Article:
HUGHES, VINCENT, LLAMAS, CARMEN and Kettig, Thomas (2026) The effects of expert conclusions and additional forensic evidence on listener evaluations in a voice comparison task. International Journal of Speech, Language and the Law. ISSN: 1748-8885

https://doi.org/10.3138/ijsll-2025-0022

The effects of expert conclusions and additional forensic evidence on listener evaluations in a voice comparison task

Abstract

Little is known about how lay people understand and evaluate forensic evidence. Yet in many jurisdictions, lay people make up juries that are responsible for determining the innocence or guilt of the accused. In this paper, we examine how lay people's evaluations of voice comparison evidence are affected by likelihood ratio (LR)-based conclusions of different magnitudes, expressed in the form of numerical and verbal statements, as well as the suggestion of the existence of additional forensic evidence in a mock case. We use a novel "jury-of-the-future" game to elicit participant judgments about whether pairs of voices belong to the same or different speakers. We then assess how those judgments are affected by expert conclusions and additional forensic evidence. Results show that verbal LRs shifted participants' responses in a consistent way relative to the strength of the conclusion presented (i.e. people thought the voices were more likely to be the same if the LR supported the prosecution). This effect was stronger for voices from the North East of England rather than Standard Southern British English (sometimes referred to as Received Pronunciation). Less consistent patterns were found for numerical LRs, suggesting that participants struggled to interpret numerical values, particularly those of the highest magnitude. When primed with the suggestion of additional forensic evidence (DNA, footprint, fingerprint), participant responses shifted towards the voices being more likely to be the same, for both same- and different-speaker pairs, with the effect being strongest for footprint and fingerprint evidence. We discuss the implications of these findings for the presentation of conclusions in forensic voice comparison cases, and for forensic science more generally.

Keywords: evidence evaluation, voice comparison, likelihood ratio, lay listeners, additional evidence

1. Introduction

Forensic voice comparison (FVC) typically involves the analysis of a recording of the voice of an unknown criminal and a recording of the voice of a known suspect (usually in a police interview). A forensic expert is tasked with comparing the speech features in the two recordings to assess the similarity between the voices and their typicality relative to a wider relevant population. In doing so, the analyst evaluates the relative strength of the evidence considering the prosecution and defence propositions. This conclusion is presented to an end-user in the form of a written report and, in some cases, oral testimony delivered to a trier-of-fact in court; in many jurisdictions around the world, this is a jury made up of lay people. It is then the responsibility of the trier-of-fact to evaluate the specific piece of FVC evidence in combination with the other evidence in the case to assess the *ultimate issue* (Lynch and McNally 2003) of the defendant's innocence or guilt.

Communicating forensic science to end-users, particularly jurors in court, remains a significant challenge. By definition, forensic expertise is requested for evidence which is "likely to be outside the experience and knowledge of a judge or jury."[1] Therefore, effective communication is essential to ensure that end-users can understand the analytical process, develop trust in the methodology being employed, and assign the appropriate weight to the evidence. Ineffective communication potentially risks miscarriages of justice, irrespective of the accuracy of the scientific work underpinning the analysis. Yet voice is an unusual form of forensic evidence in that non-experts can listen to a pair of recordings and arrive at their own view about whether they contain the voice of the same speaker or not. Voice is therefore an accessible form of evidence to lay people in a way that other forensic evidence is not (e.g. DNA, forensic chemistry, and even fingerprints). This is potentially problematic since recordings are often played to jurors in criminal trials, in addition to being presented with expert evidence in the form of a written report and/or oral testimony.

Despite its complexity and importance, little is known about the decisions made by juries and the processes that underlie those decisions. In part, this is because it is not possible to access jury deliberations in real cases, or to question jurors after the fact. In this paper, we investigate the jury decision-making process in the context of FVC evidence. Specifically, we consider the extent to which lay people's evaluations of whether two voices are the same or different are affected by expert conclusions of various forms and magnitudes, and the suggestion of additional forensic evidence in a case.

1.1 Conclusion frameworks for forensic voice comparison evidence

There is now considerable consensus within the forensic science community that the likelihood ratio (LR) is the appropriate framework for evaluating and reporting the strength of expert evidence (Aitken et al. 2011, Berger et al. 2011, Redmayne et al. 2011). Despite this, there is still variability in terms of the frameworks used in practice to

[1] https://www.cps.gov.uk/legal-guidance/expert-evidence (last viewed: 05/06/25)

express conclusions, with many forensic disciplines and jurisdictions around the world not using the LR framework at all. Where the LR is used, there is considerable variability in how those LRs are computed and then expressed. The LR can be a numerical value computed in an entirely quantitative way using statistical models and data that are representative of the conditions of the case. However, for many disciplines, principally where analysis is based on expert human judgment, it is not possible to use data and statistical models to compute a numerical LR (Evett 1998). This may be because representative data do not exist, the analytic methods do not provide easily quantifiable data, or because generating large-scale data is extremely labour intensive and time consuming. In such disciplines, experts may still use a number to represent the LR, using more subjective estimates to produce the values for the numerator and denominator. More commonly, experts use a verbal statement which expresses the extent to which the evidence supports one proposition over the other based on categorical labels (e.g. limited/moderate/strong support). Numerical LRs derived from an entirely data-driven approach may also be converted into a verbal equivalent using an established scale (see Champod and Evett 2000, Association for Forensic Science Providers 2009).

Within the field of FVC, there has long been debate about the appropriate framework to use for evaluating the strength of evidence. In many ways FVC has been ahead of many other forensic sciences in these discussions, with the first work outlining issues with posterior probability published in the 1990s and early 2000s (see Broeders 1999, Champod and Evett 2000, Rose 2002). In the late 2000s, many FVC experts adopted what is known as the *UK Position Statement* (French and Harrison 2007) which involves making separate judgments about the similarity (termed *consistency* in French and Harrison 2007) and typicality of the voices, but without a single expression of the overall strength of evidence. This approach was criticised by Rose and Morrison (2009) for ultimately falling short of using the LR framework. Since then, most experts have moved to a verbal LR scale with conclusions expressed in the form of a support statement for one proposition over another. However, there is considerable demand for experts to use numerical LRs, and to compute those LRs using representative data and statistical models (Morrison 2014, Morrison 2022). Indeed, this will become more common in casework as the use of automatic speaker recognition (ASR) systems (which output numerical LRs) increases.

1.2 Jury understanding of expert evidence

Previous work has considered the extent to which potential jurors understand forensic conclusions expressed in different ways. Some have argued against the use of numerical LRs, with Aitken et al. (2011: 1) stating that verbal expressions are “the most appropriate basis for communication”. Arscott et al. (2017) examined how verbal statements are interpreted by three groups with differing levels of experience of the criminal justice system: lay people, legal professionals, and those with forensic/ investigative knowledge. All three groups increased their evaluations of strength of evidence in line with the strength of the verbal expression. However, participants generally struggled to interpret differences at the highest end of the scale.

Further empirical work has reported other issues with verbal LR statements. Martire et al. (2013) presented lay people with three verbal LR statements of "weak or limited", "moderately strong" and "very strong" support. These categories correspond to numerical LRs of 4.5, 450, and 495,000, following the Association of Forensic Science Providers (2009) scale. They report three key findings. Firstly, people respond to shifts in the weight of evidence in the expected direction. That is, changes in belief towards guilt are greatest for *very strong* support and smallest for *weak or limited* support. Secondly, despite this, the magnitude of peoples' belief change is generally smaller than would be expected from the magnitude indicated by the verbal statement. This potentially undervalues the strength of evidence. Thirdly, results demonstrate the *weak evidence effect* (see Fernbach et al. 2011), whereby belief shifts away from guilt (i.e. towards the defence proposition) when presented with a conclusion of "weak or limited" evidence in support of the prosecution proposition. Similar patterns are replicated in Mullen et al. (2014) who found little correlation between verbal statements and perceived strength of evidence irrespective of whether participants were presented with fibre, footwear, glass, or paint evidence. This leads to the conclusion that verbal statements are "unlikely to be understood properly by lay people and it would appear that they are actually misunderstood" (p.154).

Martire et al. (2014) analysed the understanding of LRs comparing both verbal and numerical conclusions, and the use of a visual scale. Participants were presented with a small amount of information about a burglary and asked their view on the guilt of the accused. Subsequently, they were presented with high (*strong;* LR = 5500 times more likely) and low (*limited*; LR = 5.5 times more likely) magnitude LRs relating to fingerprint evidence and asked again about their view on guilt. The difference between the two responses was used as a measure of belief change. Results showed that numerical LRs "produce belief-change and implicit LRs most commensurate with the intentions of the expert" and "most resistant to weak evidence effect" (p. 61). Meanwhile, the correlation was much weaker for verbal LRs, which were also most affected by the *weak evidence effect*. A similar pattern was reported for the combination of verbal and numerical conclusions.

1.3 Factors affecting jury decision making

There are a variety of factors which potentially affect jury decision making. Jurors bring with them a unique set of life experiences and skills, which may vary according to their level of education or knowledge of the judicial process. This in turn may affect the level of understanding about forensic science, statistics and probability. Jurors also have inherently different levels of trust in legal institutions, which in some cases may come from direct interactions with the judicial system. Jurors also necessarily have a unique set of implicit and explicit biases, based on ethnicity, gender, sexuality, or other personal characteristics, which may affect the extent to which they are likely to consider someone innocent or guilty, or the extent to which they trust the evidence from a specific witness or expert (see Curley et al. 2022). As highlighted by Bornstein and Greene (2011: 64-65), the reason why individual jurors may arrive at different overall

decisions is that they “filter the evidence through their own experiences, expectations, values, and beliefs”.

Some of those biases may affect the way in which jurors perceive a defendant, and the voice has been shown to be a cue to those perceptions (see Frumkin and Stone 2020). Paver et al. (2025) found that lay people generally associate certain regional accents of the UK as more likely to be linked with criminal activity (a finding which replicates Dixon et al. 2002 amongst others). Criminality ratings on a Likert scale for Birmingham and Glasgow were generally lower than for other regional accents, such as Liverpool, London and Newcastle. Standard Southern British English (also referred to as RP) speakers were generally rated as less likely to behave in criminal ways, and those speakers were also more associated with status traits such as intelligence, education and wealth. Interestingly, ratings of criminal activity varied as a function of crime type, with the regional accents rated as more likely to be involved in crimes such as physical assault and shoplifting, but not sexual assault.

Other evidence in the case may also affect juror judgment (see Munro et al. 2024 for exploration about how information outside of the trial itself can affect juror decisions). For example, knowing that there is a lot of forensic evidence in the case may make jurors attach more weight to weaker forensic evidence, or even to dismiss contradictory evidence. Of course, other evidence in a case is not task-irrelevant for jurors. After all, it is their job to make a decision about innocence and guilt based on all of the evidence presented. However, Carlson and Russo (2001) refer to the concept of *predecisional distortion* whereby jurors attach more weight to new evidence in a trial that supports the verdict which is favoured by previous evidence. Other effects are also possible such as recency bias, whereby more weight is attached to the most recent piece of evidence, and anchoring bias, whereby a specific piece of earlier evidence is used as a basis for assessing the weight of other evidence (Tindale and Berryman 2023).

1.4 This study

The aim of the present study is to evaluate the effects of expert conclusions and the suggestion of additional forensic evidence in a mock case on lay peoples’ responses in an FVC task. More specifically, we want to know whether these sources of external information (i.e. factors not related to the voices themselves) can change listeners’ evaluations of voice sameness, as well as the direction and magnitude of any change, and how this interacts with the accent of the speaker. To do this, we asked lay people to listen to pairs of voice recordings of speakers of Standard Southern British English, Newcastle English and Middlesbrough English (two localities in the North East of England) and then to assess the likelihood that each pair contained the same or different speakers. Data collection was performed using a bespoke jury-based game allowing us to test the following research questions:

- To what extent are listener judgments in an FVC task affected by a conclusion provided by a forensic expert?

- Do listeners better understand verbal or numerical LR conclusions? And how does this interact with the magnitude of the evidence?
- Are listeners more likely to think voices belong to the same speaker (i.e. more support for guilt) when primed with the suggestion of other forensic evidence in the case?

W hile there is a body of existing research on lay peoples' understanding of expert conclusion frameworks, previous work has not assessed this issue with regard to voice evidence. More crucially, previous work has not assessed the understanding of conclusion frameworks when lay people are in a position to evaluate the evidence themselves, as is the case for voice evidence, but not for other forensics, such as DNA.

2. Methods

2.1 Stimuli

The voice recordings used as stimuli in this study were taken from two corpora. The first was the Dynamic Variability in Speech (DyViS; Nolan et al. 2009) corpus, which contains adult (18-25 years) male speakers of Standard Southern British English (SSBE). The speakers were all students at the University of Cambridge at the time of the recordings. The second corpus was collected as part of The Use and Utility of Localised Speech Forms in Determining Identity project (TUULS; Llamas et al. 2016-19). The speakers used in our study were all adult (18-65 years) male speakers from Newcastle or Middlesbrough in the North East of England. The three accents in our study were chosen to test for effects of familiarity on listeners' FVC performance. SSBE is an accent with high familiarity for all UK-based listeners. Newcastle English is also likely to be familiar to listeners, given its high exposure in the media. Middlesbrough English is segmentally similar to Newcastle English, but with lower levels of familiarity (for more, see Hughes et al. 2025). A total of 45 SSBE speakers, 15 Middlesbrough speakers, and 15 Newcastle speakers were used in the present study.

Stimuli were created from longer recordings which were representative of the conditions of typical FVC casework. Both DyViS and TUULS elicited speech through a mock police interview in which speakers had to lie about their involvement in a crime. These police interviews were all high-quality studio recordings. For the SSBE speakers, DyViS contains a landline telephone recording of a separate conversation with a supposed accomplice where each speaker discussed the mock crime again with an 'accomplice'. No 'accomplice task' speech was available from TUULS, so voice samples were taken from sociolinguistic interviews with the Middlesbrough and Newcastle speakers. Landline telephone-like quality was mimicked by applying a band-pass filter (300-3400Hz, with 100Hz smoothing at the edges of the band-pass range) and adding 25dB of white noise. From each recording, edited samples of around 10 seconds were created to use as stimuli. Identifying information and semantic content about the background of the speakers was avoided in creating the edits. Norming of the samples ensured that the content taken from all of the TUULS and DyViS tasks were similar in

style and no stimulus used in the listening task contained any overtly criminal or particularly guilty-sounding content.

Speech samples were paired to create same-speaker and different-speaker comparisons. In each case, the telephone sample was used as the nominal 'criminal' sample (i.e. the 'unknown' voice) and the high-quality interview sample was used as the nominal 'suspect' sample (i.e. the 'known' voice). This arrangement of samples replicates the most common situation in FVC casework. In total, the set of stimuli used in this study consisted of 45 same-speaker comparisons and 75 different-speaker comparisons. Most comparisons were done within accent group (i.e. SSBE, Newcastle or Middlesbrough), with the exception of 30 different-speaker comparisons involving a Newcastle voice and a Middlesbrough voice. This allowed us to test the sensitivity of listeners to regional accent variation.

2.2 Listener task

We asked our listeners to conduct something similar in design to an FVC task. This involved first listening to the nominal 'criminal' (telephone) sample and providing a judgment about the *typicality* (i.e. the distinctiveness) of the voice relative to other speakers of the same accent. In each case, the listeners were told the regional accent of this speaker. Listeners provided their typicality judgments on a 0 to 100 scale. They were then presented with the nominal 'suspect' (interview) sample and asked to provide a judgment about the *similarity* between the two voices on a 0 to 100 scale. Listeners also provided a judgment about whether they thought the voices belonged to the same speaker or not – that is, their *sameness* – again using a 0 to 100 scale. In Hughes et al. (2022), we computed LR-like scores from the similarity and typicality ratings to compare the performance of humans and automatic systems directly. However, in the present study, we focus on the results of the *sameness* ratings only, as this is more directly related to the assignment of guilt in a jury context.

Every listener was presented with a total of 24 comparisons, divided into three blocks of eight comparisons. The blocks had been curated to ensure that they consisted of three same-speaker comparisons and five different-speaker comparisons, split equally across the accent groups. Both the order of the blocks and the order of the comparisons within blocks were randomised across listeners. The experiment was a between-subjects design: each listener responded only to a subset (24) of the total number of comparisons in the study (120). This ensured that, outside of the context of same-speaker comparisons, listeners never heard the same voice or sample more than once throughout the experiment. Restrictions were not placed on the number of times a sample could be played.

2.3 Data collection

Data were collected using two different platforms. The first was a jury-based game, and the second was a Qualtrics study with no gamification.

2.3.1 Jury-based game platform

Gamification was used to immerse the participants in a ‘jury of the future’ environment. The aim here was not to replicate a real-world jury scenario. Rather, we wanted to build a context-driven narrative where it was possible for participants to respond to multiple pairs of voices (rather than a real jury scenario involving a single one-to-one comparison of voices). We also intended to increase the stakes of participant responses beyond what might typically be expected in a standard online survey. Providing the jury narrative was also used to help ensure that participants took notice of the expert conclusions and additional evidence presented to them.

The levels of the game were aligned with the three eight-comparison blocks of comparisons described in Section 2.2. In the first level, participants were presented with a standard, beige Qualtrics-style interface to provide responses to the first eight comparisons. The first level was presented to participants as a ‘Level 1: Tutorial’. At the end of the level, participants were told how many pairs of the eight they got ‘correct’ (defined as providing a *sameness* scale rating greater than 50 for same-speaker pairs and below 50 for different-speaker pairs).

Between levels one and two, participants were introduced to the game with an illustrated video mimicking a news report and court scenes describing their role on a jury of the future in which they were responsible for the fate of human jurors over machines (see Figure 1). Participants were then presented with a more futuristic, ‘technical’ looking version of the interface from level one and responded to the next block of eight comparisons. The game established that the interface was intended to represent a tablet device for the jury to provide their responses. The format of the questions remained the same as in level one, and participants were also given feedback about their accuracy after each eight-comparison level.

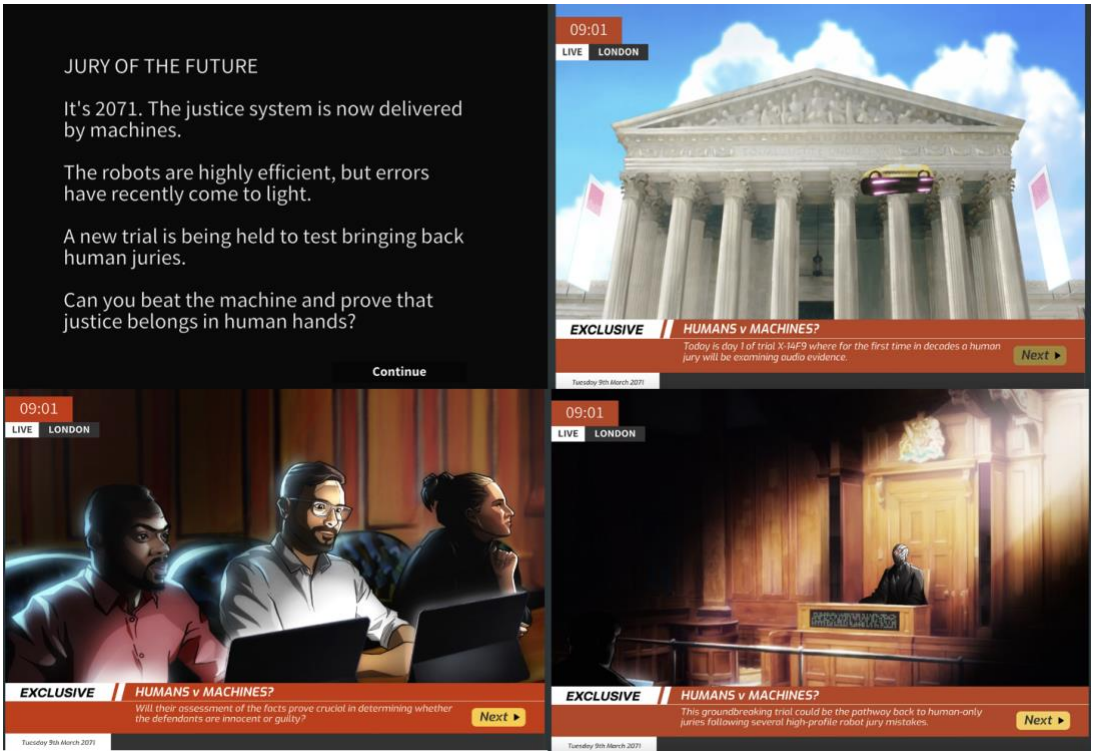

Figure 1: Screenshots of the illustrated scenes within the game-based data elicitation tool, presented to participants between levels one and two

In the third level, participants were split randomly across two conditions: (i) 'expert evidence' or (ii) 'additional evidence'. In the 'expert evidence' condition, participants were presented with an LR-based conclusion from an FVC expert (see Figure 2, left). The conclusion took the form of either a verbal statement or a numerical value. In both cases, there were three equally spaced categories (as in Martire et al. 2014), with the verbal statement and numerical values aligned following the Champod and Evett (2000) scale. The three categories were 'limited' (LR=10x), 'moderately strong' (LR=1,000x), and 'very strong' (LR=100,000x) evidence in support of either the same-speaker view or the different-speaker view. Throughout this level, the conclusion provided by the expert was always consistent with the ground truth, but the magnitude of the evidential support was randomised across comparisons. The decision to align conclusions with the ground truth was intended to maximise the extent to which the participants utilised the expert conclusion, as should be expected in a real jury trial. We considered that, for example, providing a strong conclusion in support of the same-speaker view for pairs of voices that were very clearly different would potentially confuse participants and make them more likely to entirely ignore the expert conclusions in the rest of the level.

Figure 2: Screenshots of the additional information provided to participants in level three in the form of 'expert evidence' (left) and 'additional evidence' in the case (right)

In the 'additional evidence' condition, we primed participants with the suggestion of additional forensic evidence in the case. Participants were told that there was forensic evidence in the form of DNA, fingerprints, or footprints and were provided with an image to represent each evidence type (see Figure 2, right). They were provided with no indication about the strength of evidence or whether it supported the same-speaker (guilty) or different-speaker (innocent) proposition. Given this, we expected the effects to be relatively subtle, but that, in principle, the priming could shift listener judgments more towards same-speaker rather than different-speakers; i.e. in the absence of other information the assumption would be that the additional evidence was incriminating.

Comparison pairs were counterbalanced across all three levels of the game, meaning that we have listener responses for all 120 comparisons across all levels. The median number of listener responses in levels 1 and 2 was 100. The split in level three meant that fewer listener responses were available for analysis in each condition. In the 'expert evidence' condition, the median number of listener responses per comparison was 20 for the verbal condition and 7 for the numerical condition. For the 'additional evidence' level, the median number of responses per comparison was also 7.

2.3.2 Qualtrics platform

A separate set of listeners provided responses to the same comparisons via Qualtrics. Again, listeners judged 24 pairs of comparisons each, arranged into three blocks of eight, with the same distribution of accent comparisons, same- and different-speakers, and the same process of randomisation within and between blocks as in the game. We also added a ten-second break between blocks of eight comparisons, to mimic the time course of the game. The median number of responses to each of the 120 comparisons within the Qualtrics data was 20. The Qualtrics data were intended to provide a direct

point of comparison for assessing the effects of the gamified data elicitation method. Our analysis in this paper is conducted by comparing data from the two versions of level three of the game (i.e. 'expert evidence' and 'additional evidence') with the third block of comparisons from Qualtrics (i.e. those from the same point in the chronology of the experiment).

2.4 Listeners

Participants were recruited via Prolific and sampling was conducted in the same way for both the game and Qualtrics versions of the experiment. After excluding three participants for technical issues and two participants for taking more than 71 minutes to complete the experiment, we report here data from a total of 1,804 participants: 1,503 who completed the game version and 301 who completed the Qualtrics version. Within the game, 1,205 participants completed the 'expert evidence' level (302 with numeric advice, 903 with verbal advice) and 298 participants completed the 'additional evidence' level. Participants were aged between 18 and 79 years (median=36 years) and based in the United Kingdom, with a gender split of 879 women, 902 men, and 23 non-binary or other. No additional controls over the regional background of participants were implemented. The game version of the experiment took participants an average of 14.5 minutes to complete (median=12.9 minutes, range=3.9 to 69.5 minutes), counting only time spent considering audio files (not counting pauses taken between pairs or animations presented between levels). The Qualtrics version of the experiment, inclusive of pauses taken between pairs and levels, took participants an average of 18.8 minutes to complete (median=17.3 minutes, range=8.8 to 70.8 minutes). Analysis of the data showed that the time taken to complete the task did not affect listener accuracy.

2.5 Analysis

Responses from the 'expert evidence' and 'additional evidence' levels were separately compared with the final block of eight comparisons from the Qualtrics data set. This provides a point of direct comparison involving the same pairs of voice samples, but avoids the order effects observed across the experiment as a whole. Analysis was conducted using the distribution of raw 0 to 100 responses to the question about the *sameness* of the two samples in each comparison. No normalisation of responses was applied because each listener only conducted eight comparisons within the blocks that we considered in this study. Statistical analysis was conducted using linear mixed effects models, with the response to the sameness question as the dependent variable.[2] In all models, the comparison type (same- or different-speaker) was included as an independent variable, and listener and specific comparison pair were included as random intercepts. This accounts for the fact that there will be random variation in responses across listeners and across specific pairs of speakers. For the analysis of the

[2] All statistical models took the form: SamenessResponse ~ ExperimentalCondition * GroundTruth + (1|ComparisonPair) + (1|Listener)). In the 'expert evidence' level, ExperimentalCondition was the levels of the expert conclusion including Qualtrics baseline data, with separate models fitted for verbal and numerical conclusions. In the 'additional evidence' level, ExperimentalCondition was initially responses to all of the additional evidence types pooled together and the Qualtrics baseline data. A separate model was fitted where ExperimentalCondition also broke down results by evidence type.

'expert evidence' condition, separate mixed effects models were fitted for the verbal and numerical conclusions, and the magnitude of the evidence was included as an independent variable and analysed in interaction with the comparison type (same- or different-speaker). In the 'additional evidence' condition, we report here two models: one in which all evidence types are amalgamated, with a binary Qualtrics vs. game factor analysed in interaction with the comparison type, and another in which the evidence type was included as an independent variable and analysed in interaction with the comparison type. In all cases, the Qualtrics level was dummy-coded as the baseline against which other levels were analysed; same- and different-speaker levels were relevelled as necessary to provide a baseline for significance analysis. Comparison type was included as an interaction in all models to account for the fact that the effects of the 'expert evidence' and 'additional evidence' may be different depending on whether the comparison involved the same or different speakers (note, however, that listeners did not have explicit access to the ground truth). In addition to this statistical analysis, results were also considered in terms of speaker accent.

3. Results

3.1 'Expert evidence' level

In this section, we report the results from the 'expert evidence' level, presenting the data from the verbal LR condition first, followed by the numerical LR condition.

3.1.1 Verbal LR conclusions

Figure 3 shows the distributions of responses to the sameness question on the 0 to 100 scale, as a function of the magnitude of the verbal LR conclusion presented to listeners. Qualtrics responses from the third block of comparisons (i.e. comparisons 17 to 24) are provided as a baseline where listeners were given no expert conclusion. Responses are distributed in an expected way, with listeners judging different-speaker pairs to be less likely to be the same and same-speaker pairs to be more likely to be the same when provided with an expert conclusion. The shift in listener sameness responses also aligns closely with the magnitude of the evidence presented, such that the smallest effects were found when the expert conclusion was 'limited', and the largest effects were found when the expert conclusion was 'very strong'. Smaller effects were found for different-speaker comparisons than for same-speaker pairs. For different-speaker pairs, statistical modeling demonstrated only a statistically significant difference between the Qualtrics baseline and the 'very strong' conclusion conditions (β=-2.589, p=0.048); the differences between Qualtrics and the 'limited' (β=-1.085, p=0.405) and 'moderately strong' (β=-2.274, p=0.082) conditions did not reach a p<0.05 significance level. By contrast, for same-speaker comparisons, all three verbal LR magnitudes produced statistically significant responses from the Qualtrics baseline ('limited': β=4.228, p=0.005, 'moderately strong': β=10.236, p<0.001, 'very strong': β=11.545, p<0.001).

Interestingly, slightly different patterns were found when analysing responses as a function of the regional accents involved in the comparisons. For the Newcastle and

SSBE voices, no increases in mean sameness responses were found for same-speaker pairs between the 'moderately strong' and 'very strong' conditions. For Middlesbrough voices, the difference between the 'moderately strong' and 'very strong' conditions was statistically significant. In this way, the responses to the Middlesbrough voices appear to drive the increase in overall sameness ratings for same-speaker comparisons in Figure 3. The absolute mean values for responses to the Middlesbrough voices were also lower than the responses to the Newcastle or SSBE voices across all same-speaker comparisons and all magnitudes of verbal LR conclusion.

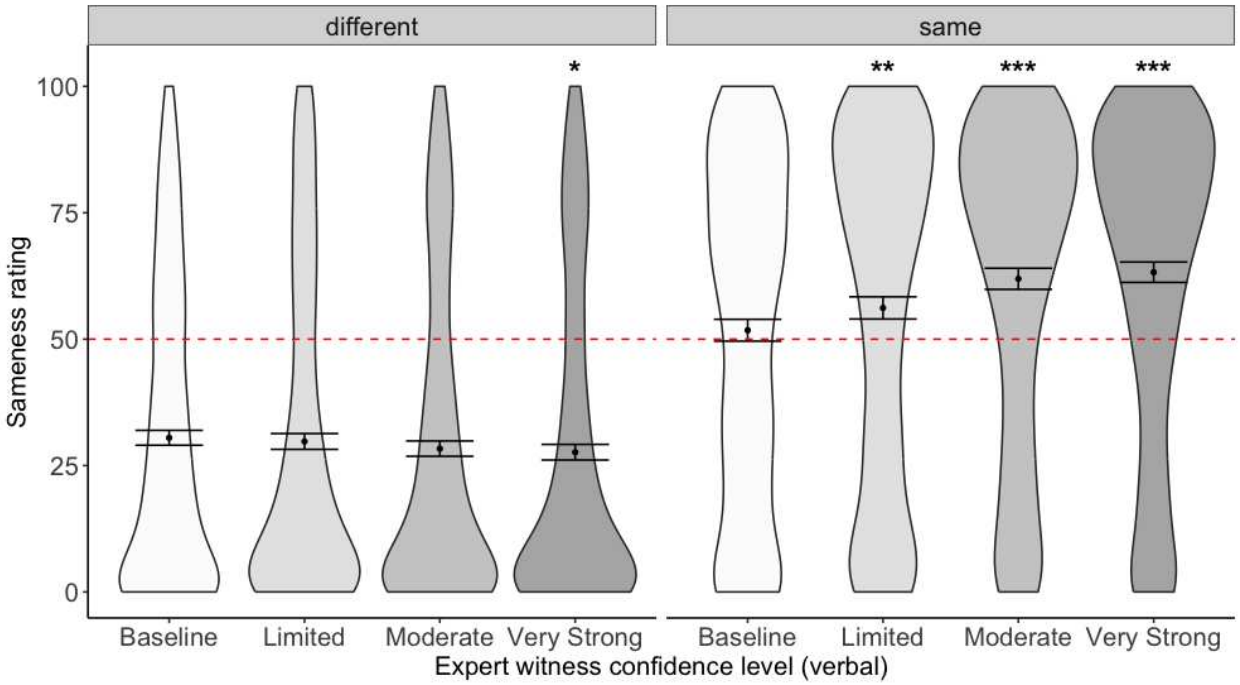

Figure 3: Distributions of sameness responses to same- (right) and different-speaker (left) pairs according to the magnitude of verbal LR conclusions (limited, moderately strong, very strong) and with responses from Qualtrics as a baseline. Whiskers represent 95% confidence intervals around the mean (significant differences from the baseline are marked: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$)

3.1.2 Numerical LR conclusions

Less consistent patterns were found in the responses to numerical LR conclusions. Figure 4 displays the distributions of sameness responses as a function of the magnitude of the numerical LR conclusion provided to listeners, with the Qualtrics responses given as a baseline.

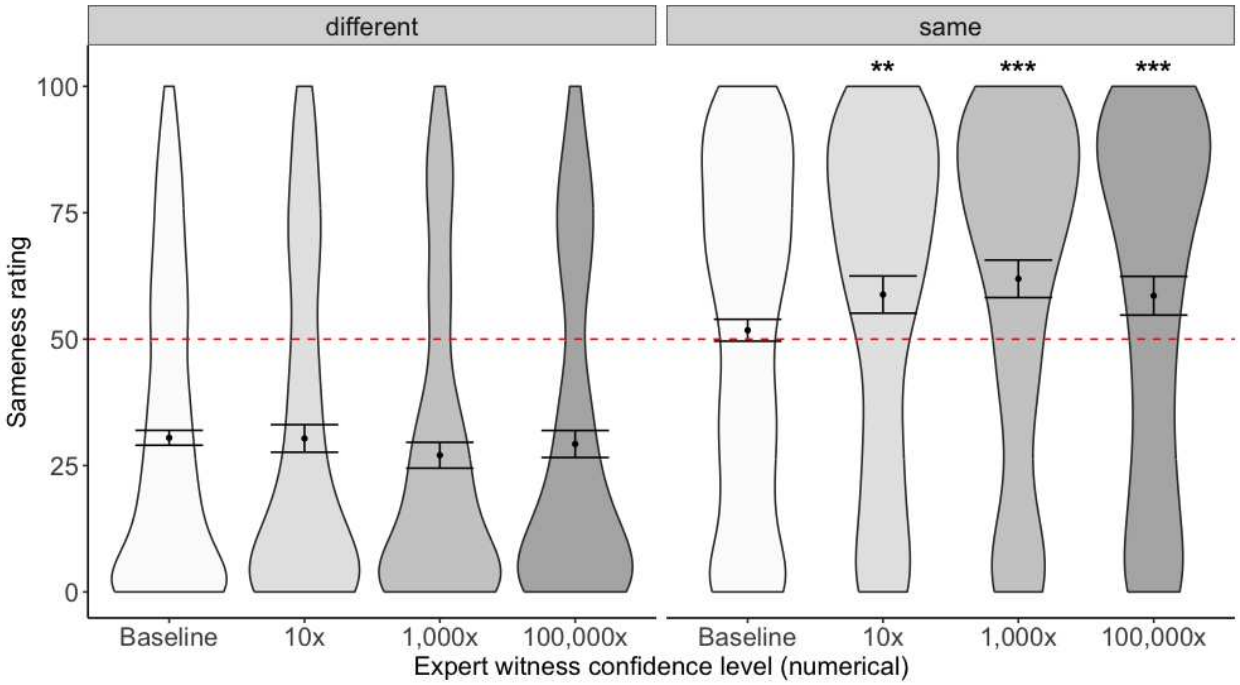

Figure 4: Distributions of sameness responses to same (right) and different-speaker (left) pairs according to the magnitude of numerical LR conclusions (10x, 1,000x, 100,000x) and with responses from Qualtrics as a baseline. Whiskers represent 95% confidence intervals around the mean (significant differences from the baseline are marked: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$)

For different-speaker comparisons, the largest shift from the Qualtrics results was found for the 1,000x condition (β=-3.378, p=0.054); that is, the numerical equivalent of 'moderately strong' conclusions, rather than the 100,000x 'very strong' conclusions (β=-0.904, p=0.606). The 1,000x condition reached close to but just shy of statistical significance, and no other significant differences were found between the Qualtrics responses and the numerical LR conditions. For same-speaker comparisons, the distributions of sameness responses in the 10x (β=6.202, p=0.003), 1,000x (β=9.383, p<0.001), and 100,000x (β=7.921, p<0.001) numerical LR conditions were significantly higher than those from Qualtrics. However, again, the biggest effect was found for the 1,000x 'moderately strong' condition, rather than the 100,000x 'very strong' condition. Unlike the verbal conclusions, no marked differences were found when analysing the results according to the regional accents involved in the comparisons.

3.2 'Additional evidence' level

Figure 5 displays the overall distribution of sameness responses on the 0 to 100 scale from the 'additional evidence' level compared with the Qualtrics baseline. Sameness responses were, on average, higher in the game compared with Qualtrics, meaning that in both same- and different-speaker conditions, listeners were more likely to think a pair of voices were the same when primed with additional forensic evidence. This effect was smaller for different-speaker pairs, where the difference was not statistically significant

(*β*=2.77, *p*=0.065), than for same-speaker pairs, where the difference was statistically significant (*β*=-5.84, *p*<0.001).

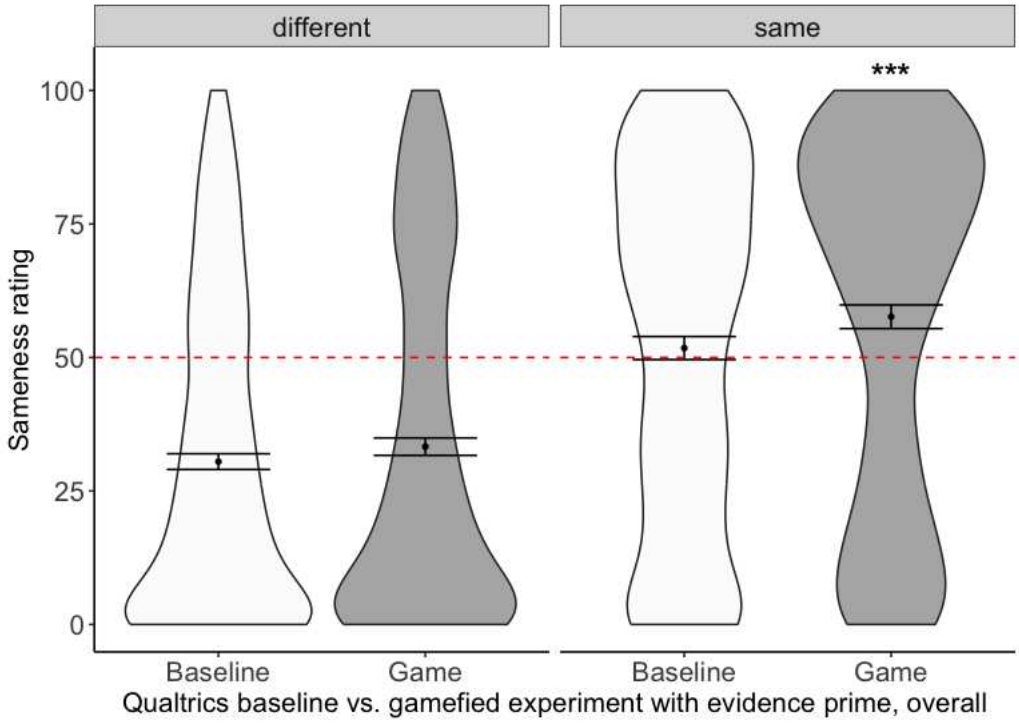

Figure 5: Distributions of sameness responses to same (right) and different-speaker (left) pairs from Qualtrics responses as a baseline versus the 'additional evidence' game level (significant differences from the baseline are marked: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$)

Variation in listener responses was also found according to the type of 'additional evidence' presented (see Figure 6). Systematic effects were found for footprint evidence which produced significantly higher sameness ratings compared with the Qualtrics baseline, for both same- (*β*=6.57, *p*=0.002) and different-speaker (*β*=3.89, *p*=0.030) comparisons. When presented with fingerprint evidence, listeners provided higher sameness ratings for same-speaker pairs compared with the Qualtrics baseline (*β*=7.74, *p*<0.001), with an effect stronger than for footprint evidence. However, no effect was found for fingerprint evidence in different-speaker comparisons (*β*=1.18, *p*=0.512). For DNA evidence, no significant effects were found for either same- (*β*=3.131, *p*=0.139) or different-speaker (*β*=3.24, *p*=0.068) comparisons; the distributions of sameness ratings provided by listeners were therefore largely unchanged irrespective of whether or not listeners were presented with additional DNA evidence.

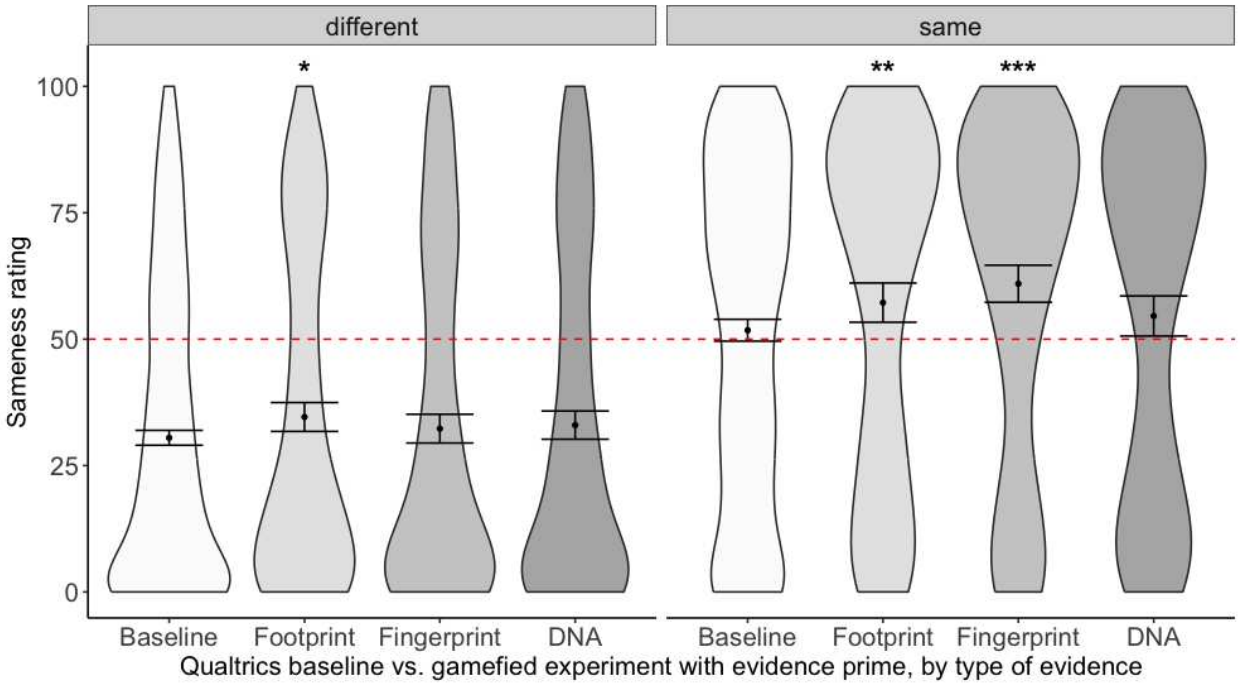

Figure 6: Distributions of sameness responses to same (right) and different-speaker (left) pairs from each of the individual 'additional evidence' types (footprint, fingerprint, DNA) with responses from Qualtrics as a baseline (white) (significant differences from the baseline are marked: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$)

4. Discussion

The present study has demonstrated a range of factors which affect the responses provided by lay listeners in a forensic voice comparison task. In contrast with previous work (see e.g. Martire et al. 2013, 2014), we found more consistent and predictable effects of expert conclusions when those conclusions were provided in the form of a verbal statement rather than a numerical value. When given a consistent-with-fact verbal conclusion, listener sameness responses on average moved in a stepwise fashion in line with the magnitude of the conclusion provided (shifting higher on the sameness scale for same-speaker comparisons and lower on the sameness scale for different-speaker comparisons). There was less systematicity in listener responses when presented with numerical LR conclusions, with no statistically significant effects found for different-speaker pairs compared with the Qualtrics baseline. While same-speaker responses to numerical LRs were all significantly higher than the Qualtrics baseline, the most extreme deviation was found for moderately strong (1,000x) evidence rather than very strong (100,000x) evidence. This finding indicates that, for our study, listeners interpreted the verbal LRs more clearly than numerical LRs. This is especially true of the highest magnitude numerical LRs, which listeners struggled to interpret. Neither the verbal nor the numerical LR responses demonstrated patterns consistent with the *weak evidence effect* reported in previous work (e.g. Fernbach et al. 2011).

A possible explanation for the differences between our results and previous work is the fact that the expert conclusions in our study interact with stimuli that listeners respond to as part of the voice comparison task. The fact that listeners were able to combine their own evaluations of the voices with the expert evidence may have provided a boost particularly in the verbal, *limited* evidence conditions. This may in turn have alleviated the *weak evidence effect*. In principle, listeners were also able to complete the task without using the expert conclusion at all, if they so wished. While it is not possible to assess the extent to which listeners utilised the expert evidence directly, the distributions of responses suggest that the expert conclusions were not attended to as consistently for the numerical LRs as for the verbal LRs. This may reflect a general difficulty in interpreting the numerical form, something which participants in previous studies could not avoid as they were responding only to the conclusion itself without any additional stimuli.

Bigger effects of expert conclusions were found for same-speaker comparisons compared to different-speaker comparisons. We believe this reflects a general pattern, demonstrated in our other work (Hughes et al. 2025), that listener performance is better (i.e. easier) with different-speaker pairs than same-speaker pairs. This is because there is more scope for differences between speakers compared with differences within speakers. Thus, the expert conclusions in our study were most useful for listeners when evaluating same-speaker pairs (without access to the ground truth).

Our results also show that the benefit of 'very strong' same-speaker conclusions is only found in responses to the Middlesbrough voices, at least for the verbal LR conclusions. This finding is consistent with the fact that people are generally much less familiar with the Middlesbrough accent, compared with SSBE or Newcastle (see Hughes et al. 2025). For SSBE and Newcastle, sameness responses were generally higher than for Middlesbrough and appear to have reached a peak by the 'moderately strong' LR condition from which they don't increase further. For Middlesbrough, there was still scope for higher overall sameness ratings in the 'very strong' LR condition, indicating that listeners struggled most with these comparisons.

The suggestion of additional forensic evidence also affected listener responses in various ways, despite us not providing any information about the direction of the evidence (i.e. whether it supported the same-speaker or different-speaker propositions). Overall, listeners evaluated the pairs of voices as being more likely to be the same when provided with additional evidence, although this effect was only significant for same-speaker pairs. As with the expert conclusion condition, the fact that different-speaker pairs were less affected by extra evidence may be due to listeners performing at ceiling already, since different-speaker pairs are generally easier than same-speaker pairs. In terms of the different types of extra evidence presented, the strongest effects were found for footprint and fingerprint evidence, with no significant effects found for DNA evidence. The explanation for our findings is not entirely clear but may be due to the choice of images used to represent each evidence type. It is possible that the footprint and fingerprint images were more clearly symbolic of criminal activity, leading listeners to shift their sameness responses towards 'guilt'. The double helix image used

to represent DNA evidence may not have had the same associations with criminality for listeners.

5. Conclusions

The results presented in this paper have demonstrated some of the factors that can affect lay peoples' judgments of voices in a forensic voice comparison task. Both the format and magnitude of expert conclusions can have substantial impact on listeners' judgments of how likely two voice recordings are to be from the same person. The suggestion of additional forensic evidence is also capable of shifting listener evaluations. Our study involved collecting data via a jury-based game. The aim was to make the experiment immersive and to increase the stakes for listeners. In this way, we wanted to capture, as far as possible within a research context, some of the elements of real jury decision-making. Our results demonstrate that even relatively subtle manipulations of key variables can have substantial effects on peoples' evaluations, to the extent that people may hear voices as being more similar when provided with additional pieces of information. Future work should further examine the interaction between various sources of case-internal and -external information on jury decision making, and how such information and biases accumulate across multiple pieces of evidence in real legal cases.

Acknowledgements

This work was supported by a UK Arts and Humanities Research Council Early Career Grant (AH/T012978/1) entitled *Humans and Machines: Novel Methods for Assessing Speaker Recognition Performance*. Thanks also to Joe Cutting and Dan Slawson who helped us develop our game and build the infrastructure to run it as an experiment.

References

Aitken, C. G. G. et al. (2011) Expressing evaluative opinion: a position statement. *Science and Justice* 51(1): 1–2.

Arscott, E., Morgan, R., Meakin, G. and French, J. (2017) Understanding forensic expert evaluative evidence: a study of the perception of verbal expressions of the strength of evidence. *Science and Justice* 57(3): 221–227.

Association of Forensic Science Providers (2009) Standards for the formulation of evaluative forensic science expert opinion. *Science and Justice* 49(3): 161–164

Berger, C. E. H., Buckleton, J., Champod, C., Evett, I. W., Jackson, G. (2011) Evidence evaluation: A response to the Court of Appeal judgment in R v T. *Science and Justice* 51: 43–49.

Bornstein, B. H. and Greene, E. (2011) Jury decision making: implications for and from Psychology. *Current Directions in Psychological Science* 20(1): 63—67.

Broeders, A. P. A. (1999) Some observations on the use of probability scales in forensic identification. *Forensic Linguistics* 6(2): 228–241.

Carlson, K. A. and Russo, J. E. (2001) Biased interpretation of evidence by mock jurors. *Journal of Experimental Psychology: Applied* 7(2): 91—103.

Champod, C. and Evett, I. W. (2000) Commentary on A. P. A. Broeders (1999) 'Some observations on the use of probability scales in forensic identification', Forensic Linguistics 6(2): 228–41. *International Journal of Speech, Language and the Law* 7(2): 239–243.

Corretge, R. (2023) Praat Vocal Toolkit. https://www.praatvocaltoolkit.com/index.html (Last viewed November 10, 2023).

Curley, L. J., Murray, J., MacLean, R., Munro, J. Lages, M., Frumkin, L. A., Laybourn, P. and Brown, D. (2022) Verdict spotting: investigating the effects of juror bias, evidence anchors and verdict system in jurors. *Psychiatry, Psychology and Law* 29(3): 323–344.

Dixon, J. A., Mahoney, B. and Cocks, R. (2002) Accents of guilt? Effects of regional accent, race, and crime type on attributions of guilt. *Journal of Language and Social Psychology* 21(2): 162—168.

Evett, I. W. (1998) Towards a uniform framework for reporting opinions in forensic science casework. *Science and Justice* 38: 198–202.

Fernbach, P. M., Darlow, A. and Sloman, S. A. (2011) When good evidence goes bad: the weak evidence effect in judgment and decision-making. *Cognition* 119: 459–467.

French, J. P. and Harrison, P. (2007) Position statement concerning the use of impressionistic likelihood terms in forensic speaker comparison cases. *International Journal of Speech, Language and the Law* 14(1): 137–144.

Fumkin, L. A. and Stone, A. (2020) Not all eyewitnesses are equal: accent status, race and age interact to influence evaluations of testimony. *Journal of Ethnicity in Criminal Justice* 18(2): 123—145.

Hughes, V., Llamas, C. and Kettig, T. (2022) Eliciting and evaluating likelihood ratios for speaker recognition by human listeners under forensically realistic channel-mismatched conditions. *Proceedings of Interspeech*. Incheon, Korea. pp. 5238–52412.

Hughes, V., Llamas, C. and Kettig, T. (2025) Human and automatic voice recognition with regionally variable speech samples. *Speech Communication* 103253.

Llamas, C., Watt, D. and French, J. P. (2016-19) *The Use and Utility of Localised Speech Forms in Determining Identity: Forensic and Sociophonetic Perspectives.* ESRC-funded project (ES/M010783/1).

Lynch, M. and R. McNally (2003) ‘Science’, ‘common sense’, and DNA evidence: a legal controversy about the public understanding of science. *Public Understanding of Science* 12: 83–103.

Martire, K. A., Kemp, R. I., Watkins, I., Sayle, M. A. and Newell, B. R. (2013) The expression and interpretation of uncertain forensic science evidence: Verbal equivalence, evidence strength, and the weak evidence effect. *Law and Human Behaviour* 37(3): 197–207.

Martire, K. A., Kemp, R. I., Sayle, M. and Newell, B. R. (2014). On the interpretation of likelihood ratios in forensic science evidence: presentation formats and the weak evidence effect. *Forensic Science International* 240: 61–68.

Morrison, G. S. (2014) Distinguishing between forensic science and forensic pseudoscience: testing of validity and reliability, and approaches to forensic voice comparison. *Science and Justice* 54(3): 245–256.

Morrison, G. S. (2022) Advancing a paradigm shift in evaluation of forensic evidence: the rise of forensic data science. *Forensic Science International: Synergy* 5: 100270.

Mullen, C., Spence, D., Moxey, L. and Jamieson, A. (2014) Perception problems of the verbal scale. *Science and Justice* 54(2): 154–158.

Munro, J., Motson, F., Turner, J., Frumkin, L. A. and Curley L. J. (2024) Double jeopardy: The effects of retrial knowledge on juror decisions. *Journal of Criminal Psychology* 14(4): 444–455.

Nolan, F., McDougall, K., de Jong, G. and Hudson, T. (2009) The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *International Journal of Speech, Language and the Law* 16(1): 31–57.

Paver, A., Wright, D., Braber, N. and Pautz, N. (2025) Stereotyped accent judgements in forensic contexts: listener perceptions of social traits and types of behaviour. *Frontiers in Communication* 9: 1462013.

Redmayne, M., Roberts, P., Aitken, C. G. G. and Jackson, G. (2011) Forensic science evidence in question. *Criminal Law Review* 5: 347–356.

Rose, P. (2002) *Forensic Speaker Identification.* Taylor and Francis: New York.

Rose, P. and Morrison, G. S. (2009) A response to the UK Position Statement on forensic speaker comparison. *International Journal of Speech, Language and the Law* 16(1): 139–163.

Tindale, R. S. and Berryman, K. (2023) Jury decision-making. In *Oxford Research Encyclopedia of Psychology*. Oxford: OUP.