A method for calculating the strength of evidence associated with an earwitness's claimed recognition of a familiar speaker

Claudia Rosas, Jorge Sommerhoff, Geoffrey Stewart Morrison

PII: S1355-0306(19)30113-3
DOI: https://doi.org/10.1016/j.scijus.2019.07.001
Reference: SCIJUS 827

To appear in: *Science & Justice*

Received date: 24 April 2019
Revised date: 1 July 2019
Accepted date: 6 July 2019

Please cite this article as: C. Rosas, J. Sommerhoff and G.S. Morrison, A method for calculating the strength of evidence associated with an earwitness's claimed recognition of a familiar speaker, Science & Justice, https://doi.org/10.1016/j.scijus.2019.07.001

A method for calculating the strength of evidence associated with an earwitness's claimed recognition of a familiar speaker

Claudia Rosas,[1] Jorge Sommerhoff,[2] Geoffrey Stewart Morrison [3,4*]

[1] Instituto de Lingüística y Literatura, Universidad Austral de Chile, Valdivia, Región de los Ríos, Chile

[2] Instituto de Acústica, Universidad Austral de Chile, Valdivia, Región de los Ríos, Chile

[3] Forensic Speech Science Laboratory, Aston Institute for Forensic Linguistics, and Centre for Forensic Data Science, Department of Computer Science, Aston University, Birmingham, England, United Kingdom

[4] Forensic Evaluation Ltd, Birmingham, England, United Kingdom

* Corresponding author. E-mail address: geoff-morrison@forensic-evaluation.net

Abstract

The present paper proposes and demonstrates a method for assessing strength of evidence when an earwitness claims to recognize the voice of a speaker who is familiar to them. The method calculates a Bayes factor that answers the question: What is the probability that the earwitness would claim to recognize the offender as the suspect if the offender was the suspect versus what is the probability that the earwitness would claim to recognize the offender as the suspect if the offender was not the suspect but some other speaker from the relevant population? By "claim" we mean a claim made by a cooperative earwitness not a claim made by an earwitness who is intentionally deceptive. Relevant data are derived from naïve listeners' responses to recordings of familiar speakers presented in a speaker lineup. The method is demonstrated under recording conditions that broadly reflect those of a real case.

Keywords

earwitness; familiar speaker recognition; strength of evidence; likelihood ratio; Bayes factor

1. Introduction

The present paper proposes and demonstrates a method for assessing strength of evidence when an earwitness claims to recognize the voice of a speaker who is familiar to them. We calculate a Bayes factor that answers the question: What is the probability that the earwitness would claim to recognize the speaker if the speaker they heard was the speaker they claimed to recognize, versus what is the probability that the earwitness would claim to recognize the speaker if the speaker they heard was actually some other speaker from a specified population? We only consider this question with respect to claims made by cooperative earwitnesses and not claims made by earwitnesses who are intentionally deceptive. We calculate Bayes factors for naïve listeners' recognition of familiar speakers under conditions broadly reflecting those of a real case. Our aim is not to provide a Bayes factor value with respect to that particular case, but to present a method that could potentially be used in other cases involving familiar speaker recognition.

Although the speaker the earwitness actually heard is not always an "offender" and the person the earwitness names is not always a "suspect", for simplicity we will hereinafter adopt these terms. Hence, using this terminology, we calculate a Bayes factor that answers the question: What is the probability that the earwitness would claim to recognize the offender as the suspect if the offender was the suspect versus what is the probability that the earwitness would claim to recognize the offender as the suspect if the offender was not the suspect but some other speaker from the relevant population? Other terms can be substituted for "offender" and "suspect" as appropriate for the particular case.

1.1. Familiar-speaker recognition and unfamiliar-speaker identification

There is a substantial amount of research literature on speaker recognition and speaker identification by naïve listeners. In this literature, "speaker recognition" refers to the scenario in which a listener hears a speaker and claims that the speaker is a person with whom they are already familiar, and "speaker identification" refers to the scenario in which a listener hears an unfamiliar speaker then later hears a speaker and claims that the latter is the same speaker they heard earlier. Reviews of the research literature from the perspective of potential application in legal contexts can be found in [1]–[5], and publications focused on related legal issues include [6]–[10].

Although the existing research literature may be informative about speaker recognition and speaker identification in general, little (if any) research appears to have been conducted under conditions that attempt to reflect those of actual legal cases, and none appears to have addressed the question that a forensic practitioner working in the likelihood-ratio framework would set out to answer in an

actual speaker recognition case: What is the strength of evidence associated with this particular earwitness's claim to recognize this particular speaker under the conditions encountered in this particular case? The method we propose is intended to answer this question, and is grounded in modern thinking on forensic inference and statistics as represented by the likelihood-ratio framework (e.g., [1], [5], [11]–[17]).

Likelihood ratios were calculated in Yarmey et al. [18] for non-familiar-speaker identification, but the conditions of the experiment did not reflect those of a legal case, calculations were based on data pooled across two speakers and multiple listeners, and likelihood ratios were not framed as quantifications of strength of evidence. In Yarmey et al. [18], the likelihood ratios were not called "likelihood ratios", but "diagnosticity indices". "Diagnosticity index", and the variants "diagnosticity ratio", "diagnostisity measure", and "diagnostisity value" are terms that appear to be peculiar to the eyewitness research literature (e.g., [19]–[21]); Yarmey et al. [18] included both earwitness and eyewitness experiments. The eyewitness literature appears to be independent of the broader forensic inference and statistics literature – we did not find references to the latter in the former. The eyewitness literature does not appear to treat a likelihood ratio value as a quantification of strength of evidence that a forensic practitioner would present to a court and that a trier of fact would (in theory) then be able to use to update their beliefs (an exception to this may be [22]).[1]

Rose [1] p. 99 proposed the calculation of likelihood ratios based on listeners' claimed recognitions of familiar speakers and the use of the resulting likelihood ratio values as quantifications of strength of evidence. The example given in Rose [1] was based on previously published data from [26]. The data were 10 listeners' claimed recognitions of one recording each of 10 familiar speakers and 2 unfamiliar speakers. The recordings were of telephone transmitted speech. In contrast to the proposal we make in the present paper to calculate a Bayes factor for a listener's claimed recognition of a particular speaker, the likelihood ratio calculations in Rose [1] were based on response data that had been either pooled across speakers for a single listener or pooled across both speakers and listeners. We are unable to relate this pooled-data approach to a question that would be of interest in an actual case – what would constitute the evidence was not made clear (we discuss

[1] As outsiders working in a different paradigm, a recent debate in the eyewitness literature about the relative merits of "diagnosticity" versus receiver operating characteristics (ROC) curves (e.g., [23] and [24]) seems to us to be misplaced (scientists working in different paradigms may be concerned with different questions, Kuhn [25]). It may be that the two communities of scientists can learn from each other, but we will have to be careful not to misunderstand each other due to apparently similar concepts and vocabulary actually having different meanings for the different communities.

this issue further in §1.3, §2.8, and §4.2).

1.2. A real case scenario

The last author of the present paper was approached to give advice on how to assess strength of evidence in relation to claimed speaker recognition in an actual case. Assessing strength of evidence did not go ahead in that case, but that enquiry inspired the research reported in the present paper. Details presented here about the case are kept to a minimum.

A call was made to emergency services, and the call was recorded at the call center. The call was made using a mobile telephone. The caller was a female victim who was in the trunk of a parked car when the call was made. Most of the recording was of sounds made by the victim. During a short section of the recording, the voices of two males could be heard in the background. The recording of the male speakers was mostly unintelligible.

A suspect was identified based on other evidence. Relatives and friends of the suspect were played the recording and asked: Do you recognize the voice of either of the male speakers? If yes, who is that speaker? Some of the listeners claimed to recognize one of the voices as that of the suspect. The section of the recording for which they claimed to recognize this speaker was approximately 3 s long.

Advice from forensic speech science practitioners was not sought until after the procedure described above had been carried out.

1.3. Speaker lineup procedures

The scenario in the original case is one of familiar-speaker recognition, i.e., a listener hears a voice and claims to recognize it as the voice of a particular speaker they already know. This differs from an unfamiliar-speaker identification scenario in which a listener, e.g., an earwitness to a crime hears the voice of a person they do not know, then later hears the voices of several speakers and is asked whether any of those speakers are the speaker they heard earlier. Best practice for unfamiliar-speaker identification involves presenting a speaker lineup in which the suspect is one of several speakers, and the listener is told that the person they heard earlier may or may not be in the lineup – note that the suspect may or may not be the offender. The other speakers in the lineup, the speakers other than the suspect, are called “foils”. The foils’ voices and the speaking style and conditions under which the voices are recorded must be such that the suspect’s voice does not stand out – listeners with no prior involvement in the case should not be able to pick out the voice of the suspect. For an example of a protocol for conducting unfamiliar-speaker-identification lineups see

Broeders & van Amelsvoort [27]. For an example of a protocol for selection of foil speakers see de Jong-Lendle et al. [28]. Further discussion of best practices for speaker lineups appears in: Yarmey [3] pp. 126–128; Sherrin [4] pp. 856–859; Morrison et al. [5] §99.960; and references cited therein. "Showups" in which the listener hears only one speaker and is asked if the speaker is the same speaker as they heard earlier have been criticized as suggestive (see [2], [4], [9], [18], [29]). Showups suggest to the earwitness that the police have reason to believe that the single speaker is the speaker that the earwitness heard earlier – showups bias the earwitness to think that the suspect is the offender.

A familiar-speaker recognition scenario could involve a true earwitness, i.e., the listener is present while the crime is being committed and recognizes the voice of the offender at that time. There is usually no recording of the offender, so a forensic voice comparison cannot be conducted. In such a scenario, the method we propose in the present paper could potentially be applied post hoc to assess strength of evidence associated with the earwitness's claim to have recognized the offender while the crime was being committed. Note that the evidence is the earwitness's claim to have recognized the offender while the crime was being committed. We use "recognize" with the implication that the earwitness not only claims that the speaker is familiar to them but also names the speaker (or otherwise indicates a particular individual). For the purposes of the present paper, we assume that the person named by the earwitness then becomes the suspect. Note that the offender and the suspect may or may not be the same person – that is the question before the court. The court will make a decision as to which is more likely to be true (beyond a reasonable doubt or on the balance of probabilities) assisted by the strength of evidence calculated for the earwitness's claim to have recognized the offender.

According to the definition given in the previous paragraph, the original case described in §1.2 is not a true earwitness scenario. The listeners were played a recording including only two male speakers and were asked whether they recognized either of the speakers. This is a form of showup. Given that the listeners were relatives and friends of the suspect and likely to a priori have a good idea of who it is that the police wanted them to recognize, the procedure was likely to have induced an expectation bias. The fact that there were two speakers rather than one is unlikely to have substantially mitigated that bias. The method we propose is one that would be preferable to such a showup procedure. It involves a lineup procedure that requires listeners to listen to and attempt to recognize the voices on multiple recordings of each of a larger number of familiar speakers mixed in with some unfamiliar speakers. If the lineup procedure were used instead of a showup procedure, the original recording of the offender would be included in the lineup, and the listener's response to the offender recording in the lineup would constitute the evidence. If a showup has already been

conducted, the offender recording would not be included in the later lineup, and the listener's response to the offender recording in the earlier showup would constitute the evidence.

Understanding what constitutes the evidence is a prerequisite to being able to calculate a strength of evidence. This fact does not appear to have been fully appreciated in earlier proposals to calculate likelihood ratios based on responses to speaker lineups.

In the original case a recording of the questioned-speaker was available, so, rather than conduct a speaker lineup, a forensic voice comparison could potentially have been performed. An empirically-validated procedure based on relevant data, quantitative measurements, and statistical models, with direct reporting of the likelihood ratio or Bayes factor output by the model would be much less susceptible to cognitive bias (see, for example, arguments in [30]).

In §2 we describe the speaker lineup procedure as we implemented it under conditions broadly reflecting those of the original case. The conditions are forensically realistic, but, as a demonstration of the method rather than an attempt to answer questions specific to the original case, we did not try to replicate all details of the original case. We did not test the same listeners. Using the categories defined by Yarmey et al. [31], in the original case the familiarity of the listeners with the speaker may have been "high", whereas in the present research the familiarity of the listeners with the speakers likely ranged from "low" to "moderate". The make and model of car used was not the same as in the original case. The language spoken was not the same. We do not know whether in the original case the male speakers were inside or outside the car. For the original case we would have tested both conditions, but for the demonstration we only tested speakers inside the car. We have not attempted to replicate the particular recording system at the emergency call center (such systems usually save the recordings in a lossy compressed format). We have not attempted to replicate the particular playback equipment and listening environment of the original case. The results given in the present paper do not therefore represent strength of evidence values for the original case. This was not our intent, our intent was to demonstrate a procedure that could potentially be used in other familiar-speaker recognition cases.

1.4. Supplementary material

The listening experiment, including the acoustic stimuli, is available at: [text redacted for blinding]

The anonymized results of the listening experiment and the Matlab code used to calculate Bayes factors based on those results are available at [32].

2. Method

2.1. Speakers

Speakers consisted of a total of 23 adult males, 18 who would be familiar to the listeners plus 5 who would not be known to the listeners. Of the 18 familiar speakers, 5 were faculty members from [text redacted for blinding], and the other 13 were famous people: [text redacted for blinding].

2.2. Recording

A telephone call was established from a mobile telephone (Samsung Note 4), and the telephone was placed in the truck of a car (Citroën Picasso) which was in a parking lot at a time of day when there was only occasional traffic. The far end of the call was recorded using a TASCAM Linear PCM Recorder DR-40 acoustically coupled to a landline telephone inside a sound-insulated box.

Each speaker was recorded separately. The speaker sat in the front of the car and responded to open questions asked by a female researcher. Audio recordings of the famous people were obtained from broadcast media, and were played from a loudspeaker (NTI Audio Talkbox, which is calibrated to 60 dBA SPL at 1 m) placed at head height on a front seat of the car and pointing toward the other front seat.

The quality of the resulting audio recordings was poor. They had low signal to noise ratios.

2.3. Stimuli

Six short sections were extracted from each speaker's recording. Each section was ~3 s long. The sections were manually selected from within each speaker's recording, with the conditions that they contain only the speech of the speaker of interest and that they not overlap or be contiguous with one another. The total number of stimuli was 138 (6 sections × 23 speakers).

2.4. Listeners

Listeners were 31 students from [text redacted for blinding]. Potential participants were asked not to participate if they had hearing problems.

2.5. Listening experiment

The listening experiment was presented online via a web browser. Listeners participated one at a time at a place convenient to them. Listeners were asked to do the experiment in a quiet place, but no constrains were placed on the audio playback equipment they used. Listeners first saw

information related to informed consent. If they agreed to continue, they then saw instructions. No personal identifying information was collected from the listeners. Listeners could take a rest at any time and resume later as long as they did not close the browser. Listeners could abandon the experiment at any time and their responses would not be submitted – to submit their responses, they had to click the "submit" button on the final screen of the experiment. Listeners were asked only to complete the experiment once, and not to discuss it with other potential participants until after the period for data collection was complete.

Listeners were presented with one recording section at a time. The sections were presented in random order. A listener saw a screen with a play button, and could listen to the section as many times as they wanted. On the screen there was also a text-entry box, a "continue" button, and the following instructions: "If you recognize the speaker, write their given name and surname in the box then press 'continue'. If you do not recognize the speaker, leave the box empty and continue to the next recording." Prior to the experiment proper, as part of the instructions, the listener saw a demonstration of how to respond. The demonstration used a good-quality recording of a famous female speaker, [text redacted for blinding].

2.6. Data coding

The raw data consisted of the written response of each listener to each recording section. The raw data were coded and anonymized by the first author. Each speaker was given a unique numeric code, and that code was used in place of the speaker's name. Coding took account of spelling variants in listeners' responses. For each speaker, a list of variant spellings of their name was created. Initial lists of variants included correct spellings and anticipated misspellings. All responses were automatically coded according to the lists of variant spellings. Responses that contained text that did not appear in the lists were automatically flagged. The first author then went through the flagged responses. When a flagged response was due to a variant spelling not already included in the lists, and the intended name was obvious, the first author added the new spelling to the appropriate list. All responses were then immediately automatically recoded using the revised list, hence any other occurrences of the same variant did not require the first author's attention. In addition to adding variant spellings for speakers who actually contributed stimuli, this process also involved adding the names of speakers who did not contribute stimuli but whom the listeners named in their responses. The anonymized version of the data was used for all subsequent analysis.

2.7. Statistical analysis

We begin this section by defining symbols. To elucidate via a concrete example, we use the name

of a famous [text redacted for blinding] as the designated speaker (he was not one of the speakers used in the present research). The observed count of responses in which the speaker was [text redacted for blinding] and a particular listener gave the name [text redacted for blinding] is c_{1+}. The observed count of responses in which the speaker was [text redacted for blinding] but the listener did not give the name [text redacted for blinding] is c_{0+}. The observed count of responses in which the speaker was not [text redacted for blinding] but the listener gave the name [text redacted for blinding] is c_{1-}. The observed count of responses in which the speaker was not [text redacted for blinding] and the listener did not give the name [text redacted for blinding] is c_{0-}. Hence, c_{1+}, c_{0+}, c_{1-}, and c_{0-}, refer respectively to the counts of "hits", "misses", "false alarms", and "correct rejections", see Table 1. The total number of times that the listener was presented with a recording section of [text redacted for blinding] is $n_+ = c_{1+} + c_{0+}$. The total number of times that the listener was presented with a recording section of someone other than [text redacted for blinding] is $n_- = c_{1-} + c_{0-}$. The listener not giving the name [text redacted for blinding] includes the listener giving the name of someone else and the listener not giving any name, i.e., stating that they do not recognize the speaker. The sets of variables $\{c_{1+},\ c_{0+},\ n_+\}$ and $\{c_{1-},\ c_{0-},\ n_-\}$ can be calculated for any specified combination of a particular listener and a designated speaker. Below we will drop the "+" and "−" subscripts when the discussion and calculations are relevant irrespective of the set of variables.

Table 1. Matrix of relationships of variables to stimulus-response pairs. Each variable in the table refers to a count. The name of any designated speaker can be substituted for "[text redacted for blinding]".

		Name given by listener		
		[text redacted for blinding]	no name or a name other that [text redacted for blinding]	total
Actual speaker	[text redacted for blinding]	c_{1+} hit	c_{0+} miss	n_+
	someone other than [text redacted for blinding]	c_{1-} false alarm	c_{0-} correct rejection	n_-

For each speaker who contributed stimuli, a count was made of the number of responses in which a listener gave that speaker's name when the stimulus was a recording of that speaker, c_{1+}, and a count was made of the number of responses in which the same listener gave that speaker's name when the stimulus was a recording of a different speaker, c_{1-}. Dividing these counts by, respectively, the number of opportunities to give a correct response, n_+, and the number of opportunities to give an incorrect response, n_-, see Eq. 1, would provide proportions that could be used as maximum likelihood estimates of the probabilities of correct and incorrect responses, θ_+ and θ_- respectively. Dividing the former by the latter, see Eq. 1, would provide a likelihood ratio answering the question: What is the probability that the listener would say the designated speaker's name if the recording they heard was of that speaker, versus what is the probability that the listener would say the designated speaker's name if the recording they heard was of some other speaker from the relevant population? The other speakers who contributed stimuli were intended to represent the population of adult male [text redacted for blinding] speakers.

(1)

$$LR = \frac{\theta_+}{\theta_-} = \frac{\left(\frac{c_{1+}}{n_+}\right)}{\left(\frac{c_{1-}}{n_-}\right)}$$

A problem occurs when there is a zero count in either the numerator or denominator, as this would give an estimated likelihood ratio of either zero or infinity. Even with non-zero counts, when n is small the proportion c_1/n may be a poor estimate of the probability for the population. To resolve this problem we will apply a Bayesian analysis; thus we will calculate Bayes factors rather than likelihood ratios.[2]

For the Bayesian analysis, we use a beta-binomial model (see, for example, [34] §3.3). The likelihood of the observed count, c_1, is given by the binomial distribution $\mathrm{Bin}(c_1|\theta, n)$. The conjugate prior is given by the beta distribution $\mathrm{Beta}(\theta|a, b)$, in which a and b are hyperparameters. Via Bayes' Theorem, the posterior distribution for θ is proportional to the

[2] Readers unfamiliar with the difference between likelihood ratios and Bayes factors and looking for a brief non-polemical introduction may wish to consult Etz [33]. The relative merits of the use of likelihood ratios and Bayes factors in quantifying strength of evidence have recently been debated in the forensic inference and statistics literature, including in a recent virtual special issue in *Science & Justice*: https://www.sciencedirect.com/journal/science-and-justice/special-issue/102F0FGVD03

likelihood multiplied by the prior distribution. The posterior distribution is therefore proportional to a beta distribution for which the posterior parameter values are the sums of the counts and the hyperparameter values, see Eq. 2 (for simplicity we have dropped constants that do not depend on θ, and thus use "proportional to" rather than "equals").

(2)

$$p(\theta|c_1, c_0, a, b) \propto \mathrm{Bin}(c_1|\theta, n)\mathrm{Beta}(\theta|a, b)$$

$$\propto (\theta^{c_1}(1-\theta)^{c_0})(\theta^{a-1}(1-\theta)^{b-1})$$

$$\propto \theta^{c_1+a-1}(1-\theta)^{c_0+b-1}$$

$$\propto \mathrm{Beta}(\theta|c_1 + a, c_0 + b)$$

The expected value for the posterior distribution, i.e., the posterior mean, $\bar{\theta}$, is given in Eq. 3, in which $m = a + b$.

(3)

$$\bar{\theta} = \int_0^1 \theta \mathrm{Beta}(\theta|c_1 + a, c_0 + b)d\theta = \frac{c_1 + a}{c_1 + a + c_0 + b} = \frac{c_1 + a}{n + m}$$

Independently calculating the posterior mean for the numerator and the denominator, we can then calculate the Bayes factor as in Eq. 4.

(4)

$$BF = \frac{\bar{\theta}_+}{\bar{\theta}_-} = \frac{\left(\frac{c_{1+} + a_+}{n_+ + m_+}\right)}{\left(\frac{c_{1-} + a_-}{n_- + m_-}\right)}$$

Given the small amount of data in the numerator ($n_+ = 6$), the results will be sensitive to the choice of prior. To reduce the potential for cognitive bias, one should specify ones priors before examining the data and not subsequently change one's choice of priors. We choose to use Jeffreys reference priors, [35]–[38], which for the beta distribution has hyperparameter values of $a_+ = 0.5$ and $b_+ = 0.5$ (hence $m_+ = 1$). Jeffreys reference priors have been proposed as "non-informative" or "objective" priors. Irrespective of arguments as to whether these or any other priors are actually non-informative or objective, Jeffreys reference priors are widely used in Bayesian statistics and thus are arguably generally accepted.

Whereas in the numerator $n_+ = 6$ is the count of 6 stimuli from 1 speaker, in the denominator $n_- = 132$ is the count of 6 stimuli from each of 22 speakers. The prior we use for the denominator is equivalent to one Jeffreys reference prior per speaker, i.e., $a_- = 0.5 \times 22 = 11$ and $b_- = 0.5 \times 22 = 11$ (hence $m_- = 22$). If we did not adjust the relative strength of the priors in the numerator and denominator to reflect the ratio of n_- to n_+ the results would be biased toward high Bayes factor values. This is illustrated in Appendix A.

Plugging the chosen values for the hyperparameters (and the values of n_+ and n_-) into Eq. 4, Eq. 5 gives the equation for calculating the Bayes factor value for each listener's claimed recognition of each speaker. Eq. 5 is specific to the present study, Eq. 4 is the general equation.

(5)

$$BF = \frac{\left(\frac{c_{1+} + 0.5}{n_+ + 1}\right)}{\left(\frac{c_{1-} + 11}{n_- + 22}\right)} = \frac{(c_{1+} + 0.5)(n_- + 22)}{(c_{1-} + 11)(n_+ + 1)} = \frac{c_{1+} + 0.5}{c_{1-} + 11} \times \frac{154}{7} = \frac{c_{1+} + 0.5}{c_{1-} + 11} \times 22$$

2.8. Comments on the evidence and hypotheses considered

Although some listeners gave the names of some speakers who did not contribute stimuli, Bayes factor values were only calculated for speakers who actually contributed stimuli. Our lineup procedure is not a database-search procedure designed to suggest candidates that may warrant further investigation, i.e., to suggest potential suspects. Instead, a suspect has already been identified and the purpose is to evaluate the strength of evidence associated with a listener's claim to have recognized the offender as the suspect. We are only interested in Bayes factors for which the hypothesis in the numerator is that the offender is a particular individual already designated as the suspect. We are not interested in Bayes factors for which the hypothesis in the numerator is that the speaker is some particular speaker other than the suspect, e.g., some other speaker that the listener may happen to name during the lineup. In an actual case there is usually only one suspect, but for the purpose of demonstrating the method we treat each speaker who contributed stimuli as a suspect, and thus calculate a Bayes factor for each speaker who contributed stimuli.

One could calculate Bayes factors for other evidence and other hypotheses. For example:

- the evidence could be that the listener did not recognize the offender, and the hypothesis in the numerator could be that the offender is a designated speaker who is familiar to the

listener; or

- the evidence could be that the listener did not recognize the offender, and the hypothesis in the numerator could be that the offender is a speaker who is not known to the listener; or
- the evidence could be that the listener claimed to recognize the offender as a particular speaker who is familiar to the listener, and the hypothesis in the numerator could be that the offender is a designated speaker who is the brother of the speaker whom the listener named.

For simplicity, in the present study we only examine Bayes factors that answer the question: What is the probability that the listener would say a designated familiar speaker's name if the recording they heard was of that speaker, versus what is the probability that the listener would say the designated speaker's name if the recording they heard was of some other speaker from the relevant population? The other speakers who contributed stimuli were intended to represent the population of adult male [text redacted for blinding] speakers.

3. Results

Table 2 provides the raw counts of the number of times each listener responded with each speaker's name. Results are only shown for speakers who contributed stimuli. Speakers 101 through 105 were faculty members familiar to the listeners, Speakers 201 through 205 were unfamiliar speakers, and Speakers 301 through 315 were famous people familiar to the listeners. In each cell, the number to the left of the vertical bar is the number of times that the listener responded with the speaker's name when the stimulus was a recording of that speaker, and the number to the right of the vertical bar is the number of times that the listener responded with the speaker's name when the stimulus was not a recording of that speaker. If the former is greater than the latter the text in the cell is blue, if the latter is greater than the former the text in the cell is red (see the electronic version of the present paper for color). The cell is blank if both values are zero. If one value is non-zero, the cell contains numbers and has a white background. If both values are non-zero, the cell contains numbers and has a gray background. Values pooled across all speakers and pooled across all listeners are provided on the margins of the table. The bottom row (labelled "total resp.") gives the total number of responses from each speaker, including responses that were the names of speakers who did not contribute stimuli.

Table 3 shows the Bayes factor values corresponding to the counts given in Table 2. Integers and fractions with integers in the denominator are exact values. Other values are given to one decimal

place. For Bayes factor values greater than 1 the text is blue, and for Bayes factor values less than 1 the text is red (see the electronic version of the present paper for color). All other formatting of Table 3 is the same as for Table 2. Trivially, the Bayes factor value for each blank cell is 1. These cells correspond to combinations of listener and speaker for which the listener never gave the speaker's name.

Table 2. Raw counts of the number of times each listener responded with each speaker's name. See main text for further explanation.

speakers \ listeners	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	all
101																	0|1			0|1		0|1						6|10			6|13
102																									0|2							0|2
103																																
104			3|0							2|0	5|0		3|0					2|0						1|0				6|0			22|0
105	2|0		1|0		2|0	2|0	5|4	2|0		2|0	6|2	3|0					5|0		1|0		5|1		5|0	5|5		1|0		6|26		5|1	58|39
201																																
202																																
203																																
204																																
205																																
301																								2|0								2|0
302																																
303																										1|0						1|0
304																																
305																										2|1						2|1
306																																
307																																
308																	1|0			2|0												3|0
309																																
310																																
311																																
312																																
313														0|1	0|1						0|1										0|3
all	2|0		4|0		2|0	2|0	5|4	2|0		4|0	11|2	3|0	3|0	0|1	0|1		1|1	7|0		3|1	0|1	5|2		7|0	6|7	3|1	1|0		18|36		5|1	94|58
total resp.	2	0	5	0	2	2	9	2	1	4	13	3	3	1	2	1	5	8	0	4	3	7	0	7	14	15	1	0	65	0	6	185

Table 3. Bayes factors based on the number of times each listener responded with each speaker's name. See main text for further explanation.

speakers \ listeners	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
101																	1/1.1			1/1.1		1/1.1							6.8		
102																									1/1.2						
103																															
104			7							5	11		7					5							3				13		
105	5		3		5	5	8.1	5		5	11	7						11		3		10.1		11	7.6		3		3.9		10.1
201																															
202																															
203																															
204																															
205																															
301																								5							
302																															
303																										3					
304																															
305																										4.6					
306																															
307																															
308																	3			5											
309																															
310																															
311																															
312																															
313														1/1.1	1/1.1						1/1.1										

4. Discussion

4.1. Bayes factor results per speaker per listener

The most obvious observation on the results is that most combinations of listeners and speakers did not produce any responses, either correct or incorrect. Of 4278 opportunities to give a response (6 sections × 23 speakers × 31 listeners), or 3348 opportunities excluding the 5 unfamiliar speakers, there were only 185 responses total (including 33 responses that were names of speakers who did not contribute stimuli). Given the short recordings with poor audio conditions it appears to have been very difficult for the listeners to recognize the speakers.

Of the 18 familiar speakers, only 7 were correctly recognized at least once by at least one of the 31 listeners. Only 2 speakers were ever correctly recognized by more than two listeners. Speaker 104 was correctly recognized at least once by 7 listeners and Speaker 105 was correctly recognized at least once by 17 listeners. Both of these familiar speakers were faculty members rather than famous people. A potential explanation for why these speakers were correctly recognized at a higher rate is that they could have been more familiar to the listeners than the other speakers. Another potential explanation is that they could have distinctive voices, i.e., voices that would be atypical with respect to the rest of the population that listeners are used to hearing (and in this case potentially atypical with respect to the other speakers who contributed stimuli). In the first author's judgment, both these speakers have atypical voices. Speaker 104, in particular, has a high-pitched voice. Previous research on unfamiliar speaker identification has found that atypical speakers are easier to recognize, e.g., [39], although speakers who are atypical in the same way (e.g., both have the same accent that is atypical from the perspective of the listeners) are more likely to be confused with each other, e.g., [40]. In [41] an unfamiliar speaker was mistakenly identified as a familiar speaker; both speakers had the same accent, and that accent was atypical from the perspective of the listeners.

In addition to variation related to speakers, there was also variation related to listeners: 6 listeners gave no responses at all, 21 gave less than 10 responses, 3 gave 10–15 responses, and 1 gave 65 responses. We have no explanation for why the latter listener (Listener 29) gave so many responses compared to the other listeners. A large number of Listener 29's responses were false alarms, but this listener was the only listener to have Bayes factors greater than 1 for more than two speakers.

Listener 29 gave 6 correct Speaker 104 responses out of 6 opportunities and 0 incorrect Speaker 104 responses out of 132 opportunities. This resulted in a Bayes factor value of 13, which was the largest Bayes factor value for any listener's claimed recognition of any speaker. The constrained

magnitude of the Bayes factor is appropriate since large values cannot be justified from small samples (see discussion in [42] and [43]). Calculating a Bayes factor using (relatively) uninformative priors is a way to take account of the sample size and constrain the magnitude of the result accordingly. If we had used a larger sample size, specifically more stimuli per speaker, we could potentially have obtained higher Bayes factor values. For actual application we would advise using more stimuli per speaker if possible (and a smaller number of speakers).

Given the large variation across speakers and listeners, it appears that it would not in general be safe to attempt to predict the strength of evidence associated with any particular listener's claimed recognition of any designated speaker based on results from any other combination of speaker and listener. How well a listener performs on other speakers would not be a good predictor, and (with the potential exception of Speakers 104 and 105) how well other listeners perform on a speaker would not be a good predictor.

There has been discussion in recent years of the issue of the precision of likelihood ratios. In Appendix B we discuss this issue in relation to the results of the present study.

4.2. Bayes factor results for groups of listeners

In the original case there were several listeners, so in such a scenario calculating a Bayes factor for a group of listeners' claimed recognitions would be appropriate for quantifying the strength of evidence. If multiple earwitnesses claim to recognize an offender then that could potentially correspond to a greater strength of evidence than if a single listener claims to recognize the offender.

One has to be clear about what would constitute the evidence (note the discussion on evidence and hypotheses in §1.3 and §2.8). If there were three earwitnesses total and all three independently claimed to recognize the offender as the suspect, then that is what would constitute the evidence. If two of the earwitnesses independently claimed to recognize the offender as the suspect and the third earwitness independently claimed not to recognize the offender, then that is what would constitute the evidence. If all three earwitnesses conferred and made a consensus claim of recognition, then that is what would constitute the evidence.

It would, in principle, be possible to calculate Bayes factors for each of the examples of evidence given above. If earwitnesses conferred on the claimed recognition of the offender then the lineup design would also have them confer and the counts would be the results of the group consensus. Note that strength of evidence should not be calculated using counts pooled from multiple listeners who independently responded to the lineup; such a calculation would not relate to any possible

evidence.

For simplicity, in the present paper we only calculate a Bayes factor for the first example given above: There are three earwitnesses total and each of them independently claimed to recognize the offender as the suspect; they had no opportunity to confer. For illustrative purposes we pick Speaker 105 as the designated speaker and Listeners 06, 07, and 11 as the three listeners. We assume that each of the three listeners has independently recognized a recording of the offender as the suspect (a recording of the offender included in the lineup or in a showup), or each heard the offender speaking during the commission of the crime and independently recognized the offender as the suspect. Note the recording of the offender should not be confused with recordings of the suspect. Since each listener independently claimed to recognize the offender as the suspect, we use naïve Bayes fusion and simply multiply together the Bayes factor values that were independently calculated for each listener, hence the resulting Bayes factor value is the product of the three values from Table 3: $5 \times 8.0\dot{6} \times 11 = 433.\dot{6}$.

4.3. Implementation issues

In the present study the audio conditions were so poor that a priori we did not consider it necessary to select speakers who sounded particularly similar to each other. Under better audio conditions, it would be necessary to follow the procedures used for unfamiliar-speaker-identification lineups and select foil speakers who sound similar to the target speaker. Foil speakers could be a mixture of speakers who are familiar to the listener(s), and speakers who are unfamiliar to the listener(s). Different protocols for unfamiliar-speaker-identification lineups suggest using between 5 and 8 foils. Such speakers are not theoretically necessarily difficult to find, as relatives or friends from a close social group may sound similar to the suspect. To further distract the listener(s) and reduce the potential for bias, we would recommend including multiple other familiar speakers, and foil speakers who sound similar to the other familiar speakers. To serve this purpose, those other familiar speakers would not necessarily have to sound particularly similar to the suspect, although it would reduce the burden elsewhere if they did. To avoid introducing a bias, if the other familiar speakers do not sound particularly similar to the suspect, the number of similar-sounding foils for each of the other familiar speakers should be (at least approximately) the same as for the suspect. If the other familiar speakers do not sound similar to the suspect then it would then be appropriate for n_{-} to be based on the number of foils that sound similar to the suspect rather than the total number of speakers in the lineup.

As we mentioned in §4.1, if possible, we would advise using more stimuli per speaker (and probably a smaller number of speakers) than were used in the listening experiment reported in the

present paper. This would potentially allow for Bayes factor values that are further from 1.

A serious practical (and potentially legal) problem arises in that to run the lineup we would need to obtain recordings of the suspect under conditions that reflect those of the case. Cooperation may not be forthcoming and may not be compellable at all or not compellable in such a way as to obtain the required recording conditions. It may also in practice be difficult to obtain recordings of other speakers who sound similar to the suspect and/or who are familiar to the listeners.

The difficulty of obtaining recordings of the suspect (and recordings of other speakers) under conditions that reflect those of the case could be mitigated if it is possible to obtain high-quality audio recordings under other circumstances (e.g., during an interview). It may then be possible to process the high-quality recordings in order to obtain conditions that reflect those of the case, e.g., in the present study rather than have the famous speakers sit in the car, high-quality recordings of them were played through a loudspeaker placed in the car.

The practical difficulties associated with unfamiliar-speaker-identification lineups are such that in some jurisdictions they are not often used, [10]. The practical difficulties associated with the familiar-speaker-recognition lineup method proposed in the present paper are somewhat greater. It is therefore unlikely that the proposed method will be used frequently, but it may be worth using in a small number of important cases in which the strength of evidence associated with the familiar-speaker recognition is pivotal.

If a recording of the offender is available, as one was in the original case, then we would recommend performing a forensic voice comparison analysis instead. The process of training/optimizing and empirically validating a forensic voice comparison under conditions that reflect those of the case would face some, but not all, of the practical challenges that would be encountered in setting up a speaker lineup. A forensic voice comparison system based on relevant data quantitative measurements, and statistical models, with direct reporting of the output of the model, would be intrinsically much more resistant to cognitive bias.

5. Appendix A: Effect of not adjusting the prior to take account of the ratio of n_- to n_+

In the calculation of the numerators of the Bayes factors in the present study $n_+ = 6$ but in the calculation of the denominators $n_- = 132$. If we did not adjust the relative strength of the priors in the numerator and denominator to reflect the ratio of n_- to n_+ the results would be biased toward high Bayes factor values. For example, if c_{1+} and c_{1-} were both 0, then the calculated Bayes

factor value should be 1. If, however, a_+, b_+, a_-, and b_- were all set to 0.5 (hence m_+ and m_- would both be 1), substituting these values into Eq. 4 would give a calculated Bayes factor value of 19 (see Eq. 6a). Likewise, $c_{1+}/n_+ = 1/6$ and $c_{1-}/n_- = 22/132 = 1/6$ should result in a calculated Bayes factor value of 1, but using priors of $a_+ = b_+ = a_- = b_- = 0.5$ would result in a calculated Bayes factor value of $1.2\dot{6}$ (see Eq. 6b).

(6a)

$$\frac{\left(\frac{0+0.5}{6+1}\right)}{\left(\frac{0+0.5}{132+1}\right)} = \frac{133}{7} = 19$$

(6b)

$$\frac{\left(\frac{1+0.5}{6+1}\right)}{\left(\frac{22+0.5}{132+1}\right)} = \frac{1.5 \times 133}{7 \times 22.5} = \frac{199.5}{157.5} = 1.2\dot{6}$$

In contrast, adjusting the hyperparameter values to reflect the ratio of n_- to n_+ as described in §2.7, i.e., $a_+ = b_+ = 0.5$ (hence $m_+ = 1$) but $a_- = b_- = 11$ (hence $m_- = 22$), leads to the correct results (see Eq. 7a and Eq. 7b).

(7a)

$$\frac{\left(\frac{0+0.5}{6+1}\right)}{\left(\frac{0+11}{132+22}\right)} = \frac{0.5 \times 154}{7 \times 11} = \frac{77}{77} = 1$$

(7b)

$$\frac{\left(\frac{1+0.5}{6+1}\right)}{\left(\frac{22+11}{132+22}\right)} = \frac{1.5 \times 154}{7 \times 33} = \frac{231}{231} = 1$$

6. Appendix B: Precision

In recent years there has been discussion of the issue of precision of likelihood ratios, including in a

virtual special issue on the topic in *Science & Justice*: https://www.sciencedirect.com/journal/science-and-justice/special-issue/102F0FGVD03

For at least some subjectivist Bayesians the idea that there could be imprecision in Bayes factor values is anathema, e.g., Taroni et al. [44] and Berger & Slooten [45]. More moderate voices, e.g., Ommen et al. [46], have pointed out that when calculations result in an approximate Bayes factor value (e.g., via Monte Carlo integration), then it would be appropriate to report an estimate of the error due to the numerical technique. The purpose of reporting the estimate of the error would be to decide whether the calculation method was sufficiently precise to go ahead and use the calculated Bayes factor value, not to report a coverage interval or adjust the reported Bayes factor value according to a coverage interval (see also Taylor et al. [47] on sensitivity).

The "objective" Bayesian approach we have adopted in the present paper makes use of "uninformative" priors which leads to Bayes factor quantifications of strength of evidence that are closer to 1 than would result from maximum-likelihood estimates of likelihood ratio values, thus addressing what may be the underlying concern related to precision of quantifications of strength of evidence: avoiding overstating strength of evidence (see Vergeer et al. [42] and Morrison & Poh [43]). The approach we have adopted has a closed-form solution so numerical imprecision per se is not an issue. In this appendix, however, we do explore the question of whether the precision of the method is good enough to report and use the calculated point-value Bayes factor values. We follow the general procedure described in van den Hout & Alberink [48] for estimating the posterior distribution of the likelihood ratio. The Bayes factor is the ratio of the expected value of the posterior beta distribution of θ_+ and the expected value of the posterior beta distribution of θ_-. Each of the latter expected values are obtained by integrating out θ_+ and θ_- respectively, see Eq. 3 and Eq. 4. To estimate the posterior distribution of the likelihood ratio, rather than integrating out θ_+ and θ_-, we independently draw a Monte Carlo sample θ_+^* and a Monte Carlo sample θ_-^* from the respective posterior beta distributions, and calculate an estimate of the posterior likelihood ratio LR^* as in Eq. 8. We repeat this for 1 million pairs of independently drawn samples, and plot the histogram of the resulting $\log_{10}(LR^*)$ values.

(8)

$$LR^* = \frac{\theta_+^*}{\theta_-^*}$$

Figs. 1 through 4 show examples based on counts found in selected cells in Table 2. These counts and their corresponding Bayes factor values (from Table 3) are repeated in Table 4. Fig. 1

corresponds to the largest Bayes factor value obtained from the data, Fig. 2 to a moderate Bayes factor value, Fig. 3 to the smallest Bayes factor value greater than 1, and Fig. 4 to the smallest Bayes factor value obtained from the data. The top panels show the posterior beta distributions for the numerator (θ_+) and denominator (θ_-). The vertical dashed lines give the analytical expected values for these distributions, $\bar{\theta}_+$ and $\bar{\theta}_-$ as calculated using Eq. 3. The bottom panels show the histograms from the Monte Carlo estimate of the posterior distribution of the likelihood ratio, the posterior distribution for LR^*. The x-axis is scaled as $\log_{10}(LR^*)$, and extends to the lowest and highest values calculated from the Monte Carlo samples (or to 0 if that is a more extreme value). A solid vertical line is drawn at $\log_{10}(LR^*) = 0$, and a dashed vertical line is drawn at the Bayes factor value calculated analytically using Eq. 5. Note that the axes are rescaled in each panel and figure. This better displays the data within each figure and panel, but should be taken into account when visually comparing across figures.

- For Fig. 1 and Fig. 2, corresponding to Bayes factors of 13 and 8.1 respectively, the precision of the posterior distribution of the likelihood ratio is good: the spread of the distribution is relatively narrow. We would decide that the precision on the method is good enough for it to be used in these instances.

- For Fig. 3, corresponding to a Bayes factor of 3, the precision of the posterior distribution of the likelihood ratio is poorer: the spread of the distribution is relatively wide. This is a borderline case as to whether we would decide that the precision of the method is good enough in this instance.

- For Fig. 4, corresponding to a Bayes factor of 1/1.2, the precision of the posterior distribution of the likelihood ratio is very poor: the spread of the distribution is very wide respectively. We would decide that the precision on the method is not good enough for it to be used in this instance.

Since the Bayes factor values associated with Figs. 3 and 4 are close to 1 anyway, deciding that the precision on the method is not good enough for it to be used in these instances would be no great loss. The values are close to 1 because the Bayes factor calculations were designed to take account of concerns regarding precision of strength of evidence / concerns regarding overstating strength of evidence.

Note that in this appendix we have only explored imprecision due to the posterior beta distributions. We have not considered imprecision due to sampling variability or sensitivity to choice of priors.

Table 4. Example counts and Bayes factors corresponding to posterior likelihood ratio distributions shown in Figs. 1 through 4.

Figure	speaker	listener	c_{1+}	c_{0+}	c_{1-}	c_{0-}	Bayes factor
1	104	29	6	0	0	132	13
2	105	07	5	1	4	128	8.1
3	105	03	1	5	0	132	3
4	102	25	0	6	2	130	1/1.2

Fig. 1. Example posterior beta distributions for θ_+ and θ_- (top panel, blue and red curves respectively, see electronic version for color), and Monte Carlo estimate of the posterior distribution of the likelihood ratio LR^* (bottom panel). Example corresponds to the largest Bayes factor value obtained from the data (dashed line).

Fig. 2. Example posterior beta distributions for θ_+ and θ_- (top panel, blue and red curves respectively, see electronic version for color), and Monte Carlo estimate of the posterior distribution of the likelihood ratio LR^* (bottom panel). Example corresponds to a moderate Bayes factor value obtained from the data (dashed line).

Fig. 3. Example posterior beta distributions for θ_+ and θ_- (top panel, blue and red curves respectively, see electronic version for color), and Monte Carlo estimate of the posterior distribution of the likelihood ratio LR^* (bottom panel). Example corresponds to the smallest Bayes factor value greater than 1 obtained from the data (dashed line).

Fig. 4. Example posterior beta distributions for θ_+ and θ_- (top panel, blue and red curves respectively, see electronic version for color), and Monte Carlo estimate of the posterior distribution of the likelihood ratio LR^* (bottom panel). Example corresponds to the smallest Bayes factor value obtained from the data (dashed line).

Declaration of interest

none

Authors' contributions

Claudia Rosas: Conceptualization, Methodology, Investigation, Data Curation, Writing – Review & Editing.

Jorge Sommerhoff: Conceptualization, Methodology, Investigation, Writing – Review & Editing.

Geoffrey Stewart Morrison: Conceptualization, Methodology, Formal Analysis, Writing – Original Draft, Writing – Review & Editing, Visualization.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Novelty Statement

To our knowledge, if accepted, this would be the first published paper to propose and demonstrate a method that could potentially actually be used to assess the strength of evidence associated with a claimed familiar-speaker recognition by an earwitness.

7. References

[1] Rose P. (2002). *Forensic Speaker Identification*. London UK: Taylor and Francis.

[2] Solan L.M., Tiersma P.M. (2003). Hearing voices: Speaker identification in court. *Hastings Law Journal*, 54, 373–435.

[3] Yarmey A.D. (2007). The psychology of speaker identification and earwitness memory. In Lindsay R.C.L., Ross D.F., Read J.D., Toglia M.P. (Eds.), *The Handbook of Eyewitness Psychology, Volume II: Memory for People* (pp. 101–136). Mahwah NJ: Lawrence Erlbaum. http://dx.doi.org/10.4324/9781315805535.ch5

[4] Sherrin C. (2016). Earwitness evidence: The reliability of voice identifications. *Osgoode Hall Law Journal*, 52, 819–862. https://digitalcommons.osgoode.yorku.ca/ohlj/vol52/iss3/3

[5] Morrison G.S., Enzinger E., Zhang C. (2018). Forensic speech science. In Freckelton, I., Selby, H., (Eds.), *Expert Evidence* (Ch. 99). Sydney, Australia: Thomson Reuters.

[6] Ormerod, D. (2001). Sounds familiar? Voice identification evidence. *Criminal Law Review*, 2001(10), 595–622.

[7] Edmond G., San Roque M. (2009). Quasi-justice: Ad hoc expertise and identification evidence. *Criminal Law Journal*, 33, 8–33.

[8] Edmond G., Martire K., San Roque M. (2011). Unsound law: Issues with ('expert') voice comparison evidence. *Melbourne University Law Review*, 35, 52–112.

[9] Laub C.E., Wylie L.E., Bornstein B.H. (2013). Can the courts tell an ear from an eye? Legal approaches to voice identification evidence. *Law and Psychology Review*, 37, 119–158.

[10] Robson J. (2018). 'Lend me your ears': An analysis of how voice identification evidence is treated in four neighbouring criminal justice systems. *International Journal of Evidence & Proof*, 22, 218–238. https://doi.org/10.1177/1365712718782989

[11] Aitken C.G.G., Taroni F. (2004). *Statistics and the Evaluation of Evidence for Forensic Scientists* (2nd Ed.). Chichester UK: Wiley. http://dx.doi.org/10.1002/0470011238

[12] Lucy D. (2005). Introduction to Statistics for Forensic Scientists. Chichester UK: Wiley.

[13] Zadora G., Agnieszka M., Castro D., Aitken C.G.G. (2014). *Statistical Analysis in Forensic Science: Evidential Value of Multivariate Physicochemical Data*. Chichester UK: Wiley. http://dx.doi.org/10.1002/9781118763155

[14] Balding D.J., Steele C. (2015). *Weight-of-evidence for forensic DNA profiles* (2nd ed). Chichester, UK: Wiley. http://dx.doi.org/10.1002/9781118814512

[15] Adam C. (2016). *Forensic Evidence in Court: Evaluation and Scientific Opinion*. Chichester UK: Wiley. http://dx.doi.org/10.1002/9781119054443

[16] Buckleton J.S., Bright J.A., Taylor D (Eds.) (2016). *Forensic DNA Evidence Interpretation* (2nd Ed.). Boca Raton, FL: CRC.

[17] Robertson B., Vignaux G.A., Berger C.E.H. (2016). *Interpreting Evidence: Evaluating Forensic Science in the Courtroom* (2nd Ed.), Chichester (UK): Wiley. http://dx.doi.org/10.1002/9781118492475

[18] Yarmey A.D., Yarmey A.L, Yarmey M. (1994). Face and voice identifications in showups and lineups. *Applied Cognitive Psychology*, 8, 453–464. https://doi.org/10.1002/acp.2350080504

[19] Wells G.L., Lindsay R.C.L. (1980). On estimating the diagnosticity of eyewitness nonidentifications. *Psychological Bulletin*, 8, 776–784. https://doi.org/10.1037/0033-2909.88.3.776

[20] Wells G.L., Luss E. (1990). The diagnosticity of a lineup should not be confused with the diagnostic value of nonlineup evidence. *Journal of Applied Psychology*, 75, 511–516. http://dx.doi.org/10.1037/0021-9010.75.5.511

[21] Wells G.L., Yang Y., Smalarz L. (2015). Eyewitness identification: Bayesian information gain, base-rate effect–equivalency curves, and reasonable suspicion. *Law and Human Behavior*, 39, 99–122. http://dx.doi.org/10.1037/lhb0000125

[22] Juslin P., Olsson N., Winman A. (1996). Calibration and diagnosticity of confidence in eyewitness identification: Comments on what can be inferred from the low confidence–accuracy correlation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22, 1304–1316. http://dx.doi.org/10.1037/0278-7393.22.5.1304

[23] Rotello C.M., Chen T. (2016). ROC curve analyses of eyewitness identification decisions: An analysis of the recent debate. *Cognitive Research: Principles and Implications*, 1, article 10. https://doi.org/10.1186/s41235-016-0006-7

[24] Wixted J.T., Mickes L. (2018). Theoretical vs. empirical discriminability: the application of ROC methods to eyewitness identification. *Cognitive Research: Principles and Implications*, 3, article 9. https://doi.org/10.1186/s41235-018-0093-8

[25] Kuhn T.S. (1970). *The Structure of Scientific Revolutions* (2nd Ed.). Chicago: University of Chicago Press.

[26] Foulkes P., Barron A. (2000). Telephone speaker recognition amongst members of a close social network. *Forensic Linguistics*, 7, 180–198. http://dx.doi.org/10.1558/sll.2000.7.2.180

[27] Broeders A.P.A., van Amelsvoort A.G. (1999). Lineup construction for forensic earwitness identification: a practical approach. *Proceedings of the International Congress of Phonetic Sciences* (pp. 1373–1376). https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS1999/papers/p14_

1373.pdf

[28] de Jong-Lendle G., Nolan F., McDougall K., Hudson T. (2015). Voice lineups: A practical guide. *Proceedings of the International Congress of Phonetic Sciences*. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0598.pdf

[29] Yarmey A.D. (2003). Earwitness identification over the telephone and in field settings. *Forensic Linguistics*, 10, 65–77. http://dx.doi.org/10.1558/sll.2003.10.1.62

[30] Morrison G.S., Thompson W.C. (2017). Assessing the admissibility of a new generation of forensic voice comparison testimony. *Columbia Science and Technology Law Review*, 18, 326–434.

[31] Yarmey A.D., Yarmey A.L., Yarmey M.J., Parliament L. (2001). Commonsense beliefs and the identification of familiar voices. *Applied Cognitive Psychology*, 15, 283–299. http://dx.doi.org/10.1002/acp.702

[32] [dataset] [text redacted for blinding] (2019) Data and software for "A method for calculating the strength of evidence associated with an earwitness's claimed recognition of a familiar speaker". [text redacted for blinding]

[33] Etz A. (2018). Introduction to the concept of likelihood and its applications. *Advances in Methods and Practices in Psychological Science*, 1, 60–69. https://doi.org/10.1177%2F2515245917744314

[34] Murphy K.P. (2012). *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT.

[35] Jeffreys H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 186, 453–461. https://doi.org/10.1098%2Frspa.1946.0056

[36] Jaynes E.T. (1968). Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics*, 4, 227–291. https://doi.org/10.1109/TSSC.1968.300117

[37] Bernardo J.M. (1979). Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society Series B*, 41, 113–147.

[38] Berger J.O., Bernardo J.M., Sun D. (2009). The formal definition of reference priors. *Annals of Statistics*, 37, 905–938. http://dx.doi.org/10.1214/07-AOS587

[39] Sørensen M.H. (2012). Voice line-ups: Speakers' F0 values influence the reliability of voice recognitions. *International Journal of Speech, Language and the Law*, 19, 145–158. http://dx.doi.org/10.1558/ijsll.v19i2.145

[40] Stevenage S.V., Clarke G., McNeill A. (2012). The "other accent" effect in voice recognition. *Journal of Cognitive Psychology*, 24, 647–653. https://doi.org/10.1080/20445911.2012.675321

[41] Ladefoged P. (1978). Expectation affects identification by listening. *Language & Speech*, 21, 373–374. http://dx.doi.org/10.1177/002383097802100412

[42] Vergeer P., van Es A., de Jongh A., Alberink I., Stoel R.D. (2016). Numerical likelihood ratios outputted by LR systems are often based on extrapolation: when to stop extrapolating? *Science & Justice*, 56 482–491. http://dx.doi.org/10.1016/j.scijus.2016.06.003

[43] Morrison G.S., Poh N. (2018). Avoiding overstating the strength of forensic evidence: Shrunk likelihood ratios / Bayes factors. *Science & Justice*, 58, 200–218. http://dx.doi.org/10.1016/j.scijus.2017.12.005

[44] Taroni F., Bozza S., Biedermann A., Aitken C.G.G. (2016). Dismissal of the illusion of uncertainty in the assessment of a likelihood ratio. *Law, Probability & Risk*, 15, 1–16. http://dx.doi.org/10.1093/lpr/mgv008

[45] Berger C.E.H., Slooten K. (2016). The LR does not exist. *Science & Justice*, 56, 388–391. http://dx.doi.org/10.1016/j.scijus.2016.06.005

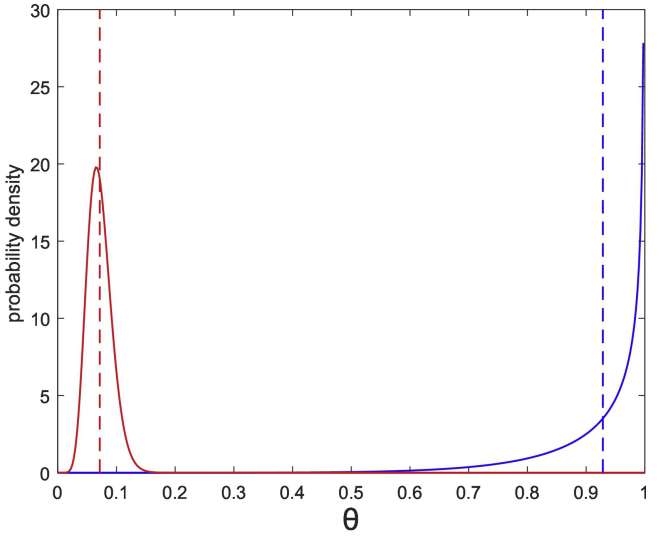
[46] Ommen D.M., Saunders C.P., Neumann C. (2016). An argument against presenting interval quantifications as a surrogate for the value of evidence. *Science & Justice*, 56, 383–387. http://dx.doi.org/10.1016/j.scijus.2016.07.001

[47] Taylor D., Hicks T., Champod C. (2016). Using sensitivity analyses in Bayesian Networks to highlight the impact of data paucity and direct future analyses: a contribution to the debate on measuring and reporting the precision of likelihood ratios. *Science & Justice*, 56, 402–410. http://dx.doi.org/10.1016/j.scijus.2016.06.010

[48] van den Hout A., Alberink I. (2016). Posterior distributions for likelihood ratios in forensic science. *Science & Justice*, 56, 397–401. http://dx.doi.org/10.1016/j.scijus.2016.06.011

Highlights

- Method to calculate strength of evidence for familiar-speaker recognition
- Bayes factors calculated using response data from speaker lineups
- Method demonstrated under forensically realistic conditions



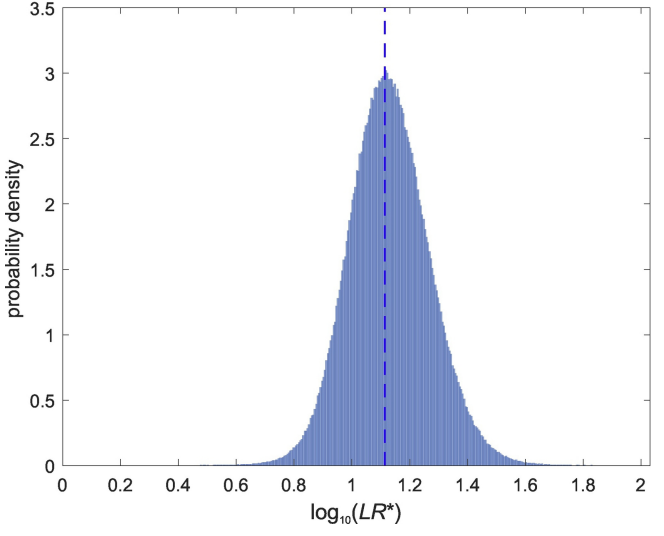
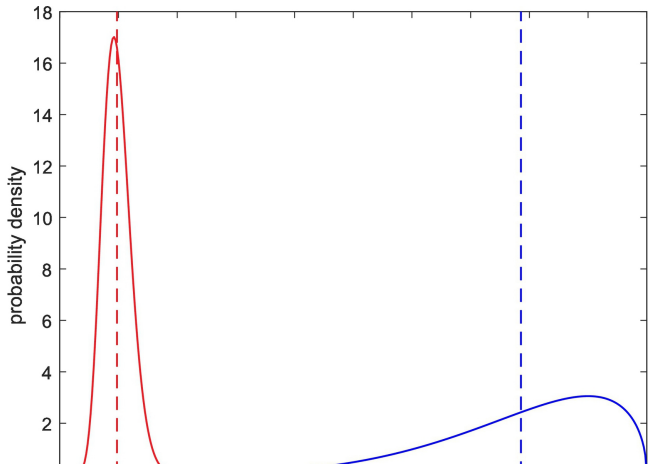


Figure 1



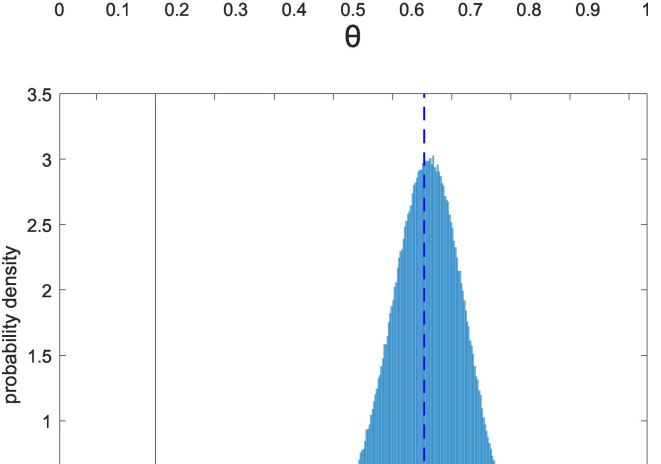
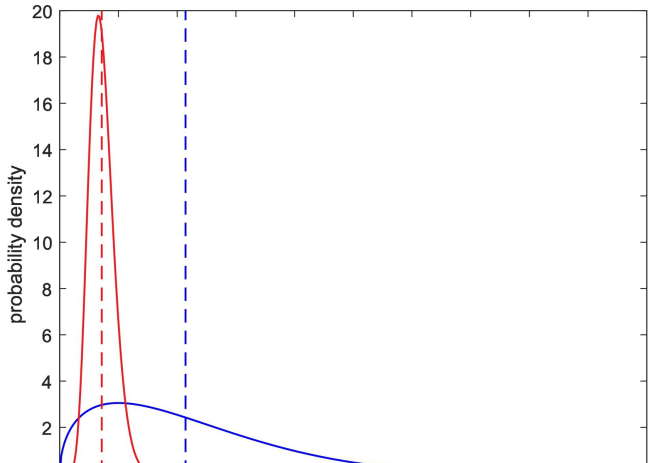


Figure 2



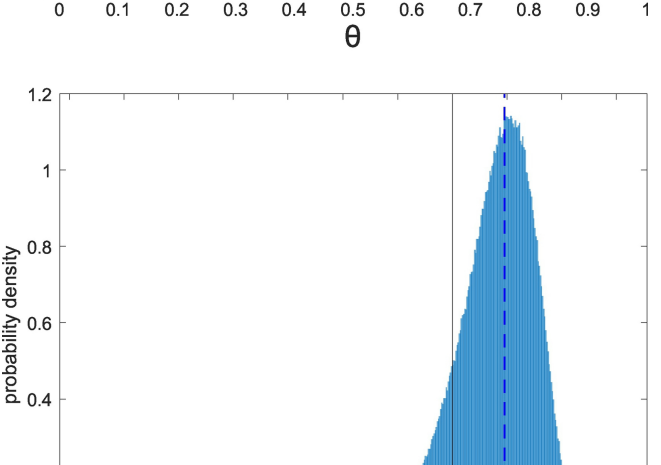
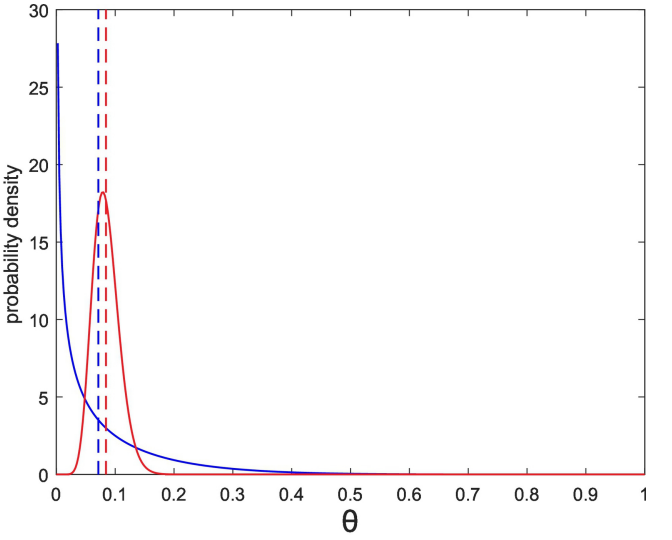


Figure 3



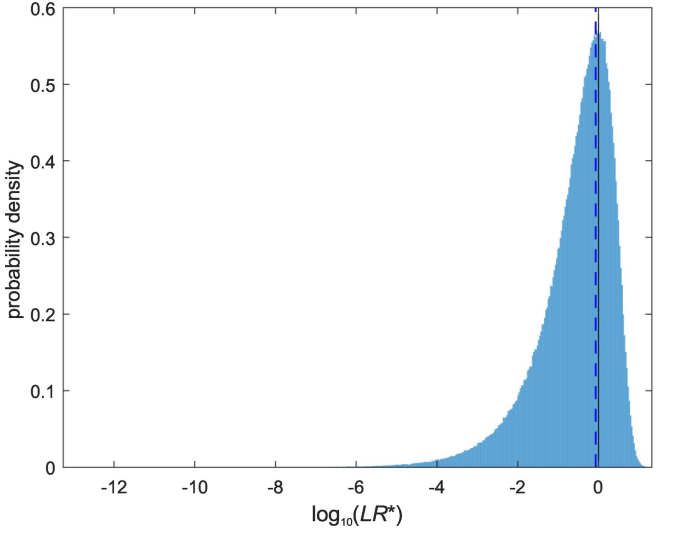


Figure 4