Forensic Science International 360 (2024) 112048

Contents lists available at ScienceDirect

Forensic Science International

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

Speaker identification in courtroom contexts – Part III: Groups of collaborating listeners compared to forensic voice comparison based on automatic-speaker-recognition technology

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

[a] School of Psychology, University of New South Wales, Sydney, New South Wales, Australia
[b] Forensic Data Science Laboratory, Aston University, Birmingham, UK
[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

Keywords:
Admissibility
Forensic voice comparison
Likelihood ratio
Speaker identification
Validation

ABSTRACT

Expert testimony is only admissible in common-law systems if it will potentially assist the trier of fact. In order for a forensic-voice-comparison expert's testimony to assist a trier of fact, the expert's forensic voice comparison should be more accurate than the trier of fact's speaker identification. "Speaker identification in courtroom contexts – Part I" addressed 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. The present paper addresses the question of whether speaker identification by a group of collaborating lay listeners (such as a jury) would be more or less accurate than the output of such a forensic-voice-comparison system. As members of collaborating groups, participants listen to pairs of recordings reflecting the conditions of the questioned- and known-speaker recordings in an actual case, confer, and make a probabilistic consensus judgement on each pair of recordings. The present paper also compares group-consensus responses with "wisdom of the crowd" which uses the average of the responses from multiple independent individual listeners.

* Corresponding author at: Forensic Data Science Laboratory, Aston University, Birmingham, UK.
E-mail address: geoff-morrison@forensic-evaluation.net (G.S. Morrison).
[1] ORCID: 0000-0002-0166–0989
[2] Now at Forensic Science Research Group, Department of Applied Sciences, Northumbria University, Newcastle upon Tyne, UK
[3] ORCID: 0000-0003-2234-2995
[4] ORCID: 0000-0002-3121-9625
[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.2024.112048
Received 18 July 2023; Received in revised form 15 January 2024; Accepted 1 May 2024
Available online 6 May 2024

1. Introduction

The present paper addresses the question of whether speaker identification[9] by a group of collaborating lay listeners (such as a jury) 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.[10] It also compares group-consensus responses with "wisdom of the crowd" which uses the average of the responses from multiple independent individual listeners.

The present paper is Part III of 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. [2]), the performance of speaker identification by individual lay listeners was 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) 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 most common-law systems if it has the potential to assist the trier of fact. 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 assist the trier of fact. Part I §1.2 discussed the legal context related to forensic voice comparison conducted by experts and speaker identification performed by triers of fact.

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.[11] Under these conditions, in terms of C_{llr} (see §2.7.2 below) and in terms of classification-error rate, all of the individual listeners in Part I performed worse than the forensic-voice-comparison system. Part II (Basu et al. [6]) explored a bias in favour of the different-speaker hypothesis that was apparent in the listeners' responses in Part I. Part II included testing listeners on high-quality non-mismatched versions of the recording pairs. Part II concluded that the bias was not due to the poor and mismatched recording conditions of the Part I stimuli.

In the present paper (Part III), we conduct experiments in which groups of lay listeners are asked to collaboratively make probabilistic same-speaker/different-speaker judgements on pairs of recordings. The stimuli used in Part III consist of 6 same-speaker pairs of recordings and 6 different-speaker pairs of recordings, which are a subset of the stimuli used in Part I. The language and accent spoken on the recordings is Australian English. The listeners in Part III are all Australian-English listeners. We compare the groups of listeners' consensus 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 [7],[8],[9], and is the same system as was used in Parts I and II.

Part I §1.3 reviewed prior research on the performance of speaker identification by lay listeners compared to the performance of automatic-speaker-recognition systems. We do not repeat that review here, but note that we were unable to find any reported research on speaker identification by groups of listeners in which the members of each group collaborated to reach a group-consensus response.[12]

A number of publications have reported on research in which multiple individual listeners each listen independently, then a function (a simple function such as mean or mode, or a more complex function such as a calibration model) is applied to the pooled responses from all the listeners, and then the output of that function is compared with the output of an automatic-speaker-recognition system. In the "wisdom of the crowd" effect, a function applied to multiple individuals' independent responses leads to a result that is close to the true answer, e.g., in the classic example, attendees at a fair were asked to guess the weight of an ox, and the median response was within 1% of the measured weight (Galton [12]). In a forensically relevant example reported in Tangen et al. [13], when fingerprint experts were given unlimited time to perform fingermark-fingerprint comparisons, the mean miss rate and mean false-alarm rate for individuals' responses were 15% and 2.8%

[9] 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.

[10] The same question could be asked with respect to other approaches to forensic voice comparison, but this is outside the scope of the present paper. For a recent summary of approaches to forensic voice comparison, see Morrison & Zhang [1].

[11] For further information about the stimuli used in Part I, see Part I §2.2. For further information about the source database, see Morrison & Enzinger [3]. For information about the simulation of the recording conditions, see Enzinger et al. [4]. For information about the data-collection protocols, see Morrison et al. [5].

[12] A research report that we inadvertently omitted from the review presented in Part I is Park et al. [10]. In that study, stimuli were high-quality read sentences less than 2 s long. Listeners gave confidence scores of 1–5 for same-speaker or different-speaker. Those scores were "unfolded" to the range 1 (positive different speaker) to 10 (positive same speaker), averaged across listeners, and calibrated using logistic regression. C_{llr} for this function fitted to pooled listener responses was 0.425, compared to 0.737 for the likelihood ratios output by an i-vector-PLDA automatic-speaker-recognition system. This study, in common with many other studies, fitted a function to pooled responses from listeners who gave independent responses to the stimuli. This does not address the questions of interest in our research with respect to the performance of individual listeners (Part I) or groups of collaborating listeners (Part III) in a context in which calibration of their responses is not possible.

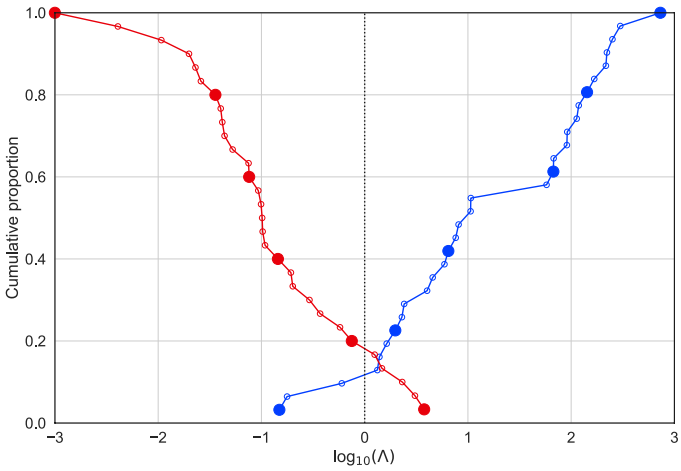

Fig. 1. Tippett plot of validation results for the forensic-voice-comparison system using the 31 same-speaker pairs of recordings and 30 different-speaker pairs of recordings from Part I. The filled circles correspond to the pairs of recordings that were presented to the groups of listeners in the present research.

respectively, but, using the modal response from sets of 3 individuals' responses, the mean miss rate and mean false-alarm rate fell to 4% and 0% respectively. For novices, however, the mean miss rate only decreased from 24% to 21% and the mean false-alarm rate increased (rather then decreased) from 60% to 71%.

Fitting a function to multiple independent listeners' speaker-identification responses may be expected to lead to a more accurate result. This does not, however, reflect the situation in which a group of people constituting a jury listen, confer, and collaboratively come to a consensus (Karpowitz & Mendelberg [11]). The lack of independence in this process is not expected to lead to as large a "wisdom of the crowd" effect: In a numerical estimation task, Lorenz et al. [14] found that exposing participants to other participants' estimates led to less accurate average estimates, and, despite a lack of improvement in accuracy, to participants having greater confidence in their estimates. In order to investigate the performance of collaborating groups versus "wisdom of the crowd" based on individual independent listeners, we compare the group-consensus responses with the geometric mean of independent-individual-listener responses. For each pair of recordings, before members of a group confer, each member gives an individual response.

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

Pairs of recordings used as stimuli in the present research consisted of a subset of the 31 same-speaker pairs of recordings and 30 different-speaker pairs of recordings (61 total pairs of recordings) used in Part I. In each pair of recordings, 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 pairs of recordings used in Part I were a subset of those in the *forensic_eval_01* validation dataset (Morrison & Enzinger [3]), 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.

A priori, we expected that groups of listeners acting collaboratively would only be able to respond to a small proportion of the number of stimulus pairs to which individual listeners had responded. Based on pilot work, we selected a subset of 12 recording pairs, 6 same-speaker pairs and 6 different-speaker pairs. The recording pairs were uniformly sampled based on the range of likelihood-ratio values that the forensic-voice-comparison system output in response to the 31 same-speaker pairs of recordings and in response to the 30 different-speaker pairs of recordings. The filled circles in Fig. 1 correspond to the pairs of recordings that were presented to the groups of listeners in the present research.

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

2.3. Participants (listeners)

Participants were recruited from online recruitment platforms operated by the School of Psychology at the University of New South Wales.[13] In exchange for their participation, first-year-undergraduate-student participants received participation credit and other participants received AUD 40.

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

1. were 18 years of age or older
2. were an Australian citizen and currently a resident of Australia
3. did not have a diagnosed hearing loss
4. were technically able and willing to use audio and video in a Zoom meeting[14]

Criteria 1 and 2 combined were a substitute for participants being jury eligible in Australia.

Participants were assigned to groups. To reflect the size of juries in most common-law jurisdictions, the target number of participants in each group was 12. The target number of groups was 30.

2.4. Experiment procedures

Bespoke software was used to run the group-of-listeners experiment. 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 was split into two parts. In part A, which lasted up to 15 minutes, participants individually logged on to the experiment software at a time of their choosing. In part A, participants were presented with participant information, and had the opportunity to give informed consent to their participation in the research. If a participant gave informed consent, they were presented with written instructions explaining the task, then a sound check to make sure they could hear audio playing on their device, then three practice trials that they completed individually (participants were told that these were practice trials), then they were provided with additional instructions about the group activity that would occur in a part B of the experiment. If they wished, participants could repeat the instructions and practice trials.

[13] https://unsw-psy.sona-systems.com/ and https://unsw-psy-paid.sona-systems.com/

[14] Potential participants who required accommodation, e.g., because of vision impairment, were asked to contact the researchers.

Participants could also access all of the instructions whenever they wanted during the practice trials and during part B of the experiment. The practice trials used two different-speaker pairs of recordings and one same-speaker pair. These recordings were different from any of the recordings used in the experiment trials. No feedback was given to participants with respect to their responses to practice trials.

After a participant had completed part A of the experiment, they were assigned to a group and sent an invitation to a scheduled Zoom[15] meeting during which they would participate in part B of the experiment. Part B of the experiment lasted up to 1 hour 45 minutes. At the appointed time, all members of a group joined the Zoom meeting and logged back in to the experiment software. In the participant information, participants had been told that they would be working in the Zoom meeting with other participants who would be able to see and hear them. For anonymization, before being admitted to the Zoom meeting, each participant's screen name was replaced by a number. Participants were instructed to refer to each other by that number, even if they happened to know other participants, and to keep the identity of other participants confidential.[16] In addition to giving informed consent, participants had to indicate that they agreed to follow these instructions.

In the Zoom meeting, the researcher running the experiment recapped the instructions and answered any questions. The group of participants were then asked to choose a foreperson. They were not given any specific instructions as to how to do this. Among the participants, only the foreperson could control the playback of the audio recordings in the experiment software (the audio could be heard by all the members of the group), and only the foreperson could enter the group response and advance to the next trial.

The first trial was a practice trial. It reused a pair of recordings that had been used in the practice trials in part A of the experiment. Participants were told that this was a practice trial, and the responses to this trial were not included in the analysis of the results. During the practice trial, the researcher walked the participants though the process. She provided guidance on what the participants had to do (but not how to do it), and helped with any technical issues.

The remainder of the trials were experiment trials.[17] Each group was presented with the experiment trials in a different random order. For the duration of the experiment trials, the researcher switched off their camera and microphone, but the participants were told that she would be observing and that the participants could ask for help if they ran into technical problems. When the group of participants had completed all the trials, or they had reached the end of the allotted time, the researcher thanked them for participating, and subsequently arranged for each participant to receive either participation-credit or payment.

Each experiment trial consisted of two stages:

- an individual-response stage, and
- a group-response stage.

Fig. 2 shows screenshots of the foreperson's view and the regular-participants' view of the individual-response stage and the group-response stage.

Each experiment trial began with an individual-response stage. During that stage, the screen included the written instruction: "Enter your individual response without conferring with the other members of your group." The foreperson had access to two sets of audio-playback controls, one labelled "questioned-speaker recording" and the other labelled "known-speaker recording". Using each set of controls, the foreperson 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 foreperson played the audio recordings. The audio was routed to the other participants' devices via Zoom.[18] Participants were asked to mute their microphones while listening. Participants could ask the foreperson to replay all or part of the recordings. Without consulting with the other participants, each participant entered their individual answer into one of two response boxes. The response boxes were identical to those in the individual-listener experiments in Parts I and II.

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.[19]

The software checked that the participant 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 a holding page which included the text: "Individual response received. Please wait

[15] https://zoom.us/

[16] We did not control whether any of the participants in a group knew each other. We do not consider this an important factor. Members of a jury would initially not know each other but would get to know each other during the course of a trial.

[17] Unlike in the individual-listener experiments in Parts I and II, the group-of-listener experiment did not include attention-checking trials.

[18] Zoom uses lossy codecs for audio transmission, so this may result in listeners hearing poorer-quality playback compared to in the individual-listener experiment in Parts I and II. The latter used uncompressed pulse-code modulation (PCM) audio files. We tried routing the uncompressed audio through our bespoke software, but could not achieve a solution that would be robust for all browsers and devices. It was also impractical to control other factors affecting sound quality, such as whether listeners used headphones. We do not know what the listening conditions would be for particular juries, but expect that this could vary substantially from courthouse to courthouse.

[19] 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 that was a subjectively assigned quantification of the listener's or group of listeners' assessment of the strength of the evidence.

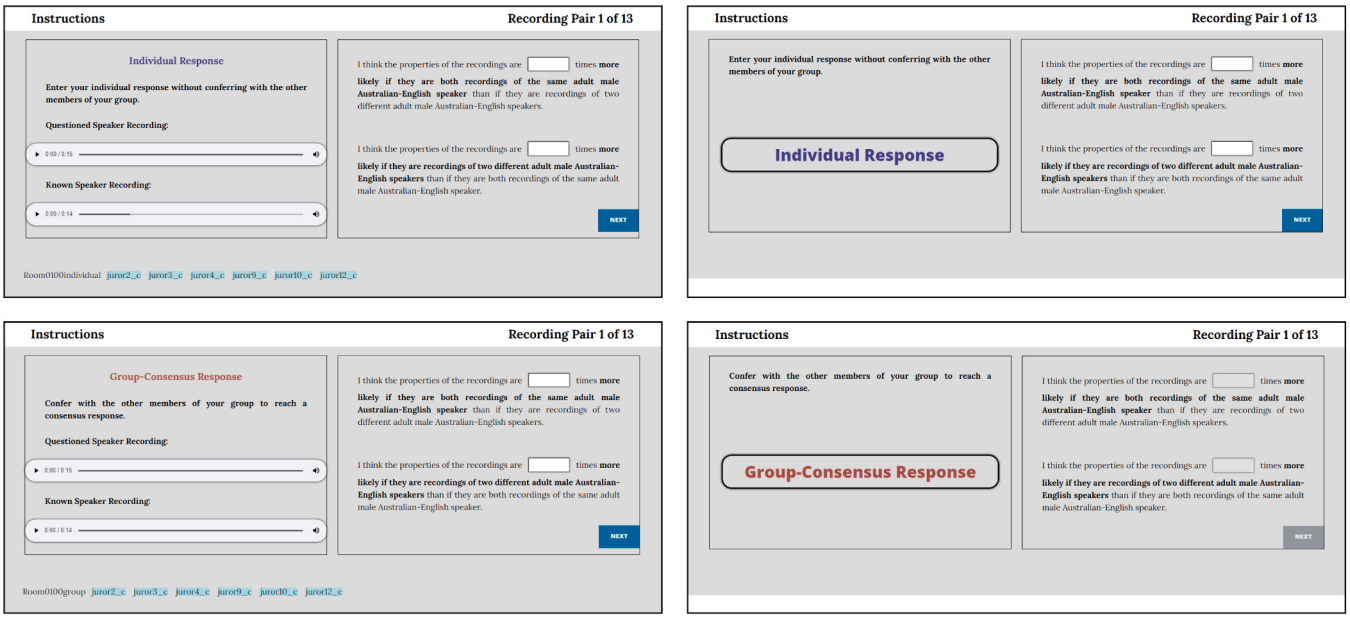

Fig. 2. Screenshots of an experimental trial in the group-of-listeners experiment. Top left: Foreperson's view of the individual-response stage. Bottom left: Foreperson's view of the group-response stage. Top right: Regular-participants' view of the individual-response stage. Bottom right: Regular-participants' view of the group-response stage.

to begin the group-response phase." 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 holding page, they could not return to the individual-response page. If the foreperson was still on the individual-response page, they could see how many participants were still on that page, and if foreperson moved to the holding page, they could see how many participants had reached that page. Via Zoom, the foreperson could ask whether anyone needed more time to provide an individual-response. Once all participants had given an individual response (or the foreperson decided that any who had not responded were not going to), the foreperson could press a "next" button that moved all participants to the group-response stage of the trial.

During the group-response stage, the screen included the written instruction: "Confer with the other members of your group to reach a consensus response." During group discussion, participants could ask the foreperson to replay the recordings as many times as they wanted. They were not given instructions as to how to come to a consensus, but they were told that it had to be a response that everyone was willing to accept. They were told that they could not enter a response that was simply a mathematically calculated average of the individual responses, and they could not choose a response by simple majority vote. Once the group had agreed on a response, the foreperson entered that response, pressed "next", and the experiment software moved to the next trial.

When the group of participants had completed all the trials, or they had reached the end of the allotted time, the researcher thanked them for participating, and subsequently arranged for each participant to receive course-credit or payment.[20]

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. For more detailed descriptions of this system, see Morrison et al. [9] and Weber et al. [7],[15].[21]

The system, how it was trained and calibrated for the poor-quality mismatched recording conditions, and how validation was conducted, were described in Part I §2.5. As calibration data, all recordings in the *forensic_eval_01* validation set were used.[22] Three 15 s long non-overlapping sections were randomly selected from each of the latter recordings. Leave-one-speaker-out / leave-two-speakers-out cross validation was then employed.[23] The validation data for the present research consisted of the same ~15 s-long recordings as had been used with the groups of human listeners.

2.6. "Wisdom of the crowd"

For "wisdom of the crowd", we used a geometric mean. We took the participants' individual responses to a stimulus pair (the responses they

[20] We examined within-group individual-listener responses arranged in the order in which the stimulus pairs were presented to the group. We did not observe any obvious learning patterns such as within-group individual-listener responses converging with one another.

[21] More information about E^3FS^3 is available from http://forensic-voice-comparison.net/E3FS3/

[22] We applied regularized logistic regression with a regularization weight, κ, equivalent to 1 pseudo-speaker relative to the number of speakers used for training the logistic-regression model (see Morrison & Poh [16]).

[23] 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).

gave before conferring with the other members of their group), converted each participant's individual response to a likelihood ratio (see §2.7.1 below), then to a log likelihood ratio, calculated the mean of the log likelihood ratios for the group, converted the mean log likelihood ratio to a linear scale, i.e., to a likelihood ratio, and used that likelihood ratio as the group response.

2.7. Metrics for analysis of response data

2.7.1. Introduction

For each individual listener's response and for each consensus response by a group of listeners: 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:

- C_{llr} (log-likelihood-ratio cost) (§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} (a difference metric) (§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} (a bias metric) (§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's responses and for the group-consensus responses of each group of listeners. 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).

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

2.7.2. C_{llr}

For the forensic-voice-comparison system's responses, for each listener's independent responses, for each group of listeners' consensus responses, and for each group of listeners' "wisdom of the crowd" responses, a C_{llr} value was calculated [17]. 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.

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

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* [18]. 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 a group of listeners' consensus responses (or "wisdom of the crowd" 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 a group of listeners did not respond to all pairs of recordings, only the responses which they did provide were used for calculating D_{llr}. 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.

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

2.7.4. B_{llr}

In order to compare a group of listeners' consensus responses (or "wisdom of the crowd" 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). If a group of listeners did not respond to all pairs of recordings, only the responses which they did provide were used for calculating B_{llr}. 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.

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

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 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 [19]). 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. For example, if, in response to the 6 same-speaker pairs of recordings, a listener gave 2 responses of "1", 3 responses in the top box greater than 1 (corresponding to likelihood ratios greater than 1), and 1 response in the bottom box greater than 1 (corresponding to likelihood ratios less than 1), the miss rate would be calculated as $(2+1)/6 = 50\%$.
- 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 6 same-speaker pairs of recordings, a listener gave 2 responses of "1", 3 responses in the top box greater than 1 (corresponding to likelihood ratios greater than 1), and 1 response in the bottom box greater than 1 (corresponding to likelihood ratios less than 1), the miss rate would be calculated as $1/(3+1) = 25\%$.

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 and discussion

3.1. Performance of the forensic-voice-comparison system

In Part I, which used 31 same-speaker and 30 different-speaker pairs of recordings, the C_{llr} value for E^3FS^3 was 0.42. Given the 6 same-speaker and 6 different-speaker pairs of recordings in the present research, the C_{llr} value for E^3FS^3 was 0.60.

The same forensic-voice-comparison system was used in Part I and in Part III, and the underlying performance of the system is therefore unchanged. The difference in measured performance is due to the difference in validation data (test data). The validation data for Part III were not a random sample, but a uniform subsample based on the ranking of the likelihood-ratio values that the system had provided in response to the 31 same-speaker pairs and a uniform subsample based on the ranking of the likelihood-ratio values that the system had provided in response to the 30 different-speaker pairs (see §2.2 above). A random sample of log-likelihood-ratio values resulting from same-speaker comparisons would result in a relatively large number of values close to the mode of the same-speaker log-likelihood-ratio values, and relatively few values far from the mode, including few log-likelihood-ratio values that are small and positive and even fewer that are negative. *Mutatis mutandis*, a random sample of log-likelihood-ratio values resulting from different-speaker comparisons would result in a relatively large number of values close to the mode of the different-speaker log-likelihood-ratio values, and relatively few values far from the mode, including few log-likelihood-ratio values that are small and negative and even fewer that are positive. In contrast, the uniform sampling based on the already-known output of the system resulted in a relatively larger number of same-speaker log-likelihood-ratio values that are negative or small and positive, and a relatively larger number of different-speaker log-likelihood-ratio values that are positive or small and negative. It is the low same-speaker log-likelihood-ratio values and the high different-speaker log-likelihood-ratio values that contribute greatest to the magnitude of a calculated C_{llr}, see Equation (1), and their over-representation in the Part III validation sample leads to a high C_{llr} value which is misrepresentative of the forensic-voice-comparison system's underlying performance. Repeatedly randomly sampling 6 values from the 31 same-speaker log-likelihood-ratio values and 6 values from the 30 different-speaker log-likelihood-ratio values resulted in a modal C_{llr} value of 0.41 (the distribution was positively skewed).

Note that the stimulus-pair selection procedure was biased against the forensic-voice-comparison system because it overrepresented pairs of recordings on which it was already known the system performed poorly, whereas it was not known how well human listeners would perform on these pairs. We will proceed, however, to compare the C_{llr} values for the human listeners' responses with the C_{llr} value of 0.60 obtained for the forensic-voice-comparison system's responses to the same stimuli.

A Tippett plot of the validation results from E^3FS^3 is provided in Fig. 3. This is identical to Fig. 1 above except that it only shows results for the same recording pairs to which the groups of listeners responded in the present research. For an explanation of how to interpret Tippett plots, see Appendix C.1 of the *Consensus on validation of forensic voice comparison* [18] and the references cited therein.

3.2. Groups of listeners – demographics

Reflecting the size of juries in most common-law jurisdictions, the target number of participants per group was 12. The target number of groups was 30. Recruiting participants who went on to attend the collaborative group activity to which they were invited was, however, very difficult. Only one group actually ended up having 12 attendees. We excluded from analysis data from groups with fewer than 5 members. We also excluded data from groups that did not provide responses to all 12 stimulus pairs. After these exclusions, we had 23 groups. Fig. 4, Fig. 5, and Fig. 6 show, respectively, the number of participants in each group split by self-reported gender, the number of participants in each group split by first language, and the ages of the participants in each group. Because a substantial proportion of participants were recruited from undergraduate university students, the age distribution was skewed toward younger ages than might be expected for juries, but we

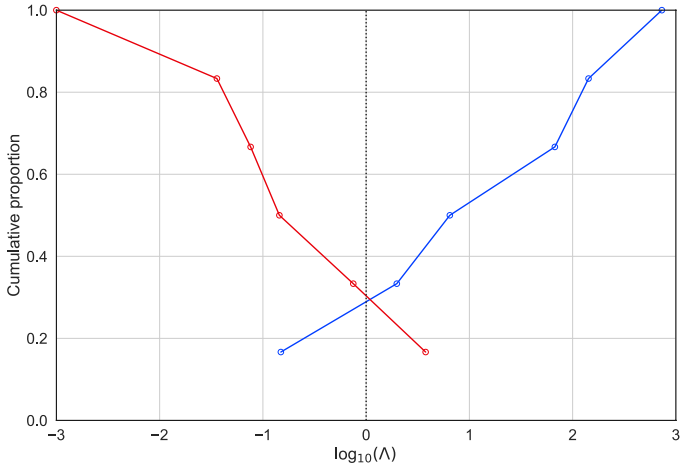

Fig. 3. Tippett plot of validation results for the forensic-voice-comparison system using 6 same-speaker pairs of recordings and 6 different-speaker pairs of recordings.

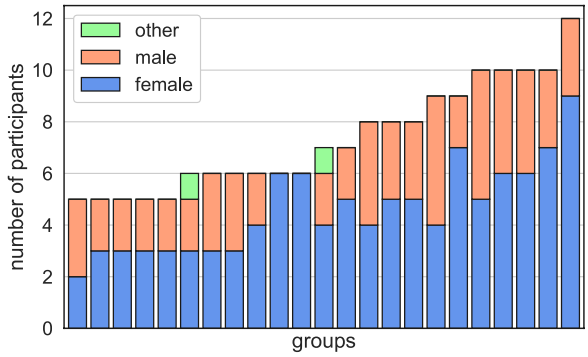

Fig. 4. Number of participants in each group of listeners, split by self-reported gender of participants.

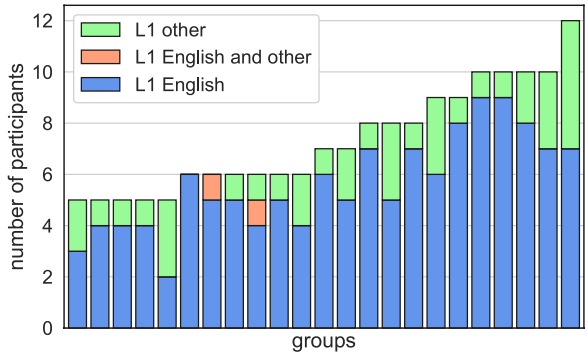

Fig. 5. Number of participants in each group of listeners, split by first language of participants.

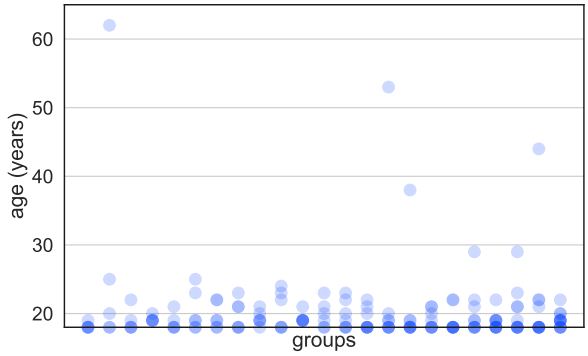

Fig. 6. Ages of participants in each group of listeners. Darker symbols represent multiple participants with the same age.

do not expect this to have meaningfully impacted the results.

Having different numbers of participants in different groups, with some groups having substantially fewer than the target number of 12 participants, is not ideal. To explore whether there was a relationship between the number of participants in a group and the C_{llr} value calculated for the group's consensus responses, we plotted C_{llr} values against group sizes (see Fig. 7). There was no apparent relationship between C_{llr} value versus group size, nor was there any apparent relationship between D_{llr} value or B_{llr} value versus group size (plots not shown in the present paper). We therefore proceed with presentation and discussion of results pooled across groups of different sizes. We have no reason to believe that the pattern of results would have been substantially different if all the groups had had 12 participants.

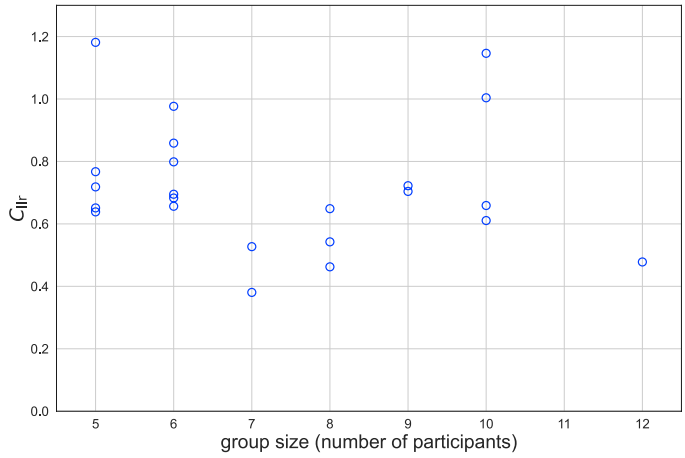

Fig. 7. Plot of C_{llr} values for group-consensus responses versus group sizes.

3.3. Performance of groups of listeners

3.3.1. C_{llr} values

A C_{llr} value was calculated separately, for each listener's independent responses, for each group of listeners' consensus responses, and for each group of listeners' "wisdom of the crowd" responses. Fig. 8 shows violin plots of the resulting C_{llr} values. The heavy black horizontal line indicates the C_{llr} value for the forensic-voice-comparison system.

For the group-consensus results, 5 of the 23 groups (22% of the groups) had lower C_{llr} values than the C_{llr} value for the forensic-voice-comparison system, the remaining 18 groups (78% of the groups) had C_{llr} values greater than the C_{llr} value for the forensic-voice-comparison system. We therefore conclude that a state-of-the-art forensic-voice-comparison system would outperform most groups of collaborating lay listeners.

As expected, the variability of C_{llr} values that resulted from group-consensus responses was less than that for individual-listener responses. The group-consensus responses also resulted in substantially lower C_{llr} values than for the individual-listener responses. We conclude that collaborating groups of listeners outperform individual listeners.

The "wisdom of the crowd" responses also resulted in substantially lower C_{llr} values than for the individual-listener responses, but, unexpectedly, the "wisdom of the crowd" responses resulted in higher C_{llr} values than the group-consensus responses. The within-group pairwise difference in C_{llr} values (C_{llr} for "wisdom of the crowd" minus C_{llr} for

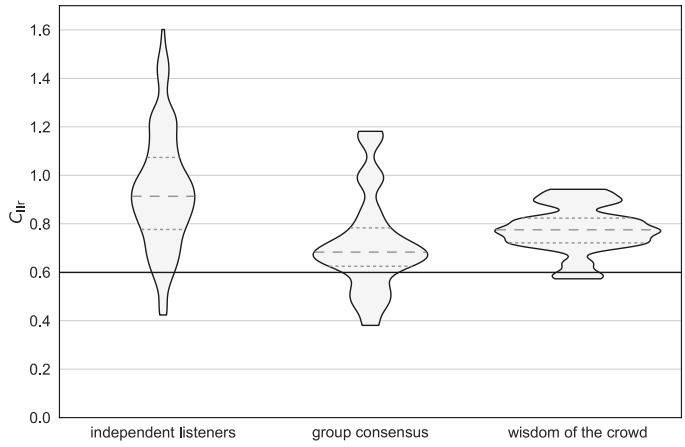

Fig. 8. Violin plots of the C_{llr} values for individual listeners' independent responses, for the group-consensus responses, and for the "wisdom of the crowd" responses. The heavy black horizontal line indicates the C_{llr} value for the forensic-voice-comparison system.

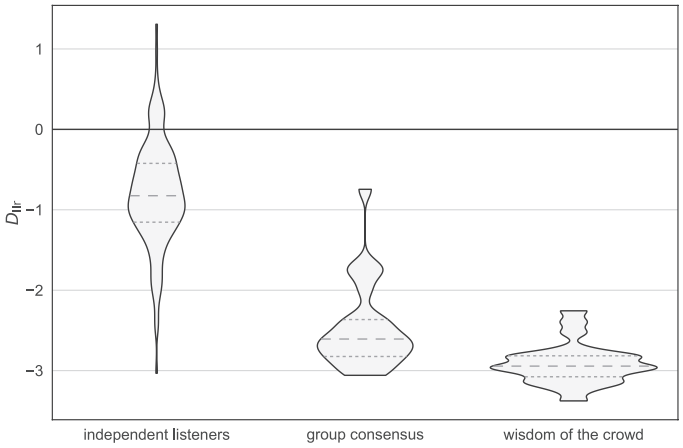

Fig. 9. Violin plots of the D_{llr} values for individual listeners' independent responses, for the group-consensus responses, and for the "wisdom of the crowd" responses. D_{llr} values were calculated relative to the forensic-voice-comparison system.

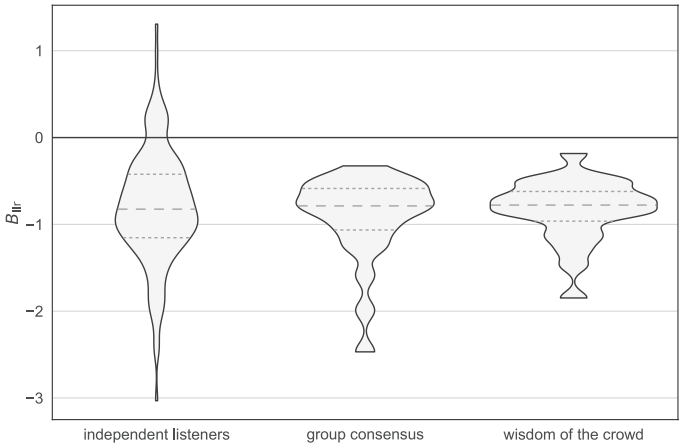

Fig. 10. Violin plots of the B_{llr} values for individual listeners' independent responses, for the group-consensus responses, and for the "wisdom of the crowd" responses. B_{llr} values were calculated relative to the forensic-voice-comparison system.

group-consensus responses) are shown in Fig. 11. We conclude that, for the same sets of listeners, collaborating groups of listeners outperform the "wisdom of the crowd". In making this conclusion, however, we note the limitation that the number of individuals in each group was small. It could be that a larger "wisdom of the crowd" effect would occur for larger numbers of independent listeners.[24]

3.3.2. D_{llr} values

A D_{llr} value was calculated separately, for each listener's independent responses, for each group of listeners' consensus responses, and for each group of listeners' "wisdom of the crowd" responses. D_{llr} values were calculated relative to the forensic-voice-comparison system. Fig. 9 shows violin plots of the resulting D_{llr} values.

For most individual-listener responses, and for all the group-consensus responses and all the "wisdom of the crowd" responses, 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 of the group-consensus responses was about a sixth that of the forensic-voice-comparison system.

The D_{llr} values for the group-consensus responses were much more negative than those of individual listeners (in addition to Fig. 9, see Fig. 11), i.e., on average the different-source likelihood ratios and same-source likelihood ratios are closer to one another for the groups of collaborating listeners than for the individual listeners. It appears that groups of collaborating listeners are more conservative than individual listeners.[25]

The D_{llr} values for the "wisdom of the crowd" responses were much more negative than those of individual listeners and more negative than those for the group-consensus responses. If between-listener variability were such that some listeners gave high likelihood-ratio values and others low likelihood-ratio values to the same stimuli, then the average values for these stimuli would be moderate (close to a likelihood-ratio value of 1), thus accounting for the large negative D_{llr} values.

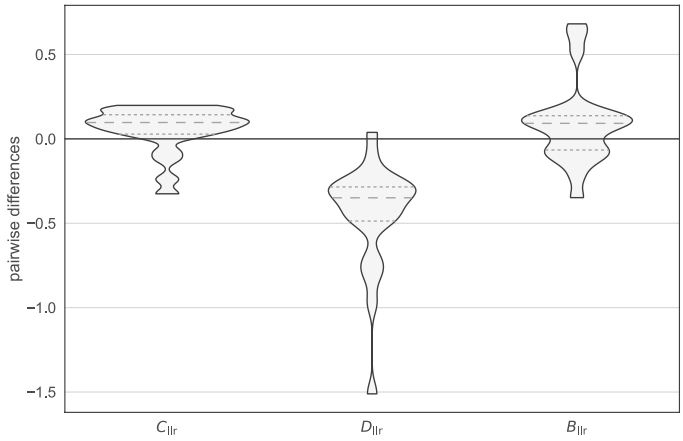

Fig. 11. Violin plots of pairwise differences (metric for "wisdom of the crowd" responses minus metric for group-consensus responses) for C_{llr}, D_{llr}, and B_{llr}.

3.3.3. B_{llr} values

A B_{llr} value was calculated separately, for each listener's independent responses, for each group of listeners' consensus responses, and for each group of listeners' "wisdom of the crowd" responses. B_{llr} values were calculated relative to the forensic-voice-comparison system. Fig. 10 shows violin plots of the resulting B_{llr} values.

For most individual-listener responses, and for all the group-consensus responses and all the "wisdom of the crowd" responses, B_{llr} values were negative, i.e., relative to the forensic-voice-comparison system, they were biased in favour of the different-speaker hypothesis. For the group-consensus responses, the median bias was such that, on average, likelihood-ratio values were about 40% lower than of those of the forensic-voice-comparison system. The median bias was similar for

[24] In Feichter & Kornell [20], on a task to estimate the number of stars that appeared on a screen, substantial improvements in the accuracy of estimates were observed as the number of participants increased from 1 to 5, and, albeit at a slower rates, accuracy continued to improve as the number of participants increased from 5 to 10. In White et al. [21], on a categorical unfamiliar-face matching task, modal response with 8 or more participants outperformed the best individual, but there was only slight improvement for modal response beyond 10 participants.

[25] C_{llr}, D_{llr}, and B_{llr} measure different things. C_{llr} is an absolute measure of accuracy. D_{llr} and B_{llr} are measures of difference and bias relative to the forensic-voice-comparison system. It is possible, as in this case, for one set of results to have a lower D_{llr} value (indicating that they are more conservative, i.e., on average their different-source likelihood ratios and same-source likelihood ratios are closer to one another) than another set, but for the first set to have a lower C_{llr} value (indicating better accuracy).

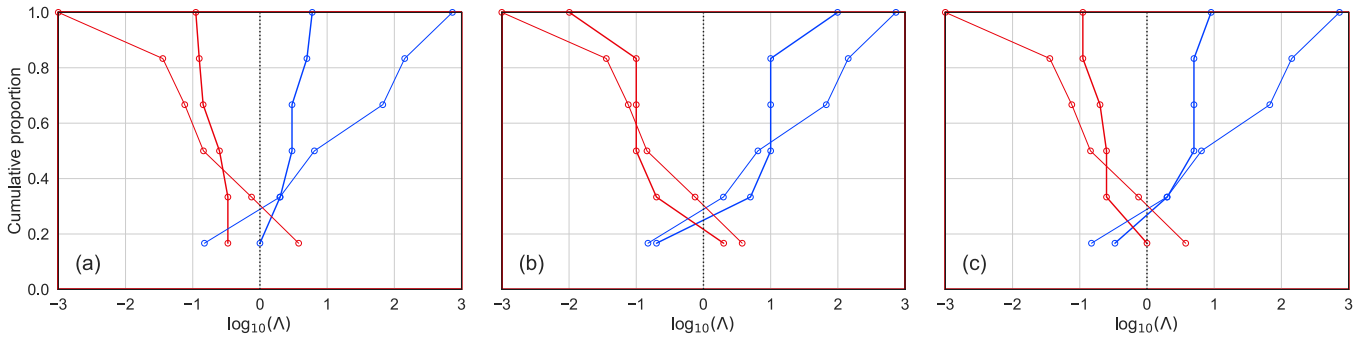

Fig. 12. Tippett plots of the group-consensus responses from the three best-performing groups of listeners, i.e., those with the lowest C_{llr} values. Drawn using lighter weight lines and symbols, each panel also includes the Tippett plot for the forensic-voice-comparison system. (a) $C_{llr} = 0.38$, $D_{llr} = -1.69$, $B_{llr} = -0.75$ (7 participants) (b) $C_{llr} = 0.46$, $D_{llr} = -0.75$, $B_{llr} = -0.43$ (8 participants) (c) $C_{llr} = 0.48$, $D_{llr} = -1.77$, $B_{llr} = -0.58$ (12 participants).

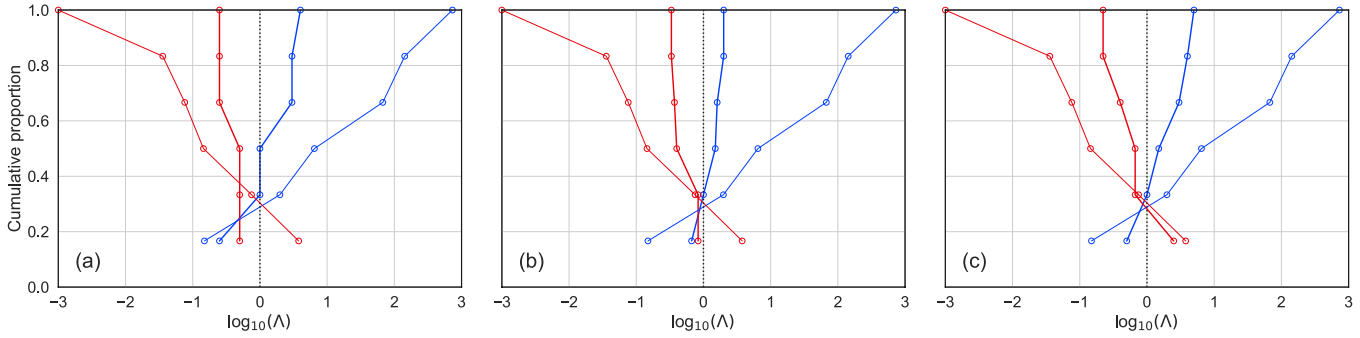

Fig. 13. Tippett plots of the group-consensus responses from groups with C_{llr} values close to the median C_{llr} value, 0.68. Drawn using lighter weight lines and symbols, each panel also includes the Tippett plot for the forensic-voice-comparison system. (a) $C_{llr} = 0.68$, $D_{llr} = -2.61$, $B_{llr} = -0.81$ (6 participants) (b) $C_{llr} = 0.70$, $D_{llr} = -2.86$, $B_{llr} = -0.64$ (9 participants) (c) $C_{llr} = 0.72$, $D_{llr} = -2.70$, $B_{llr} = -0.33$ (9 participants).

individual-listener responses, for the group-consensus responses, and for the "wisdom of the crowd" responses.

Part II explored whether such bias in favour of the different-speaker hypothesis was due to the poor and mismatched quality recording conditions, and concluded that it was not.

3.3.4. *Pairwise differences for C_{llr}, D_{llr}, and B_{llr}*

As a supplement to §3.3.1, §3.3.2, and §3.3.3, Fig. 11 shows violin plots of pairwise differences (metric for "wisdom of the crowd" minus metric for group-consensus responses) for C_{llr}, D_{llr}, and B_{llr}.

3.3.5. *Tippett plots*

Fig. 12, Fig. 13, and Fig. 14 show example Tippett plots of group-consensus responses for, respectively, groups with the best performance (lowest C_{llr}), average performance (close to the median C_{llr}), and the worst performance (highest C_{llr}). For comparison, each panel also

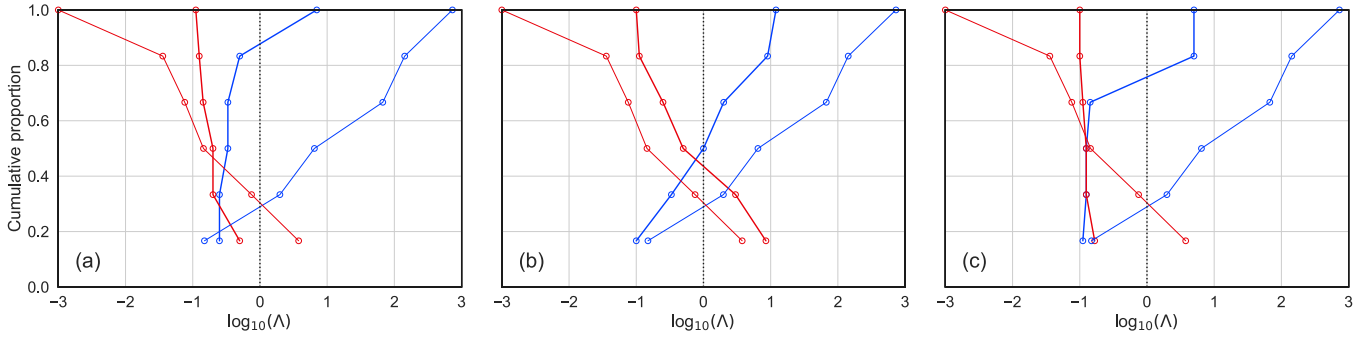

Fig. 14. Tippett plots of the group-consensus responses from the three worst-performing groups of listeners, i.e., those with the highest C_{llr} values. Drawn using lighter weight lines and symbols, each panel also includes the Tippett plot for the forensic-voice-comparison system. (a) $C_{llr} = 1.00$, $D_{llr} = -2.85$, $B_{llr} = -1.99$ (10 participants) (b) $C_{llr} = 1.15$, $D_{llr} = -2.98$, $B_{llr} = -0.49$ (10 participants) (c) $C_{llr} = 1.18$, $D_{llr} = -2.70$, $B_{llr} = -2.47$ (5 participants).

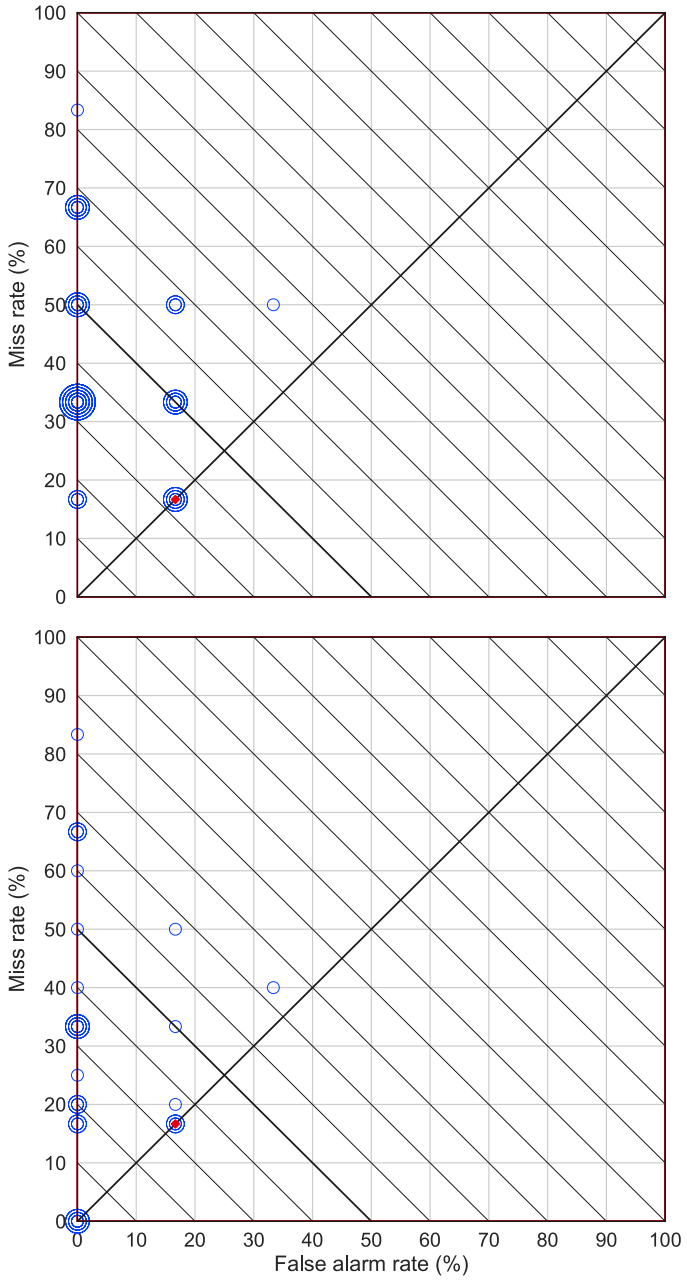

Fig. 15. Plots of miss rate versus false-alarm rate for the group-consensus responses. Top panel: Responses of "1" treated as errors. Bottom panel: Responses of "1" ignored. Concentric circles represent multiple groups with the same combination of miss rate and false-alarm rate. The filled diamonds represent the results for the forensic-voice-comparison system.

includes the Tippett plot for the forensic-voice-comparison system drawn using lighter weight lines and symbols.[26] Across the three figures, for the group-consensus response, many of the example Tippett plots have same-speaker likelihood-ratio responses and different-speaker likelihood-ratio responses that are substantially closer together than those of the forensic-voice-comparison system (and hence have large negative D_{llr} values). Fig. 14 includes two example Tippett plots in which, relative to the forensic-voice-comparison system, the group-consensus responses have large biases in favour of the different-speaker hypothesis (and hence have large negative B_{llr} values).[27]

3.3.6. *Miss rate and false-alarm rate*

Fig. 15 shows plots of miss rates versus false-alarm rates for group-

[26] D_{llr} and B_{llr} were calculated pairwise; however, the group-consensus Tippett plots and the forensic-voice-comparison Tippett plots are drawn independently of one another. In the Tippett plots, group-consensus responses and forensic-voice-comparison responses at the same cumulative proportion are, therefore, generally not responses to the same stimulus pairs.

[27] Note that in Fig. 12a, the group-consensus likelihood ratios are more conservative (on average the different-source likelihood ratios and same-source likelihood ratios are closer to one another) than for the forensic-voice-comparison system (D_{llr} is negative), and, compared to the forensic-voice-comparison system, the group-consensus likelihood ratios have a slight bias in favour of the different-speaker hypothesis (B_{llr} is negative), but the group-consensus likelihood ratios are more accurate than the forensic-voice-comparison system's likelihood ratios (C_{llr} is smaller).

consensus responses. The top panel shows results calculated using the procedure that counted responses of "1" as errors, and the bottom panel 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. 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 would 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. The circles show the group-consensus results from groups of listeners. Concentric circles represent multiple groups with the same combination of miss rate and false-alarm rate. The filled diamonds represent the results for the forensic-voice-comparison system.

Treating responses of "1" as errors, 3 groups of listeners had equal miss and false-alarm rates, but the other 20 groups were biased in favour of the different-speaker hypothesis. 8 groups of listeners had the same classification-error rate as the forensic-voice-comparison system, and 2 groups had a lower classification-error rate. The other 13 groups (57%) had higher error rates than the forensic-voice-comparison system.

Ignoring responses of "1", 5 groups of listeners had equal miss and false-alarm rates, but the other 18 groups were biased in favour of the different-speaker hypothesis. 5 groups of listeners had the same error rate as the forensic-voice-comparison system, and 8 groups had lower error rates (3 groups had no errors). The other 10 groups (43%) had higher error rates than the forensic-voice-comparison system.

In contrast to the C_{llr} results reported in §3.3.1 above, in which the forensic-voice-comparison system outperformed 18 (78%) of the groups of listeners, in terms of classification-error rates, the forensic-voice-comparison system only outperformed about half the groups of listeners (a little more than half if responses of "1" were counted as errors and a little less than half if responses of "1" were ignored).

4. General discussion and conclusion

Expert testimony should only be admitted if it has the potential to assist the trier of fact. 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 should not be admitted.

In Part I, 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. In terms of C_{llr}, all listeners performed worse than a forensic-voice-comparison system that was based on state-of-the-art automatic-speaker-recognition technology.

In Part III, we tested the accuracy of speaker identification by collaborating groups of listeners who came to a consensus response. This was intended to be informative with respect to a context in which a jury 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 listeners' group-consensus 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. In terms of C_{llr}, the forensic-voice-comparison system outperformed 18 of the 23 groups (78% of the groups). We therefore conclude that a state-of-the-art forensic-voice-comparison system would outperform most groups of collaborating lay listeners.

In contrast to C_{llr} results, in terms of classification-error rates, the forensic-voice-comparison system only outperformed about half the groups of listeners.[28] We included results in the form of classification-error rates in order to allow for comparison with others studies that have collected categorical responses (see §2.7.1 above). Classification-error rates, however, ignore the strength of the evidence, e.g., the magnitude of the likelihood ratio output by the forensic-voice-comparison system or the degree of confidence that a jury has in the correctness of their speaker identification. We think that strength of evidence matters for how the voice evidence will contribute to the ultimate decision made by the trier of fact: a likelihood ratio just above 1 or a low-confidence same-speaker identification will not have the same effect as a likelihood ratio that is much greater then 1 or a high-confidence same-speaker identification. We therefore consider the C_{llr} results to be more meaningful than the classification-error rates.

In terms of C_{llr}, the group-consensus responses outperformed independent individual listener responses. The group-consensus responses also outperformed "wisdom of the crowd" responses consisting of the geometric means of independent-individual-listener responses. The latter result was unexpected – previous research has found that accuracy suffers when participants are exposed to other participants' responses. In that prior research, however, individuals were exposed to other participants' responses without the opportunity for discussion with those other participants, and the individuals made their own individual responses.[29]

The number of members of a jury in most common-law jurisdictions is 12. A limitation of the Part III research was that, because of recruitment and retention problems, all but one of the groups had fewer then 12 participants, and many had substantially fewer. There was, however, no apparent relationship between number of participants in a group and the resulting C_{llr} value. We therefore believe that the pattern of results of the present research can be extrapolated to groups of 12 collaborating listeners, such as juries.

Based on the results in Part I and Part III, at least under the particular case conditions tested, we infer that the forensic-voice-comparison system would satisfy the admissibility criterion of being able to assist the trier of fact, specifically of being more accurate than speaker identification performed by a judge and being more accurate than speaker identification performed by most juries. Taking into consideration the results of previous research (which was summarized in Part I §1.3) and the Part II results (which included a high-quality-audio condition), 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 and by most groups of collaborating listeners, we advise judges and juries against attempting to perform their own speaker identification. Instead, we recommend they rely on expert testimony that is based on a validated forensic-voice-comparison system.

[28] Note, however, that the selection of stimulus pairs used in Part III was biased against the forensic-voice-comparison system. See discussion in §3.1.

[29] Prior numerical information can tether or anchor and distort subsequent discussions (Rachlinski et al. [22]; Bystranowski et al. [23]). In the group-consensus experiment, participants first gave their own numerical response then discussed a consensus response. It could be though that each individual participant would therefore be tethered by their own individual response, but participants' own numerical responses may not have had the same tethering effect as has been observed for exposure to external sources of numerical information. Individual variability in error types and cognitive processes may have survived and contributed to the deliberations leading to a "wisdom of the crowd" advantage through to the consensus response. Had all the participants in a group been exposed to the same external numerical information, the results could have been different.

Over the last two decades, advances in automatic-speaker-recognition technology have resulted in substantial improvements in the performance of forensic-voice-comparison systems, and we expect further advances to be made leading to additional improvements in performance. In contrast, one would not expect the performance of untrained lay listeners to change over time. We therefore expect that the performance of forensic-voice-comparison systems will continue to improve relative to the speaker-identification performance of judges and juries.

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.

CRediT authorship contribution statement

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

References

[1] G.S. Morrison, C. Zhang, Forensic voice comparison – Overview, in: M. Houck, L. Wilson, H. Eldridge, S. Lewis, K. Lothridge, P. Reedy (Eds.), 3rd Ed.Encyclopedia of Forensic Sciences, 2, Elsevier, 2023, pp. 737–750, https://doi.org/10.1016/B978-0-12-823677-2.00130-6.

[2] 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. 341 (2022) 111499, https://doi.org/10.1016/j.forsciint.2022.111499.

[3] 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.

[4] 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.

[5] 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.

[6] N. Basu, P. Weber, A.S. Bali, C. Rosas-Aguilar, G. Edmond, K.A. Martire, G. S. Morrison, Speaker identification in courtroom contexts Part II: Investigation of bias in individual listeners' responses, Forensic Sci. Int. 349 (2023) 111768, https://doi.org/10.1016/j.forsciint.2023.111768.

[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] 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.

[9] 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.), 3rd Ed.Encyclopedia of Forensic Sciences, 2, Elsevier, 2023, pp. 720–736, https://doi.org/10.1016/B978-0-12-823677-2.00182-3.

[10] S.J. Park, A. Afshan, J. Kreiman, G. Yeung, A. Alwan, Target and non-target speaker discrimination by humans and machines. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 6326–6330, https://doi.org/10.1109/ICASSP.2019.8683362.

[11] C.F. Karpowitz, T. Mendelberg, Groups and deliberation, Swiss Political Sci. Rev. 13 (2007) 645–662, https://doi.org/10.1002/j.1662-6370.2007.tb00092.x.

[12] F. Galton, Vox populi, Nature 75 (1907) 450–451, https://doi.org/10.1038/075450a0.

[13] J. Tangen, K. Kent, R. Searson, Collective intelligence in fingerprint analysis, Cogn. Res.: Princ. Implic. 5 (2020) 23, https://doi.org/10.1186/s41235-020-00223-8.

[14] J. Lorenz, H. Rauhut, F. Schweitzer, D. Helbing, How social influence can undermine the wisdom of crowd effect, Proc. Natl. Acad. Sci. 108 (2011) 9020–9025, https://doi.org/10.1073/pnas.1008636108.

[15] 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 (2024). Available at ⟨http://e3fs3.forensic-voice-comparison.net/⟩.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] J.L. Fiechter, N. Kornell, How the wisdom of crowds, and of the crowd within, are affected by expertise, Cogn. Res.: Princ. Implic. 6 (2021) 5, https://doi.org/10.1186/s41235-021-00273-6.

[21] D. White, A.M. Burton, R.I. Kemp, R. Jenkins, Crowd effects in unfamiliar face matching, Appl. Cogn. Psychol. 27 (2013) 769–777, https://doi.org/10.1002/acp.2971.

[22] J.J. Rachlinski, A.J. Wistrich, C. Guthrie, Can judges make reliable numeric judgments: distorted damages and skewed sentences, Indiana Law J. 90 (2015) 695–739. ⟨https://www.repository.law.indiana.edu/ilj/vol90/iss2/6/⟩.

[23] P. Bystranowski, B. Janik, M. Próchnicki, P. Skórska, Anchoring effect in legal decision-making: a meta-analysis, Law Hum. Behav. 45 (2021) 1–23, https://doi.org/10.1037/lhb0000438.