Ramos, J. González-Rodríguez, A. Lozano-Díez, Statistical models in forensic voice comparison, in: D.L. Banks, K. Kafadar, D.H. Kaye, M. Tackett (Eds.), Handbook of Forensic Statistics, CRC, Boca Raton, FL, 2020, pp. 451–497, https://doi.org/10.1201/9780367527709
[7] G.S. Morrison, E. Enzinger, Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (forensic_eval_01) – Conclusion, Speech Commun. 112 (2019) 37–39, https://doi.org/10.1016/j.specom.2019.06.007
[8] C. Greenberg, A. Martin, L. Brandschain, J. Campbell, C. Cieri, G. Doddington, J. Godfrey, Human assisted speaker recognition in NIST SRE10. Proceedings of Odyssey 2010: The Speaker and Language Recognition Workshop (2010) 180–185. http://isca-speech.org/archive_open/odyssey_2010/od10_032.html.
[9] J. Kahn, N. Audibert, S. Rossato, J.F. Bonastre, Speaker verification by inexperienced and experienced listeners vs. speaker verification system, Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP) (2011) 5912–5915, https://doi.org/10.1109/ICASSP.2011.5947707
[10] D. Ramos, J. Franco-Pedroso, J. González-Rodríguez, Calibration and weight of the evidence by human listeners. The ATVS-UAM submission to NIST human-aided speaker recognition 2010, Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2011) 5908–5911 http://dx.doi.org/10.1109/ICASSP.2011.5947706.
[11] W. Shen, J.P. Campbell, D. Straub, R. Schwartz, Assessing the speaker recognition performance of naive listeners using Mechanical Turk, Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2011) 5916–5919 https://doi.org/10.1109/ICASSP.2011.5947708.
[12] R. González-Hautamäki V. Hautamäki P. Rajan T. Kinnunen Merging human and automatic system decisions to improve speaker recognition performance Proceedings of Interspeech (2013) pp. 2519–2523. http://isca-speech.org/archive/interspeech_2013/i13_2519.html.
[13] Matějka P., Glembek O., Plchot O., Schwarz M., Cipr T., Cumani S., Kudla R., Szöke I., Svobodová M., Malý K., Černocký J., 2012. BUT HASR'12 experience: Are developers of SRE systems naïve listeners? Technical Report, Brno University of Technology. http://www.fit.vutbr.cz/research/view_pub.php?id=10777.
[14] Schwartz R., Campbell J.P., Shen W., Sturim D.E., Campbell W.M., Richardson F.S., Dunn R.B., Granville R. (2011). USSS-MITLL 2010 Human assisted speaker recognition. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing* (ICASSP) (2011), pp. 5904–5907. https://dx.doi.org/10.1109/ICASSP.2011.5947705.
[15] Saeidi R., van Leeuwen D.A., (2012). The Radboud University Nijmegen submission to NIST SRE-2012. Technical Report. https://users.aalto.fi/~saeidir1/file_library/SRE12.pdf.
[16] C. Sherrin, Earwitness evidence: The reliability of voice identifications, Osgoode Hall. Law J. 52 (2016) 819–862 https://digitalcommons.osgoode.yorku.ca/ohlj/vol52/iss3/3.
[17] G.S. Morrison, E. Enzinger, C. Zhang, Forensic speech science, in: I. Freckelton, H. Selby (Eds.), Expert Evidence (Ch. 99). Thomson Reuters, Sydney, Australia, 2018. Preprint at http://expert-evidence.forensic-voice-comparison.net/.
[18] C. Rosas, J. Sommerhoff, G.S. Morrison, A method for calculating the strength of evidence associated with an earwitness's claimed recognition of a familiar speaker, Sci. Justice 59 (2019) 585–596, https://doi.org/10.1016/j.scijus.2019.07.001
[19] A. Schmidt-Nielsen, T.H. Crystal, Speaker verification by human listeners: experiments comparing human and machine performance using the NIST 1998 speaker evaluation data, Digit. Signal Process. 10 (2000) 249–266, https://doi.org/10.1006/dspr.1999.0356
[20] A. Alexander, F. Botti, D. Dessimoz, A. Drygajlo, The effect of mismatched recording conditions on human and automatic speaker recognition in forensic applications, Forensic Sci. Int. 146S (2004) S95–S99, https://doi.org/10.1016/j.forsciint.2004.09.078
[21] S.S. Kajarekar H. Bratt E. Shriberg R. de Leon A study of intentional voice modifications for evading automatic speaker recognition Proc. Odyssey: Speak. Lang. Recognit. Workshop 2006 https://doi.org/10.1109/ODYSSEY.2006.248123.
[22] V. Hautamäki T. Kinnunen M. Nosratighods K.A. Lee B. Ma H. Li Approaching human listener accuracy with modern speaker verification Proc. Inter. (2010) 1473–1476. http://isca-speech.org/archive/interspeech_2010/i10_1473.html.
[23] J. Lindh G.S. Morrison Forensic voice comparison by humans and machine: forensic voice comparison on a small database of Swedish voice recordings Proc. 17th Int. Congr. Phon. Sci. (2011) 1254–1257.
[24] M. van Dijk R. Orr D. van der Vloed D.A. van Leeuwen A human benchmark for automatic speaker recognition Proc. Biom. Technol. Forensic Sci., BTFS (2013) 39–45. https://repository.ubn.ru.nl/handle/2066/119388.
[25] L. Fernández Gallardo, Human and Automatic Speaker Recognition Over Telecommunication Channels, Springer, Singapore, 2016, https://doi.org/10.1007/978-981-287-727-7
[26] S.J. Park, G. Yeung, N. Vesselinova, J. Kreiman, P.A. Keating, A. Alwan, Towards understanding speaker discrimination abilities in humans and machines for text-independent short utterances of different speech styles, J. Acoust. Soc. Am. 144 (2018) 375–386, https://doi.org/10.1121/1.5045323
[27] D. Snyder, D. García-Romero, D. Povey, S. Khudanpur, Deep neural network embeddings for text-independent speaker verification, Proc. Inter. (2017) 999–1003, https://doi.org/10.21437/Interspeech.2017-620
[28] K.A. Lee, H. Yamamoto, K. Okabe, Q. Wang, L. Guo, T. Koshinaka, J. Zhang, K. Shinoda, NEC-TT system for mixed-bandwidth and multi-domain speaker recognition, Comput. Speech Lang. 61 (2020), https://doi.org/10.1016/j.csl.2019.101033 article 101033.
[29] P. Matějka, O. Plchot, O. Glembek, L. Burget, J. Rohdin, H. Zeinali, L. Mošner, A. Silnova, O. Novotný, M. Diez, J.H. Černocký, 13 years of speaker recognition research at BUT, with longitudinal analysis of NIST SRE, Comput. Speech Lang. 63 (2020) 101035, https://doi.org/10.1016/j.csl.2019.101035
[30] J. Villalba, N. Chen, D. Snyder, D. García-Romero, A. McCree, G. Sell, J. Borgstrom, L.P. García-Perera, F. Richardson, R. Dehak, P.A. Torres-Carrasquillo, N. Dehak, State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and Speakers in the Wild evaluations, Comput. Speech Lang. 60 (2020), https://doi.org/10.1016/j.csl.2019.101026 article 101026.
[31] G.S. Morrison, P. Weber, E. Enzinger, B. Labrador, A. Lozano-Díez, D. Ramos, J. González-Rodríguez, Forensic voice comparison – Human-supervised-automatic approach, in: M. Houck, L. Wilson, S. Lewis, H. Eldridge, P. Reedy, K. Lothridge (Eds.), Encyclopedia of Forensic Sciences, third ed., Elsevier, 2022 (in press), https://www.elsevier.com/books/encyclopedia-of-forensic-sciences/houck/978-0-12-823677-2.
[32] V. Hughes, C. Llamas, T. Kettig, Eliciting and evaluating likelihood ratios for speaker recognition by human listeners under forensically realistic channel-mismatched conditions, Proc. Inter. (2022) 5238–5242, https://doi.org/10.21437/Interspeech.2022-490
[33] G.S. Morrison, E. Enzinger, Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (forensic_eval_01) – Introduction, 2016, Speech Commun. 85 (2016) 119–126, https://doi.org/10.1016/j.specom.2016.07.006
[34] D. van der Vloed, Evaluation of Batvox 4.1 under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01), Speech Commun. 85 (2016) 127–130, https://doi.org/10.1016/j.specom.2016.10.001 [Errata in: Speech Communication, 92, 23. http://dx.doi.org/10.1016/j.specom.2017.04.005].
[35] G.D. da Silva, C.A. Medina, Evaluation of MSR identity toolbox under conditions reflecting those of a real forensic case (forensic_eval_01), Speech Commun. 94 (2017) 42–49, https://doi.org/10.1016/j.specom.2017.09.001
[36] C. Zhang, C. Tang, Evaluation of Batvox 3.1 under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01), Speech Commun. 100 (2018) 13–17, https://doi.org/10.1016/j.specom.2018.04.008
[37] M. Jessen, G. Meir, Y.A. Solewicz, Evaluation of nuance forensics 9.2 and 11.1 under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01), Speech Commun. 110 (2019) 101–107, https://doi.org/10.1016/j.specom.2019.04.006
[38] M. Jessen, J. Bortlík, P. Schwarz, Y.A. Solewicz, Evaluation of Phonexia automatic speaker recognition software under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01), Speech Commun. 111 (2019) 22–28, https://doi.org/10.1016/j.specom.2019.05.002
[39] F. Kelly, A. Fröhlich, V. Dellwo, O. Forth, S. Kent, A. Alexander, Evaluation of VOCALISE under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01), Speech Commun. 112 (2019) 30–36, https://doi.org/10.1016/j.specom.2019.06.005
[40] P. Weber, E. Enzinger, B. Labrador-Serrano, A. Lozano-Díez, D. Ramos, J. González-Rodríguez, G.S. Morrison, Validation of the alpha version of the E^3

Forensic Speech Science System (E^3FS^3) core software tools, Forensic Sci. Int.: Synerg. 4 (2022) 100223, https://doi.org/10.1016/j.fsisyn.2022.100223
[41] P. Weber, E. Enzinger, G.S. Morrison, E^3 Forensic Speech Science System (E^3FS^3): Technical report on design and implementation of software tools, 2022. Available at http://e3fs3.forensic-voice-comparison.net/.
[42] T.K. Perrachione, Speaker recognition across languages, in: S. Frühholz, P. Belin (Eds.), The Oxford Handbook of Voice Perception, Oxford: Oxford University Press, 2019, pp. 515–538 Preprint at https://hdl.handle.net/2144/23877.
[44] G.S. Morrison, N. Poh, Avoiding overstating the strength of forensic evidence: shrunk likelihood ratios / Bayes factors, Sci. Justice 58 (2018) 200–218, https://doi.org/10.1016/j.scijus.2017.12.005
[45] C. Greenberg, O. Sadjadi, E. Singer, K. Walker, K. Jones, J. Wright, S. Strassel, NIST Speaker Recognition Evaluation Test Set (LDC2020S04). Linguistic Data Consortium, https://catalog.ldc.upenn.edu/LDC2020S04.
[46] B. Sun, J. Feng, K. Saenko, Correlation alignment for unsupervised domain adaptation, in: G. Csurka (Ed.), Domain Adaptation in Computer Vision Applications. Advances in Computer Vision and Pattern Recognition, Springer, Cham, 2017, , https://doi.org/10.1007/978-3-319-58347-1_8
[47] J. Alam G. Bhattacharya P. Kenny Speaker verification in mismatched conditions with frustratingly easy domain adaptation Proc. Odyssey 2018: Speak. Lang. Recognit. Workshop (2018) pp. 176–180. https://doi.org/10.21437/Odyssey.2018-25.
[48] N. Brümmer, J. du Preez, Application independent evaluation of speaker detection, Comput. Speech Lang. 20 (2006) 230–275, https://doi.org/10.1016/j.csl.2005.08.001
[49] G.S. Morrison, E. Enzinger, V. Hughes, M. Jessen, D. Meuwly, C. Neumann, S. Planting, W.C. Thompson, D. van der Vloed, R.J.F. Ypma, C. Zhang, A. Anonymous, B. Anonymous, Consensus on validation of forensic voice comparison, Sci. Justice 61 (2021) 229–309, https://doi.org/10.1016/j.scijus.2021.02.002
[50] G.S. Morrison, In the context of forensic casework, are there meaningful metrics of the degree of calibration? Forensic Sci. Int.: Synerg. 3 (2021) 100157, https://doi.org/10.1016/j.fsisyn.2021.100157
[51] L.M. Solan, P.M. Tiersma, Hearing voices: speaker identification in court, Hastings Law J. 54 (2003) 373–436 https://repository.uchastings.edu/hastings_law_journal/vol54/iss2/2.
[52] G. Edmond, K.A. Martire, Just cognition: scientific research on bias and some implications for legal procedure and decision-making, Mod. Law Rev. 82 (4) (2019) 633–664, https://doi.org/10.1111/1468-2230.12424
[53] B.J. Dietvorst, J.P. Simmons, C. Massey, Overcoming algorithm aversion: people will use imperfect algorithms if they can (even slightly) modify them, Manag. Sci. 64 (2016) 1155–1170, https://doi.org/10.1287/mnsc.2016.2643