Forensic Science International 349 (2023) 111768

Contents lists available at ScienceDirect

Forensic Science International

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

Speaker identification in courtroom contexts – Part II: Investigation of bias in individual listeners' responses

Nabanita Basu [a,1,2], Philip Weber [a,3], Agnes S. Bali [b,4], Claudia Rosas-Aguilar [c,5], Gary Edmond [d,6], Kristy A. Martire [b,7], Geoffrey Stewart Morrison [a,e,*,8]

[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 17 October 2022
Received in revised form 18 May 2023
Accepted 20 June 2023
Available online 22 June 2023

Keywords:
Bias
Forensic voice comparison
Likelihood ratio
Recording condition
Speaker identification

ABSTRACT

In "Speaker identification in courtroom contexts – Part I" individual listeners made speaker-identification judgements on pairs of recordings which reflected the conditions of the questioned-speaker and known-speaker recordings in a real case. The recording conditions were poor, and there was a mismatch between the questioned-speaker condition and the known-speaker condition. No contextual information that could potentially bias listeners' responses was included in the experiment condition – it was decontextualized with respect to case circumstances and with respect to other evidence that could be presented in the context of a case. Listeners' responses exhibited a bias in favour of the different-speaker hypothesis. It was hypothesized that the bias was due to the poor and mismatched recording conditions. The present research compares speaker-identification performance between: (1) listeners under the original Part I experiment condition, (2) listeners who were informed ahead of time that the recording conditions would make the recordings sound more different from one another than had they both been high-quality recordings, and (3) listeners who were presented with high-quality versions of the recordings. Under all experiment conditions, there was a substantial bias in favour of the different-speaker hypothesis. The bias in favour of the different-speaker hypothesis therefore appears not to be due to the poor and mismatched recording conditions.

1. Introduction

The present paper explores whether a bias in speaker identification by lay listeners in favour of the different-speaker hypothesis, a bias that was observed in an earlier paper, is due to poor and mismatched recording conditions.

The present paper is Part II of what we expect to be a three part report on a research project that compares the performance of speaker identification by lay listeners with the performance of a forensic-voice-comparison system that makes use of state-of-the-art automatic-speaker-recognition technology. In Part I (Basu et al. [1]), the performance of speaker identification by individual lay listeners was compared with the performance of the forensic-voice-comparison system. In Part III, the performance of speaker identification by groups of lay listeners who, within each group, work collaboratively to arrive at a judgement will be compared with the performance of the forensic-voice-comparison system. The research in Part I was intended to be informative with respect to a courtroom context in which a judge (an individual) listens to a recording of a speaker whose identity is in question and to a recording of a known speaker (or to that speaker speaking live in court), and makes a decision as to whether the recordings are of the same speaker or are of different speakers. The research in Part III is intended to be informative with respect to a courtroom context in which members of a jury (a group)

* 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] Now at Forensic Science Research Group, Department of Applied Sciences, Northumbria University, Newcastle upon Tyne, UK.
[3] ORCID: 0000-0002-3121-9625.
[4] ORCID: 0000-0002-0166-0989.
[5] ORCID: 0000-0002-8544-7965.
[6] ORCID: 0000-0003-2609-7499.
[7] ORCID: 0000-0002-5324-0732.
[8] ORCID: 0000-0001-8608-8207.

https://doi.org/10.1016/j.forsciint.2023.111768

listen and collaboratively make a decision as to whether the recordings are of the same speaker or are of different speakers. The questions of whether a forensic-voice-comparison system is more or less accurate than a judge listening and making a judgement alone, or whether a forensic-voice-comparison system is more or less accurate than a jury listening and making a judgement as a collaborative group, are important because expert testimony is only admissible in common law if it will potentially assist the trier of fact to make a decision. If the trier of fact's speaker identification were equally accurate or more accurate than the output of the forensic-voice-comparison system, then testimony based on the output of the forensic-voice-comparison system would not be admissible. Part I §1.2 discussed the legal context related to forensic voice comparison conducted by experts and speaker identification performed by triers of fact. Part I §1.3 described prior research on speaker identification by lay listeners.

The stimuli in Part I consisted of pairs of recordings that reflected the conditions of a questioned-speaker recording and a known-speaker recording in a real case. The recording conditions were poor and there was a mismatch in recording conditions between the questioned-speaker recording and the known-speaker recording. The questioned-speaker condition reflected a landline-telephone call, with background babble noise, saved using lossy compression, and the known-speaker condition reflected an interview recorded in a reverberant room, with background ventilation-system noise. Each questioned-speaker-condition recording and each known-speaker-condition recording was ∼15 s long. There were 31 same-speaker pairs of recordings and 30 different-speaker pairs of recordings.[9] Under these conditions, in terms of C_{llr} (see §2.7.2 below), all of the individual listeners in Part I performed worse than the forensic-voice-comparison system.

One way in which the listeners in Part I performed worse than the forensic-voice-comparison system was with respect to bias. The forensic-voice-comparison system was calibrated using a logistic-regression model which was trained using data that reflected the same conditions as the validation data.[10] Relative to the likelihood-ratio values output by the forensic-voice-comparison system, the responses of more than 90% of the listeners exhibited bias that favoured the different-speaker hypothesis. There was substantial inter-listener variability, but Fig. 1(a) gives an example Tippett plot showing a common pattern of bias in a listener's responses. For comparison purposes, Fig. 1(b) shows the validation results from the forensic-voice-comparison system.

In Part I, the speakers spoke Australian-accented English, and speaker identification by listeners with three different language backgrounds (Australian-English listeners, North-American-English listeners, and Spanish-language listeners) was tested. Within each language background, there was substantial inter-listener variability, but bias in favour of the different-speaker hypothesis was observed for listeners from all three language backgrounds (see Part I §3.2.4). In terms of B_{llr}, a metric assessing the bias in the listeners' responses relative to the likelihood-ratio values output by the forensic-voice-comparison system (see §2.7.4 below), the median and quartile values for listeners from the different language backgrounds were similar. Across language backgrounds, the median relative bias was such that likelihood-ratio values were, on average, just above half those of the forensic-voice-comparison system.

The experiments in Part I did not include any contextual information that would be expected to bias the listeners – they did not contain any information about the circumstances of a case or information about any other evidence in the case. In Part I, we hypothesized that the observed bias in favour of the different-speaker hypothesis could have been due to the poor recording conditions and the mismatches in recording conditions between the questioned-speaker-condition and 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 present paper investigates:

- Whether providing information about the recording conditions ahead of time would result in a reduction in observed bias, i.e., if the listeners are warned ahead of time that the recordings conditions will make the questioned-speaker and known-speaker recordings sound more different from one another, will this reduce the degree of bias in favour of the different-speaker hypothesis?
- Whether the observed bias is actually due to the recording conditions, i.e., if listeners are presented with high-quality versions of the questioned-speaker and known-speaker recordings, will this eliminate the average bias in favour of the different-speaker hypothesis?

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: Behavioural Sciences, and from the Aston Institute for Forensic Linguistics Research Ethics Committee.

2.2. Stimuli

Stimuli consisted of recordings of adult male speakers of Australian English. Two different sets of stimuli were used. Both sets consisted of 31 same-speaker pairs of recordings and 27 different-speaker pairs of recordings.[11]

One set of stimuli was the same as used in Part I, except that it had three fewer different-speaker pairs. In each stimulus pair, the questioned-speaker condition reflected a landline-telephone call, with background babble noise, saved using lossy compression, and the known-speaker condition reflected an interview recorded in a reverberant room, with background ventilation-system noise. The recordings were taken from the *forensic_eval_01* dataset (Morrison and Enzinger [2]), which had previously been used to validate multiple forensic-voice-comparison systems. The pairs of recordings used were a subset of those used for validating forensic-voice-comparison systems, and each recording was shortened to ∼15 s in duration (∼15 s long sections were randomly selected from within each recording). See Part I §2.2 for further details about the construction of these stimuli.

The other set of stimuli consisted of recordings of exactly the same speech by the same speakers as in the first set of stimuli, but all recordings, both questioned-speaker and known-speaker condition, were high-quality recordings. These recordings were manually extracted from the AusEng 500+ database (Morrison et al. [5]). The AusEng 500+ database was the base that had been used to create the *forensic_eval_01* dataset, which was in turn the base for the stimuli used in Part I. In creating the *forensic_eval_01* dataset, noise and reverberation had been added and codecs applied to high-quality recordings from the AusEng 500+ database (see Enzinger et al. [3] for details).

[9] For further information about the stimuli used in Part I, see Part I §2.2. For further information about the source database, see Morrison and Enzinger [2]. For information about the simulation of the recording conditions, see Enzinger et al. [3]. For information about the data-collection protocols, see Morrison et al. [4].

[10] Details of the training of the calibration model are provided in Part I §2.5.

[11] Part I used 30 different-speaker pairs. We were unable to locate high-quality versions of three recordings, resulting in us using only 27 different-speaker pairs for the present research.

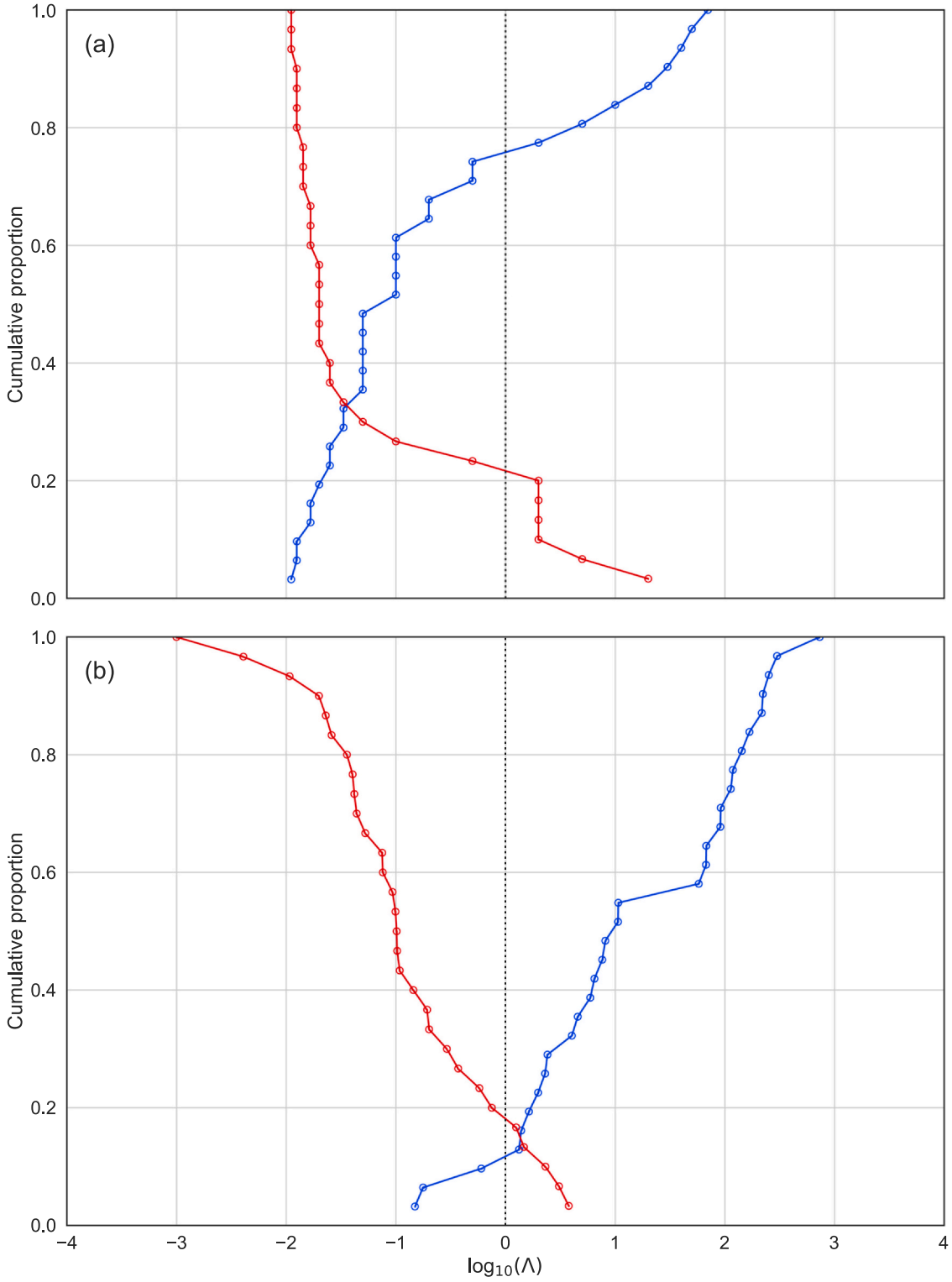

Fig. 1. (a) Example Tippett plot of the results from a listener in Part I whose responses were biased toward the different-speaker hypothesis. $C_{llr} = 1.90$, $D_{llr} = -2.5$, $B_{llr} = -3.5$. (b) Tippett plot of the validation results in Part I for the forensic-voice-comparison system. $C_{llr} = 0.42$. (Explanations of C_{llr}, D_{llr}, and B_{llr} are provided in §2.7 below.)

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

2.3. Forensic-voice-comparison system

The E^3 Forensic Speech Science 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. For detailed descriptions of the forensic-voice-comparison system, see Morrison et al. [6] and Weber et al. [7,8].

The system, and how it was trained and calibrated for the original poor-quality recording conditions, was described in Part I §2.5. For the high-quality recording conditions, the system was retrained and recalibrated using 15 s long non-overlapping sections that were selected from the AusEng 500+ database. The recordings selected were the high-quality versions of the poor-quality recordings used in Part I, except for the exclusion of three recordings for which we were unable to locate high-quality versions.

2.4. Participants (listeners)

Participants were recruited using an online recruitment platform, Prolific.[12] The experiments were advertised as taking up to 2 h to complete, and participants who completed an experiment were paid GBP 21 (at the time the research began, this was the amount recommended by Prolific for 2 h of participant time).

For Part I, we recruited Australian-English listeners, North-American-English listeners, and Spanish-language listeners. For the present research, we recruited North-American-English listeners and Australian-English listeners. We recruited three non-overlapping groups of North-American-English listeners (including the original group of North-American-English listeners in Part I), and two non-overlapping groups of Australian-English listeners (including the original group of Australian-English listeners in Part I).[13] The target number of listeners to recruit for each group 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
3. were currently a resident of the United States or Canada / were currently a resident of Australia
4. had lived for at least 4 years in the United States or Canada, or were a citizen of the United States or Canada / had lived for at least 4 years in Australia, or were a citizen of Australia
5. had completed at least an undergraduate degree
6. did not have a diagnosed hearing loss

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, see Part I §2.3 for details.

2.5. 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.

The experiment procedures were identical to those in the main experiments in Part I: After completing eligibility questions, providing informed consent, and answering demographic questions, each participant was presented with written instructions explaining the task,[14] 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 58 experiment trials in random order, a different random order for each participant. Randomly mixed with the experiment trials were 4 attention-check trials. Each trial screen included a counter showing the number of trials completed out of the total (63 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.

A screenshot of an experiment trial is shown in Fig. 2. 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

[12] https://prolific.co/.

[13] We did not attempt to recruit three groups of Australian-English listeners because the pool of Australian-English listeners on Prolific was small. In contrast, of the three language backgrounds, North-American-English listeners constituted the largest pool of potential participants on Prolific.

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

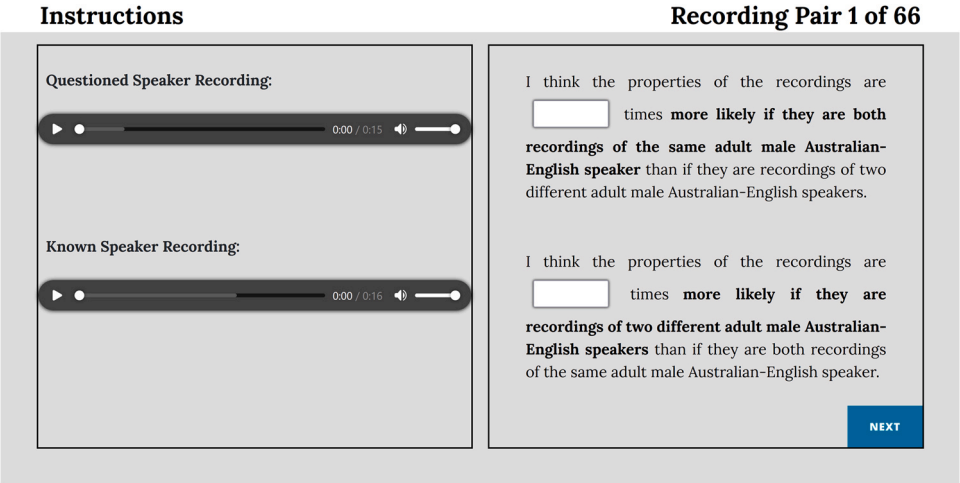

Fig. 2. Screenshot of an experiment trial in the individual-listener experiment.

of the same speaker or recordings of different speakers, they should enter 1 in either one of the boxes.[15]

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.

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.

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 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.

2.6. Experiment conditions

We ran three different versions of the experiment, each with a different set of conditions:

1. The original experiment condition in Part I. Recording conditions were poor and there was a mismatch between the questioned-speaker condition and the known-speaker condition. Before they started the experiment, listeners were not told what the recording conditions would be.
2. An experiment condition that was identical to Experiment Condition 1, except that, before they started the experiment, listeners were told about the recording conditions and their potential effect.
3. An experiment condition that was identical to Experiment Condition 1, except that the recordings were all high-quality audio and there was no mismatch between the questioned-speaker-condition and known-speaker-condition recordings.

For Experiment Condition 2, the following text was included near the end of the instructions:

> The two recordings in each pair were recorded under different conditions. The questioned-speaker recording is a recording of a telephone call and the known-speaker recording was made in a police interview room. Because both are relatively poor-quality recordings, and because there is a mismatch in recording conditions, the voices on the two recordings will sound more different from one another than they would if they were both recorded under good conditions.

This is one of the types of information that might be provided by an expert witness testifying about speaker identification by lay listeners, and that might be included by a judge in instructions to jurors.

Each group of listeners was exposed to a different experiment condition. No listener was exposed to more than one experiment condition. Hereinafter, each group of listeners is referred to using its accent and the number of the experiment condition to which they were exposed, i.e., Group 1, Group 2, and Group 3, correspond to Experiment Condition 1, Experiment Condition 2, and Experiment Condition 3 respectively. North-American-English listeners completed Experiment Conditions 1, 2, and 3. Australian-English listeners completed Experiment Conditions 1 and 3. The pool of Australian-English listeners on Prolific was small, so we concentrated recruitment of Australian-English listeners for only the latter two conditions.

[15] 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 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.

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; 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:[16]

- 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.

These are the same metrics as were used in Part I. See Part I §3.2.5 for examples of Tippett plots showing a range of different values for each of these metrics.

In addition to these likelihood-ratio-based metrics, we also calculated the miss rate and the false-alarm rate for the forensic-voice-comparison system and for each individual listener's responses. We would not do this in the context of forensic casework. We do it here only to allow for potential comparison with other studies that have collected categorical responses, which is the case for almost all previous studies (see Part I §1.3).

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 (Brümmer and de Preez [9]). C_{llr} was calculated using Eq. (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.

$$C_{\text{llr}} = \frac{1}{2}\left(\frac{1}{N_{\text{s}}}\sum_{i=1}^{N_{\text{s}}}\log_2\left(1+\frac{1}{\Lambda_{\text{s}_i}}\right)+\frac{1}{N_{\text{d}}}\sum_{j=1}^{N_{\text{d}}}\log_2(1+\Lambda_{\text{d}_j})\right) \tag{1}$$

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* [10]. 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 Eq. (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.

$$D_{\text{llr}} = \frac{1}{2}\left(\frac{1}{N_{\text{s}}}\sum_{i=1}^{N_{\text{s}}}\left(\log_2(\Lambda_{\text{h,s}_i})-\log_2(\Lambda_{\text{f,s}_i})\right)+\frac{1}{N_{\text{d}}}\sum_{j=1}^{N_{\text{d}}}\left(\log_2(\Lambda_{\text{f,d}_i})-\log_2(\Lambda_{\text{h,d}_i})\right)\right) \tag{2}$$

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 Eq. (3).[17] 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.

$$B_{\text{llr}} = \frac{1}{2}\left(\frac{1}{N_{\text{s}}}\sum_{i=1}^{N_{\text{s}}}\left(\log_2(\Lambda_{\text{h,s}_i})-\log_2(\Lambda_{\text{f,s}_i})\right)+\frac{1}{N_{\text{d}}}\sum_{j=1}^{N_{\text{d}}}\left(\log_2(\Lambda_{\text{h,d}_i})-\log_2(\Lambda_{\text{f,d}_i})\right)\right) \tag{3}$$

2.7.5. Miss rate and false-alarm rate

In order to calculate miss rates and false-alarm rates, we ignored the magnitudes of the likelihood-ratio responses and counted values greater than 1 as if they were categorical same-speaker responses

[16] 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.

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

and values less than 1 as if they were categorical different-speaker responses.

There has been debate in the literature as to how to treat "inconclusive" responses in error-rate calculations for contexts in which practitioners give "same-source", "inconclusive", or "different-source" conclusions (traditionally, "identification", "inconclusive", "exclusion"). Some argue that "inconclusives" should be counted as errors. Others argue that "inconclusives" should not be counted at all. Our perspective on this is that forensic-evaluation systems should output likelihood ratios, and that the appropriate metric to calculate is therefore C_{llr}, not classification-error rate (or its components miss rate and false-alarm rate), hence the debate is misplaced (see Morrison [11]). In the current research, listeners had the option to respond with a likelihood ratio of 1. With respect to responses of "1", we calculated miss rates and false-alarm rates using two different procedures:

- Responses of "1" treated as errors. If the pair of recordings was a same-speaker pair, and the listener responded "1", the response was treated as a miss. If the pair of recordings was a different-speaker pair, and the listener responded "1", the response was treated as a false alarm.
- Responses of "1" ignored. Only responses for which listeners gave values other than 1 were included in the miss rate and false-alarm rate calculations. For example, if, in response to the 31 same-speaker pairs of recordings, a listener gave 11 responses of "1", 15 responses in the top box greater than 1 (corresponding to likelihood ratios greater than 1), and 5 responses in the bottom box greater than 1 (corresponding to likelihood ratios less than 1), the miss rate was calculated as 5/(15+5) = 25%.

For the stimuli in the current research, the forensic-voice-comparison system never gave a likelihood ratio of exactly 1, or even a value that rounded to 1.

3. Results

3.1. Demographics

3.1.1. Exclusions

We excluded from analysis the submissions from listeners who did not answer all of the attention-check trials correctly,[18] 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.[19] The results reported below are those after the removal of these submissions.

3.1.2. North-American-English listeners

For North-American-English listeners there were:

- 61 submissions from Group 1
  - for reported ages, minimum, lower quartile, median, upper quartile, and maximum were 22, 27, 32, 39, and 72 years respectively
  - 23 identified as females, 34 as males, and the remainder responded "other" or "prefer not to say"
  - 53 identified as first-language English speakers, and the remainder as having other first languages
  - 43 stated that they were "somewhat familiar" and 18 that they were "not familiar" with Australian English
- 53 submissions from Group 2
  - for reported ages, minimum, lower quartile, median, upper quartile, and maximum were 20, 28, 36, 48, and 68 years respectively
  - 25 identified as females, 26 as males, and the remainder responded "other" or "prefer not to say"
  - 48 identified as first-language English speakers, and the remainder as having other first languages
  - 34 stated that they were "somewhat familiar", and 19 that they were "not familiar" with Australian English
- 51 submissions from Group 3
  - for reported ages, minimum, lower quartile, median, upper quartile, and maximum were 20, 28, 34, 40, and 60 years respectively
  - 19 identified as females and 32 as males
  - 49 identified as first-language English speakers, and the remainder as having other first languages
  - 1 stated that they were "very familiar", 37 that they were "somewhat familiar", and 13 that they were "not familiar" with Australian English

Group 1 was the same group of North-American-English listeners who participated in the main experiment in Part I, and the responses analyzed were the same responses except that they only included responses to 27 different-speaker pairs.[20]

3.1.3. Australian-English listeners

For Australian-English listeners there were:

- 53 submissions from Group 1
  - 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, and the remainder as having other first languages
- 55 submissions from Group 3
  - for reported ages, minimum, lower quartile, median, upper quartile, and maximum were 22, 25, 30, 36, and 65 years respectively
  - 30 identified as females and 25 as males
  - 53 identified as first-language English speakers, and the remainder as having other first languages

Group 1 was the same group of Australian-English listeners who participated in the main experiment in Part I, and the responses analyzed were the same responses except that they only included responses to 27 different-speaker pairs.

3.2. Performance metrics

3.2.1. Forensic-voice-comparison system

Given the 31 same-speaker pairs and the 27 different-speaker pairs of recordings, the C_{llr} value for the forensic-voice-comparison system under the original poor-quality recording conditions (which reflected the conditions of a real case) was 0.40. Under the high-

[18] 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".

[19] Although a first-accent Australian-English listener who had lived in the US or Canada for more than 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).

[20] The three additional pairs of recordings to which listeners in Conditions 1 and 2 responded means that the context in Conditions 1 and 2 is different from Condition 3. We do not, however, expect this small difference in context to substantially impact results.

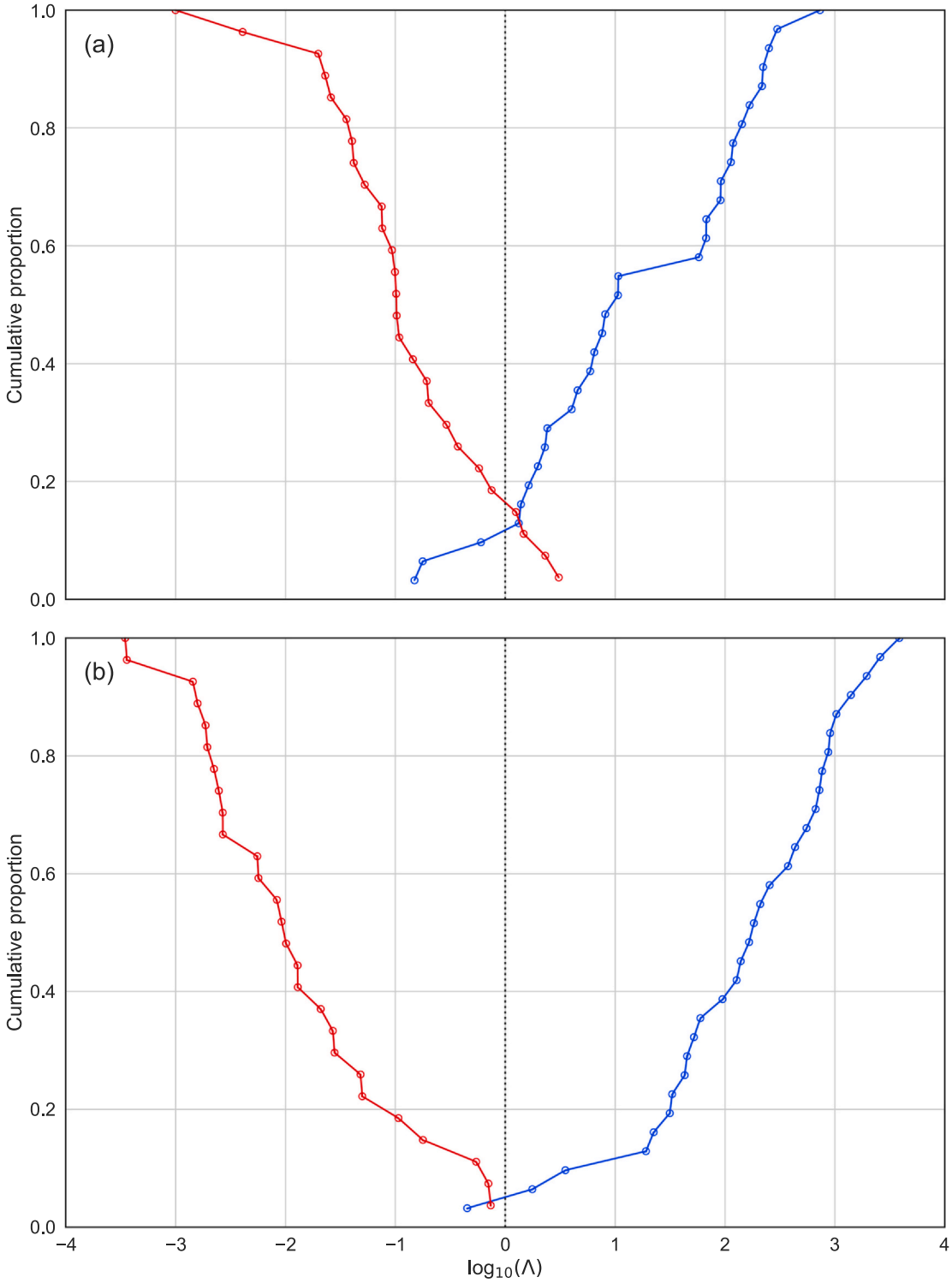

Fig. 3. Tippett plot of validation results from the forensic-voice-comparison system under (a) the original poor-quality recording conditions but including responses to only 27 different-speaker pairs (C_{llr} = 0.40), and (b) the high-quality recording conditions (C_{llr} = 0.10).

quality recording conditions, the C_{llr} value for the forensic-voice-comparison system was 0.10. Improved performance under better conditions is as would be expected. Tippett plots of results from the forensic-voice-comparison system under the original poor-quality recording conditions and under the high-quality recording conditions are given in Fig. 3.

3.2.2. North-American-English listeners

A C_{llr} value, a D_{llr} value, and a B_{llr} value was calculated separately for each individual North-American-English listener's responses.

Fig. 4 shows violin plots of the C_{llr} values grouped by experiment condition. The heavy black horizontal lines indicate the C_{llr} value for the forensic-voice-comparison system under the original poor-quality

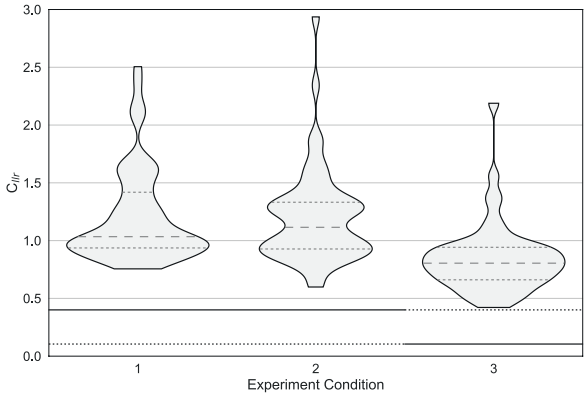

Fig. 4. Violin plots of the C_{llr} values for the responses from individual North-American-English listeners. The heavy black horizontal lines indicate the C_{llr} value for the forensic-voice-comparison system under the original poor-quality recording conditions (line at C_{llr} = 0.40) and under the high-quality conditions (line at C_{llr} = 0.10).

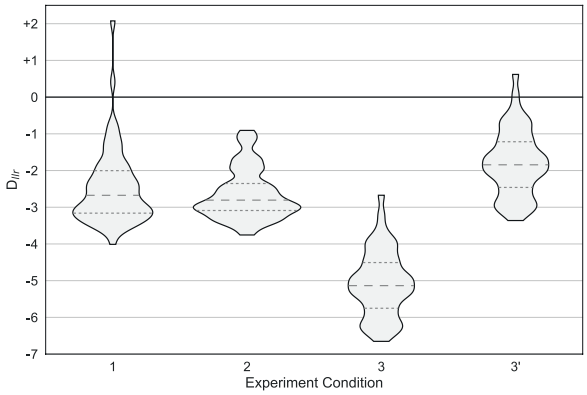

Fig. 5. Violin plots of the D_{llr} values for the comparison of individual North-American-English listeners' responses with the responses of the forensic-voice-comparison system.

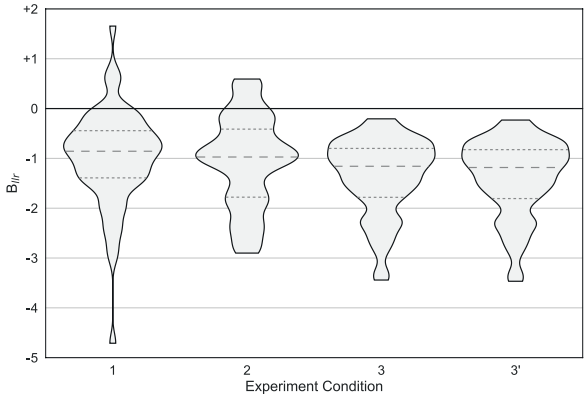

Fig. 6. Violin plots of the B_{llr} values for the comparison of individual North-American-English listeners' responses with the responses of the forensic-voice-comparison system.

recording conditions (for Experiment Conditions 1 and 2) and under the high-quality recording conditions (Experiment Condition 3). The heavy black horizontal lines are drawn solid under the relevant experiment condition(s) and dotted under the other experiment condition(s).

Figs. 5 and 6 show, respectively, violin plots of the D_{llr} values grouped by experiment condition and the B_{llr} values grouped by experiment condition. For Experiment Conditions 1 and 2, D_{llr} and B_{llr} values were calculated relative to the likelihood-ratio values output by the forensic-voice-comparison system under the original poor-quality recording conditions, and, for Experiment Condition 3, they were calculated relative to the likelihood-ratio values output by the forensic-voice-comparison system under the high-quality recording conditions. In addition, for Experiment Condition 3, D_{llr} and B_{llr} values were calculated relative to the likelihood-ratio values output by the forensic-voice-comparison system under the original poor-quality recording conditions. In Figs. 5 and 6, the latter results are labelled 3′. The latter results do not constitute a comparison of the performance of listeners against the performance of the forensic-voice-comparison-system under the same conditions, but, instead, allow for a comparison of the performance of (different) listeners in different conditions.

Fig. 7 shows plots of miss rates versus false-alarm rates for each of Experiment Conditions 1, 2, and 3. The top row shows results calculated using the procedure that counted responses of "1" as errors, and the bottom row shows results calculated using the procedure that ignored responses of "1". Diagonal grid lines running from upper left to lower right indicate combinations of miss rates and false-alarm rates with the same classification-error rates (the classification-error rate was calculated as the mean of the miss rate and the false-alarm rate). Symbols above and to the right of the 50% diagonal represent classification-error rates that are worse than what would be expected from randomly guessing same-speaker or different-speaker (note that axes do not extend to 100%). The filled diamonds show the results for the forensic-voice-comparison system. The unfilled circles show the results for individual listeners. Symbols in the upper left of a panel indicate bias in favour of the different-speaker hypothesis. Symbols in the lower right of a panel indicate bias in favour of the same-speaker hypothesis. The further the symbol from the heavy black diagonal line running bottom left to top right, the greater the bias.

3.2.3. Australian-English listeners

A C_{llr} value, a D_{llr} value, and a B_{llr} value was calculated separately for each individual Australian-English listener's responses.

Fig. 8 shows violin plots of the C_{llr} values grouped by experiment condition. The heavy black horizontal lines indicate the C_{llr} value for the forensic-voice-comparison system under the original poor-quality recording conditions (Experiment Condition 1) and under the high-quality recording conditions (Experiment Condition 3). The heavy black horizontal lines are drawn solid under the relevant experiment condition and dotted under the other experiment condition.

Figs. 9 and 10 show, respectively, violin plots of the D_{llr} values grouped by experiment condition and the B_{llr} values grouped by experiment condition. For Experiment Condition 1, D_{llr} and B_{llr} values were calculated relative to the likelihood-ratio values output by the forensic-voice-comparison system under the original poor-quality recording conditions, and, for Experiment Condition 3, they were calculated relative to the likelihood-ratio values output by the forensic-voice-comparison system under the high-quality recording conditions. In addition, for Experiment Condition 3, D_{llr} and B_{llr} values were calculated relative to the likelihood-ratio values output by the forensic-voice-comparison system under the original poor-quality recording conditions. In Figs. 9 and 10, the latter results are labelled 3′. The latter results do not constitute a comparison of the performance of listeners against the performance of the forensic-voice-comparison-system under the same conditions, but, instead, allow for a comparison of the performance of (different) listeners in different conditions.

Fig. 11 shows plots of miss rates versus false-alarm rates for each of Experiment Conditions 1 and 3. The top row shows results calculated using the procedure that counted responses of "1" as errors, and the bottom row shows results calculated using the procedure that ignored responses of "1". Diagonal grid lines running from

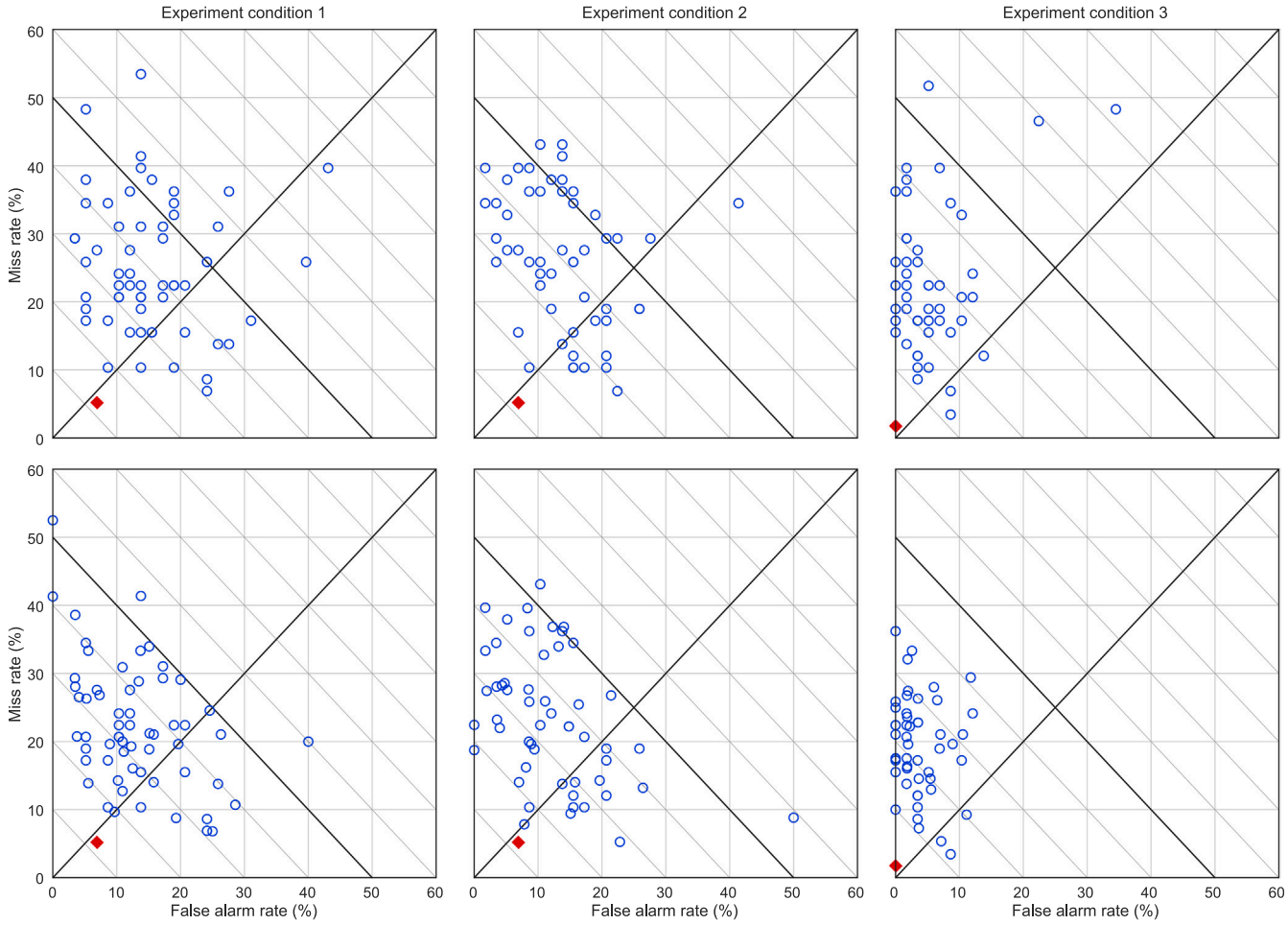

Fig. 7. Plots of miss rate versus false-alarm rate for the forensic-voice-comparison system (filled diamond) and for individual North-American-English listeners' responses (unfilled circles). Top panels: Responses of "1" treated as errors. Bottom panels: Responses of "1" ignored.

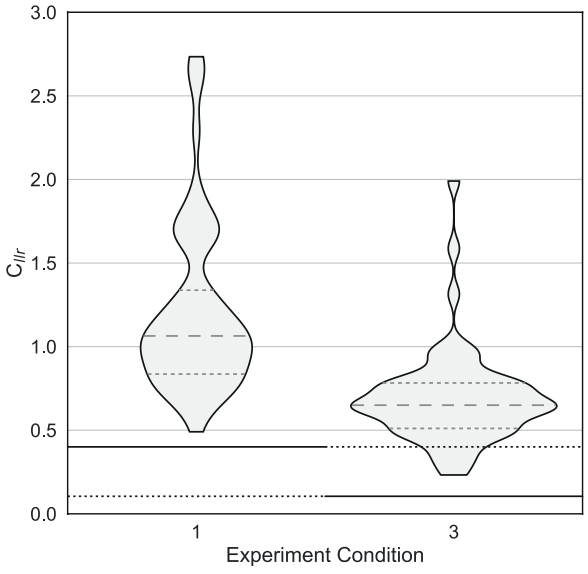

Fig. 8. Violin plots of the C_{llr} values for the responses from individual Australian-English listeners. The heavy black horizontal lines indicate the C_{llr} value for the forensic-voice-comparison system under the original poor-quality recording conditions (line at C_{llr} = 0.40) and under the high-quality conditions (line at C_{llr} = 0.10).

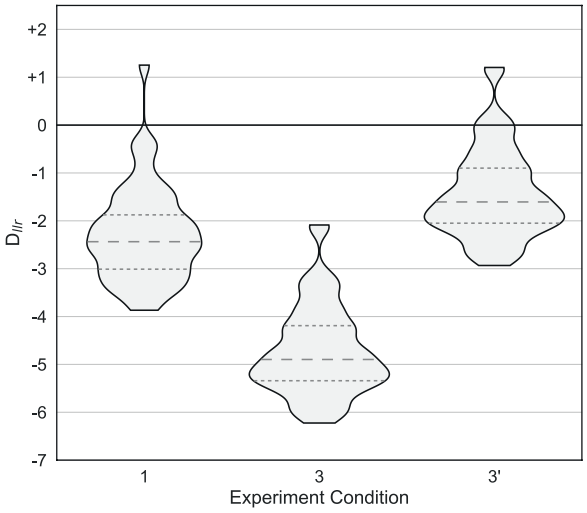

Fig. 9. Violin plots of the D_{llr} values for the comparison of individual Australian-English listeners' responses with the responses of the forensic-voice-comparison system.

upper left to lower right indicate combinations of miss rates and false-alarm rates with the same classification-error rates. Symbols above and to the right of the 50% diagonal represent classification-error rates that are worse than what would be expected from randomly guessing same-speaker or different-speaker (note that axes

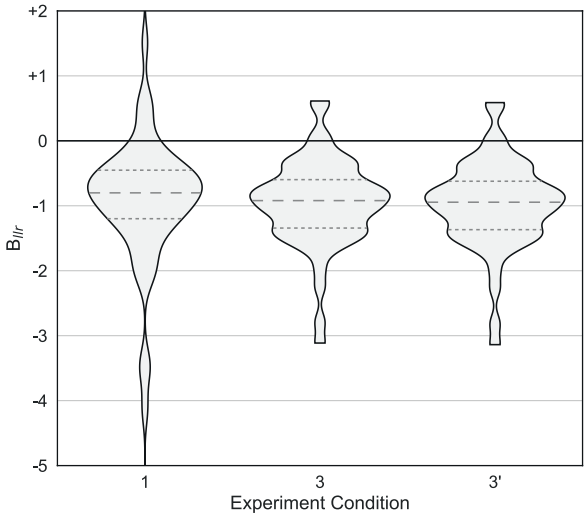

Fig. 10. Violin plots of the B_{llr} values for the comparison of individual Australian-English listeners' responses with the responses of the forensic-voice-comparison system.

do not extend to 100%). The filled symbol shows the results for the forensic-voice-comparison system. The unfilled symbols show the results for individual listeners. Symbols in the upper left of a panel indicate bias in favour of the different-speaker hypothesis. Symbols in the lower right of a panel indicate bias in favour of the same-speaker hypothesis. The further the symbol from the heavy black diagonal line running bottom left to top right, the greater the bias. For Condition 1, results from one listener, who always responded "1", are not shown.

4. Discussion

In terms of C_{llr}, as expected, speaker-identification performance was better for listeners who heard high-quality versions of the recordings than for listeners who heard poor-quality versions of the recordings with a mismatch between the questioned-speaker condition and the known-speaker condition. This was the case for both North-American-English listeners and for Australian-English listeners. For North-American-English listeners, the lower quartile, median, and upper quartile for C_{llr} values under Experiment Condition 3 were all substantially lower than those under Experiment Conditions 1 and 2 (see Fig. 4). There was no obvious consistent difference in C_{llr} values between Experiment Conditions 1 and 2. For Australian-English listeners, the lower quartile, median, and upper quartile for C_{llr} values under Experiment Condition 3 were all substantially lower than those under Experiment Condition 1 (see Fig. 8) Australian-English listeners did not provide responses for Experiment Condition 2.

A key finding in Part I was that, under poor-quality recordings conditions, in terms of C_{llr}, all listeners performed worse than the forensic-voice-comparison system. Although, in the present research, listeners who heard high-quality recordings (Experiment Condition 3) performed better than listeners who heard poor-quality recordings (Experiment Conditions 1 and 2), all listeners who heard high-quality recordings performed worse than the forensic-voice-comparison system did with respect to the high-quality recordings. Under the high-quality condition, the lowest C_{llr} for a North-American-English listener was 0.42, and the lowest C_{llr} for an Australian-English listener was 0.23, compared to 0.10 for the forensic-voice-comparison system.

Under the high-quality condition, the forensic-voice-comparison system also outperformed all the listeners with respect to classification-error rate, if only just. The forensic-voice-comparison system made 1 error out of 58 responses (a single miss), an error rate of 1.7%.[21] When responses of "1" were counted as errors, the lowest error rate for a North-American-English listener was 12%, 7 out of 58 responses (one listener had 5 misses and 2 false alarms, and another had 2 misses and 5 false alarms, see Fig. 7), and the lowest error rate for an Australian-English listener was 3.5%, 2 out of 58 responses (1 miss and 1 false alarm, see Fig. 11). When responses of "1" were ignored, the lowest error rate for a North-American-English listener was 10%, 2 out of 20 responses (2 misses), and the lowest error rate for an Australian-English listener was 1.8%, 1 out of 57 responses (a single false alarm).

D_{llr} results indicated that, compared to the forensic-voice-comparison system, listeners' 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, i.e., compared to the forensic-voice-comparison system, listeners under-discriminated between same-speaker and different-speaker pairs of recordings. Under-discrimination was greater for listeners who heard high-quality versions of the recordings than for listeners who heard poor-quality versions of the recordings. For North-American-English listeners, the lower quartile, median, and upper quartile for D_{llr} values under Experiment Condition 3 were all substantially lower than those under Experiment Conditions 1 and 2 (see Fig. 5). For Experiment Condition 2, the upper quartile and median for D_{llr} values were slightly lower than they were for Experiment Condition 1, but the lower quartile was slightly higher. For Australian-English listeners, the lower quartile, median, and upper quartile for D_{llr} values under Experiment Condition 3 were all substantially lower than those under Experiment Condition 1 (see Fig. 9). Australian-English listeners did not provide responses for Experiment Condition 2.

The difference in D_{llr} values between the poor-quality condition (Experiment Conditions 1 and 2) and the high-quality condition (Experiment Condition 3) is due to the improved performance of the forensic-voice-comparison system on the high-quality condition. When listeners' responses under the high-quality condition were compared with the forensic-voice-comparison system's responses under the poor-quality condition (see results labelled 3′ in Figs. 5 and 9) D_{llr} values were actually higher than for Experiment Conditions 1 and 2,[22] i.e., listeners who heard high-quality versions of the recordings discriminated between same-speaker pairs and different-speaker pairs better than listeners who heard poor-quality versions of the recordings, but the improvement in discrimination for the forensic-voice-comparison system under the high-quality recording condition was much greater.

The key metric to consider with respect to whether Experiment Condition 2 or Experiment Condition 3 resulted in reduction or elimination of the bias observed in Experiment Condition 1 (the original poor-quality condition) is B_{llr}. If providing the information about the poor-quality recording conditions in Experiment Condition 2 resulted in a reduction in the bias in favour of the different-speaker hypothesis compared to Experiment Condition 1 in which no information about the recording conditions was provided, then the B_{llr} values would be expected to be less negative for Experiment Condition 2 than they were for Experiment Condition 1. Similarly, if providing high-quality versions of the recordings in Experiment Condition 3 resulted in a reduction in the bias in favour of the different-speaker hypothesis compared to providing poor-quality versions of the recordings in Experiment Condition 1, then

[21] Percentage error rates reported in the present paragraph are calculated as the number of errors divided by the total number of responses (all 58 responses, or all non "1" responses), without giving equal weight to miss rate and false-alarm rate.

[22] D_{llr} values for Experiment Condition 3′, however, were still almost all negative.

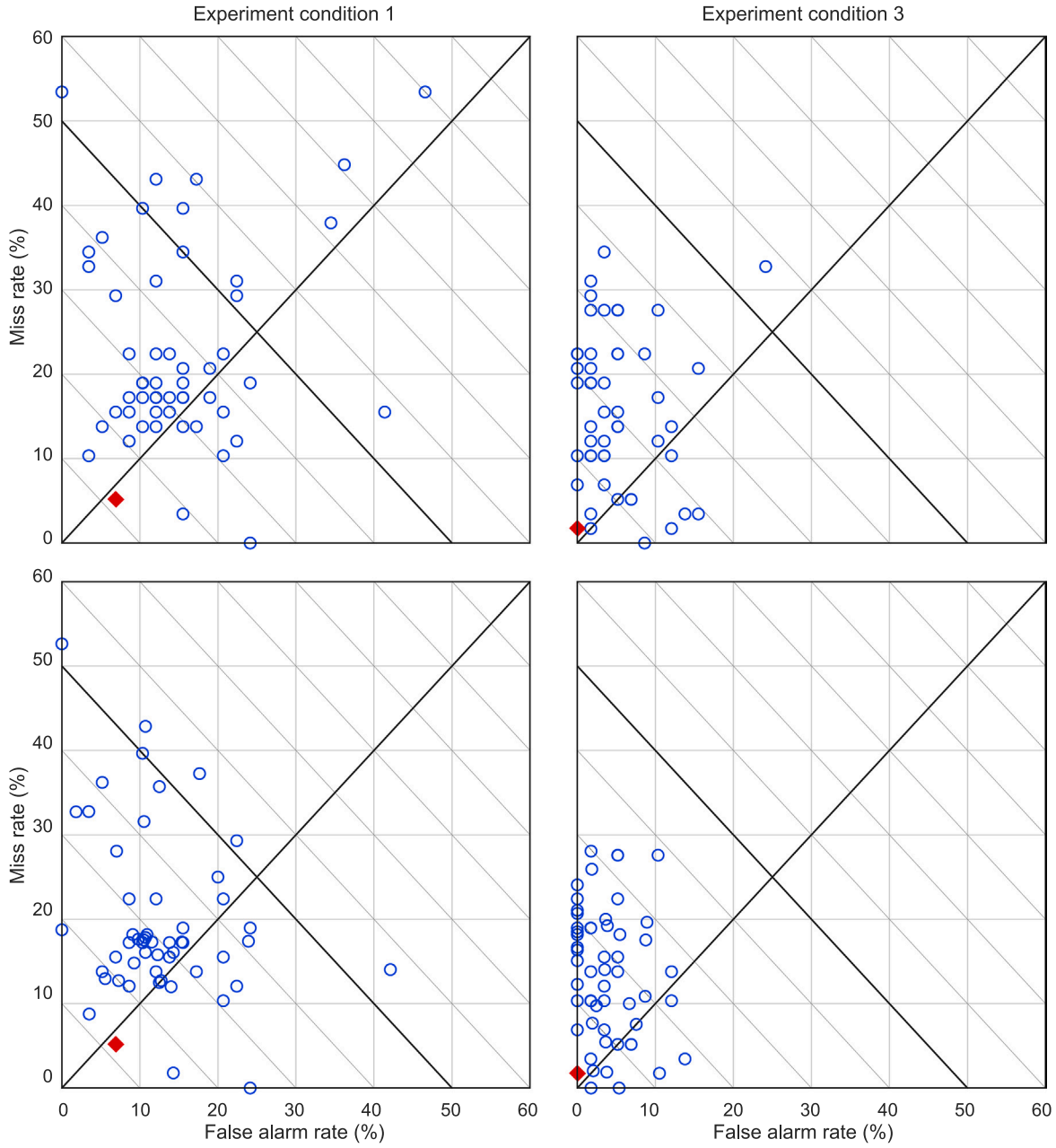

Fig. 11. Plots of miss rate versus false-alarm rate for the forensic-voice-comparison system (filled diamond) and for individual Australian-English listeners' responses (unfilled circles). Top panels: Responses of "1" treated as errors. Bottom panels: Responses of "1" ignored.

the B_{llr} values would be expected to be less negative for Experiment Condition 3 than they were for Experiment Condition 1.

For North-American-English listeners, in terms of B_{llr} values, compared to Experiment Condition 1, for Experiment Condition 2 the upper quartile was almost identical, the median was slightly lower, and the lower quartile was substantially lower (see Fig. 6). The difference between B_{llr} results is small, but what difference there is in the opposite direction to that expected if there had been a reduction in bias. We conclude that the information we provided about recording conditions in Experiment Condition 2 did not result in a reduction in the bias in favour of the different-speaker hypothesis.

In terms of B_{llr} values, compared to Experiment Condition 1, for Experiment Condition 3 the upper quartile, the median, and the lower quartile were all substantially lower for North-American-English listeners, and slightly lower for Australian-English listeners (see Figs. 6 and 10). The differences in B_{llr} results between Experiment Conditions 1 and 3 are in the opposite direction to what would be expected if there had been a reduction in bias. We conclude that providing high-quality recordings in Experiment Condition 3 instead of the low-quality recordings provided in Experiment Condition 1 did not result in a reduction in the bias in favour of the different-speaker hypothesis. If anything, it resulted in an increase in the bias in favour of the different-speaker hypothesis.

Bias in favour of the different-speaker hypothesis was also apparent in terms of miss rates and false-alarm rates: Most listeners had a miss rate that was larger than their false-alarm rate, and listeners' miss rates were often substantially larger than their false-alarm rates (see Figs. 7 and 11). This was the case for both North-American-English listeners and for Australian-English listeners, and was the case irrespective of whether responses of "1" were counted as errors or whether they were ignored.[23]

In terms of miss rates and false-alarm rates, for North-American-English listeners, there was no obvious different in bias between Experiment Condition 1 and Experiment Conditions 2. We conclude that the information we provided about recording conditions in Experiment Condition 2 did not result in a reduction in the bias in favour of the different-speaker hypothesis.

In terms of miss rates and false-alarm rates, for both North-American-English listeners and Australian-English listeners, the bias in favour of the different-speaker hypothesis was actually greater for Experiment Condition 3 than for Experiment Condition 1: When listening to high-quality recordings, more listeners had a miss rate that was larger than their false-alarm rate, and listeners' miss rates were often substantially larger than their false-alarm rates. Even though both miss rates and false-alarm rates tended to be lower for Experiment Condition 3 compared to Experiment Condition 1, the reductions in false-alarm rates tended to be greater than the reductions in miss rates, hence the miss rates were still relatively high compared to the false-alarm rates. We conclude that providing high-quality recordings in Experiment Condition 3 instead of the low-quality recordings provided in Experiment Condition 1 did not result in a reduction in the bias in favour of the different-speaker hypothesis.

In a series of studies by Lavan and colleagues [12–15], individual listeners were asked to cluster recordings by speaker. The recordings were short utterances (1–4 s long), included natural variation in linguistic content and speaking style, and were high quality (they had as little background noise as possible). Each listener was exposed to 30 recordings. Listeners who were unfamiliar with the speakers tended to create larger numbers of clusters than the true number of speakers, and the clusters tended to be similar in size to each other, i.e., the listeners appeared to perceive the number of speakers present in the set of recordings as being larger then the actual number of speakers present. The actual number of speakers that each listener heard was 2, but, across listeners, the median number of clusters created was 4–14, depending on the particular pair of speakers. These studies present additional examples of bias in favour of the different-speaker hypothesis for listeners' identification of unfamiliar speakers in high-quality recording conditions. Njie et al. [15] tested listeners who were more familiar and less familiar with the accent spoken by the speakers, but, for listeners who were unfamiliar with the speakers, that study did not find a difference in the numbers of clusters that listeners created.

5. Conclusion

Speaker-identification performance by individual listeners who heard high-quality versions of recordings was better than speaker-identification performance by individual listeners who heard poor-quality versions of the recordings with a mismatch between the questioned-speaker condition and the known-speaker condition. Listeners who heard high-quality versions of the recordings, however, still performed worse than a forensic-voice-comparison system that was based on state-of-the-art automatic-speaker-recognition technology.

In Part I, listeners heard mismatched poor-quality versions of the recordings. A bias in favour of the different-speaker hypothesis was observed in those listeners' responses. The research questions that were the focus of the present paper were:

- Whether providing information about the recording conditions ahead of time would result in a reduction in observed bias, i.e., if the listeners are warned ahead of time that the recordings conditions will make the questioned-speaker and known-speaker recordings sound more different from one another, will this reduce the degree of bias in favour of the different-speaker hypothesis?
- Whether the observed bias is actually due to the recording conditions, i.e., if listeners are presented with high-quality versions of the questioned-speaker and known-speaker recordings, will this eliminate the average bias in favour of the different-speaker hypothesis?

The answer to both research questions is a definitive "no". Under all experiment conditions, there was a substantial bias in favour of the different-speaker hypothesis, and, if anything, the bias was greater for listeners who heard high-quality versions of recordings than for listeners who heard poor-quality versions. The bias in favour of the different-speaker hypothesis therefore appears not to be due to the poor and mismatched recording conditions.

In the literature (e.g., Edmond [16]), the primary concern regarding bias has been about the circumstances of a courtroom case and other evidence presented in the case potentially biasing judges and juries in favour of the same-speaker hypothesis. Bias in either direction, however, either in favour of the same-speaker hypothesis or in favour of the different-speaker hypothesis, should be of concern for legal decision making. In Part I and in Part II, no case context was provided to the listeners – all experiment conditions were decontextualized with respect to case circumstances and with respect to other evidence that could be presented in the context of a case. Under these decontextualized conditions, the bias was in the opposite direction to that which has been of primary concern in the literature, i.e., the bias was in favour of the different-speaker hypothesis. In the current research, we have not investigated the effect of case context on speaker identification, and so, on the basis of the current research, we cannot draw any conclusions with respect to whether or how case context might bias speaker identification by judges or juries.

CRediT authorship contribution statement

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

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

[23] Although they were still in the minority, more listeners had false-alarm rates larger than miss rates than had positive B_{llr} values. Although both of these are indicators of bias in favour of the same-speaker hypothesis, unlike B_{llr}, false-alarm rate versus miss rate does not take account of a pattern commonly observed in Part I §3.2.5 in which, even if same-speaker responses are likelihood-ratio values greater than 1, those values are too low – the listeners' likelihood-ratio responses to same-speaker pairs undervalue the strength of evidence relative to the calibrated likelihood-ratio values output by the forensic-voice-comparison system in response to same-speaker pairs.

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 2019–2024.

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.

References

[1] N. Basu, A.S. Bali, P. Weber, C. Rosas-Aguilar, G. Edmond, G. Martire, G.S. Morrison, Speaker identification in courtroom contexts – Part I: Individual listeners compared to forensic voice comparison based on automatic-speaker-recognition technology, Forensic Sci. Int. (2022) 111499, https://doi.org/10.1016/j.forsciint.2022.111499

[2] 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, Speech Commun. 85 (2016) 119–126, https://doi.org/10.1016/j.specom.2016.07.006

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

[4] G.S. Morrison, P. Rose, C. Zhang, Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice, Aust. J. Forensic Sci. 44 (2012) 155–167, https://doi.org/10.1080/00450618.2011.630412

[5] G.S. Morrison, C. Zhang, E. Enzinger, F. Ochoa, D. Bleach, M. Johnson, B.K. Folkes, S. De Souza, N. Cummins, D. Chow, A. Szczekulska (2015/2022), Forensic Database of Voice Recordings of 500+ Australian English Speakers (AusEng 500+). Available at: ⟨http://databases.forensic-voice-comparison.net/⟩.

[6] 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, H. Eldridge, S. Lewis, K. Lothridge, P. Reedy (Eds.), Encyclopedia of Forensic Sciences (third ed.), vol. 2, pp. 737–750. Elsevier, 2023. ⟨https://doi.org/10.1016/B978-0-12-823677-2.00130-6⟩. Preprint at ⟨http://forensic-voice-comparison.net/encyclopedia/⟩.

[7] 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

[8] 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://forensic-voice-comparison.net/e3fs3/⟩.

[9] 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

[10] 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

[11] G.S. Morrison, A plague on both your houses: the debate about how to deal with "inconclusive" conclusions when calculating error rates, Law Probab. Risk 21 (2022) 127–129, https://doi.org/10.1093/lpr/mgac015

[12] N. Lavan, L.F.K. Burston, L. Garrido, How many voices did you hear? Natural variability disrupts identity perception in unfamiliar listeners, Br. J. Psychol. 110 (2018) 576–593, https://doi.org/10.1111/bjop.12348

[13] N. Lavan, S.E. Merriman, P. Ladwa, L. Burston, S. Knight, C. McGettigan, "Please sort these sounds into 2 identities": effects of task instructions on performance in voice sorting studies, Br. J. Psychol. 111 (2020) 556–569, https://doi.org/10.1111/bjop.12416

[14] J. Johnson, C. McGettigan, N. Lavan, Comparing unfamiliar voice and face identity perception using identity sorting tasks, Q. J. Exp. Psychol. 73 (2020) 1537–1545, https://doi.org/10.1177/1747021820938659

[15] S. Njie, N. Lavan, C. McGettigan, Talker and accent familiarity yield advantages for voice identity perception: a voice sorting study, Mem. Cogn. 51 (2022) 175–187, https://doi.org/10.3758/s13421-022-01296-0

[16] G. Edmond, Against jury comparisons, Aust. Law J. 96 (2022) 315–346.