Introduction to Forensic Voice Comparison

Geoffrey Stewart Morrison & *Ewald Enzinger*

Abstract

This chapter provides a brief introduction to forensic voice comparison. It describes different approaches that have been used to extract information from voice recordings: auditory, spectrographic, acoustic-phonetic, and automatic approaches. It also describes different frameworks that have been used to draw inferences from such information: likelihood-ratio, posterior-probability, identification/exclusion/inconclusive, and the UK framework. In addition, the chapter describes empirical validation of forensic voice comparison systems and briefly discusses legal admissibility.

Preprint of:
Morrison, G.S., Enzinger, E. (2019). Introduction to forensic voice comparison. In Katz W.F., Assmann P.F. (Eds.) *The Routledge Handbook of Phonetics* (ch. 21, pp. 599–634). Abingdon, UK: Taylor & Francis. https://doi.org/10.4324/9780429056253-22

1 Introduction

In a court of law there is sometimes a dispute as to the identity of a speaker on an audio recording. For example, the prosecution contends that the speaker is the defendant, whereas the defense contends that the speaker is not the defendant. Other scenarios are possible, for example, the issue could be whether the questioned speaker is a kidnap victim. The court may call on a forensic practitioner to compare the recording of the speaker of questioned identity with a recording of the speaker of known identity. The task of the forensic practitioner is to help the court decide whether the voices on the two recordings were produced by the same speaker or by different speakers.

The present chapter provides an introduction to *forensic voice comparison* (aka forensic speaker comparison, forensic speaker recognition, and forensic speaker identification). A general knowledge of phonetics is assumed, but previous knowledge of forensic inference and statistics is not. It is our hope that this chapter can provide phoneticians with a working sense of the challenges faced when conducting forensic voice comparison analyses for presentation in the courtroom. We also wish to address some current controversies concerning how such forensic voice comparison may be best implemented.

Forensic voice comparison is challenging because the quality of speech recordings in casework is often poor and there is often mismatch between the speaking style in and the recording conditions of the known- and questioned-speaker recordings. Recordings may contain only a few seconds of speech, they may include background noise of various sorts (e.g., babble, ventilation system noise, vehicle noise) at varying intensities, they may have been recorded in reverberant environments, they may have been recorded using microphones that are distant from the speaker of interest (e.g., covert microphones, telephone microphones picking up speakers other than the caller), they may have been transmitted through different transmission channels (e.g., landline telephone, mobile telephone, voice over internet protocol) that distort the signal, and they may have been saved in compressed formats (e.g., MP3) that also distort the signal. Mismatches in speaking style and recording conditions can make two recordings of the same speaker appear more different than they if they were recorded under the same conditions. Mismatches in recording conditions can also mask or be mistaken for differences due to recordings actually being of different speakers. See Ajili (2017) ch 3 for a review of the effects of speaker intrinsic and speaker extrinsic factors on the performance of automatic speaker recognition systems.

The present chapter is structured as follows:

- Section 2 describes different approaches to extracting information from voice recordings.
- Section 3 describes frameworks for making inferences based on that information.
- Section 4 describes empirical validation.
- Section 5 briefly discusses legal admissibility and case law in some common-law jurisdictions.

The present chapter has some overlap with Morrison, Enzinger, & Zhang (2018). The latter covers some

topics in more detail and also covers other topics not included in the present chapter.

2 Approaches to forensic voice comparison

We use the term *approaches* to refer to broadly different ways of extracting information from speech recordings. Historical and current practice in forensic voice comparison can be described in terms of four different approaches:

- auditory
- spectrographic
- acoustic-phonetic
- automatic

This section describes the different approaches, then presents information about their popularity. Many practitioners use combinations of approaches, e.g., auditory-spectrographic or auditory-acoustic-phonetic, but for simplicity we will describe each approach separately.

We will make a distinction between modes of practice in which the conclusion as to the strength of evidence is directly based on subjective judgment (subjective mode), and in which it is based on relevant data, quantitative measurements, and statistical models (statistical mode). Auditory and spectrographic approaches are intrinsically subjective, automatic approaches intrinsically statistical, and acoustic-phonetic approaches can be practiced in either a subjective or a statistical mode.

2.1 Auditory approaches

Auditory approaches, as the name implies, are based on listening. The practitioner listens to the known- and questioned-speaker recordings in search of similarities in speech properties that they would expect if the two recordings were produced by the same speaker but not if they were produced by different speakers, and in search of differences in speech properties that they would expect if the two recordings were produced by different speakers but not if they were produced by the same speaker. Properties that a practitioner attends to may include vocabulary choice, pronunciation of particular words and phrases, segmental pronunciation, intonation patterns, stress patterns, speaking rate, and voice source properties. Practitioners who have training in auditory phonetics can transcribe and describe segmental and suprasegmental properties, including attributing putative articulatory or physiological origins of what they perceive auditorily.

Some (perhaps most) practitioners listen only to the known- and questioned-speaker recordings, and rely on their training and experience to make judgments as to whether perceived differences are more likely

to be due to same- or different-speaker origin. Some practitioners also listen to a set of *foil speakers*, i.e., speakers who sound broadly similar to the known and questioned speakers (including same sex, language spoken, and accent spoken), speaking in a similar speaking style and under similar recording conditions. One approach, known as *blind grouping* (see Cambier-Langeveld et al., 2014) is for one practitioner to prepare recordings of foil speakers, and present the known-speaker, questioned-speaker, and foil recordings to a second practitioner without telling the second practitioner the origin of each recording or how many speakers there are total. Each original recording may have been cut into a number of smaller recordings, and the recordings presented to the second practitioner are randomly labeled. The task of the second practitioner is to group the recordings by speaker. The correctness of grouping of foil speakers serves as a test of performance. Care must be taken that speaking style, linguistic content, or recording conditions for the known- and questioned-speaker recordings do not make them stand out relative to the foil recordings.

Descriptions of the auditory approach are provided in: Nolan (1997); Rose (2002); Nolan (2005); Jessen (2008); Hollien (2016); Hollien et al. (2016).

2.2 Spectrographic approaches

Practitioners of *spectrographic approaches* visually compare spectrograms of words or phrases that occur in both the known- and questioned-speaker recordings. Some protocols require the known speaker to be recorded saying the same phrases in the same speaking style as on the questioned-speaker recording. Practitioners look for similarities in the spectrograms that they would expect if the two recordings were produced by the same speaker but not if they were produced by different speakers, and for differences that they would expect if the two recordings were produced by different speakers but not if they were produced by the same speaker. Practitioners rely on their training and experience to make judgments as to whether differences they perceive between the known- and questioned-speaker spectrograms are more likely to be due to same- or different-speaker origin. Practitioners may attend to segmental and suprasegmental properties visible in the spectrograms, including fundamental frequency (f0), formants, spectral tilt, word duration, and the effect of nasal anti-resonances (American Board of Recorded Evidence, 1999). Rather than documenting multiple visually perceptible properties, some practitioners use a Gestalt approach (Poza & Begault, 2005).

As with auditory approaches, some practitioners (perhaps most) only look at spectrograms from the known- and questioned-speaker recordings, but some also look at spectrograms from foil speakers (the latter is advocated in Gruber & Poza, 1995, and Poza & Begault, 2005).

In the early 1970s, there was a debate about whether a visual only or a visual plus auditory approach was better. *Auditory-spectrographic approaches* (aka *aural-spectrographic approaches*) won out. Spectrographic / auditory-spectrographic approaches have also been called “voiceprint” or “voicegram”

approaches. The term "voiceprint" has been criticized as suggesting a false analogy with fingerprint. Whereas fingerprints are images of friction ridge patterns which are relatively stable anatomical features, spectrograms are not images of anatomical features and the acoustic properties that are graphically represented in spectrograms are subject to considerable variation due to speaker behavior and recording conditions. In the 1960s and 70s, unsubstantiated claims of near perfect performance were made by some advocates of spectrographic approaches, and the term "voiceprint" fell into disrepute.

Descriptions of the spectrographic approach are provided in: Kersta (1962); Tosi (1979); and National Research Council (1979). There has been substantial controversy surrounding the use of the spectrographic approach. Reviews of the controversy around its use and admissibility are included in: Gruber & Poza (1995); Solan & Tiersma (2003); Meuwly (2003a,b); Morrison (2014); and Morrison & Thompson (2017).

2.3 Acoustic-phonetic approaches

Practitioners of *acoustic-phonetic approaches* may examine many of the same acoustic properties that practitioners of auditory and spectrographic approaches examine via auditory or visual perception, but practitioners of acoustic-phonetic approaches make quantitative measurements of those acoustic properties. The most widely used acoustic-phonetic properties are fundamental frequency and formant frequencies, but quantitative measurements can also be made of voice onset time (VOT), fricative spectra, nasal spectra, voice source properties, speaking rate, etc. (Gold & French, 2011; French & Stevens, 2013). A common approach involves first finding and marking the beginning and end of realizations of particular phonemes or of major allophones of particular phonemes, for example, all tokens of /i/ or all tokens of /aɪ/ not adjacent to a nasal, lateral, rhotic, or labiovelar. Human supervised formant measurements are then made using the same sort of signal-processing algorithms and procedures as are used in acoustic-phonetic research in general, e.g., linear predictive coding (LPC) plus a peak picking algorithm with the optimal number of LPC coefficients selected by the human supervisor. Measurements may be made at multiple points in time during the vowel to capture information about formant trajectories.

Since questioned-speaker recordings are often telephone recordings and traditional telephone systems have bandpasses of around 300 Hz – 3.4 kHz, first formant (F1) frequencies close to 300 Hz (e.g., in [i] and [u]) are often distorted, and high frequency spectral information in bursts and fricatives (e.g., in [t^h] and [s]) is often missing. Fundamental frequency is below the bandpass, but can be recovered from harmonic spacing. Practitioners of acoustic-phonetic approaches have to take such transmission-channel effects into account, especially when there is a mismatch in recording channel between the known- and questioned-speaker recordings. One solution is to not use certain measurements, such as measurements of F1 in realizations of vowels with intrinsically low F1; however, even in the middle of the bandpass, the codecs used in mobile telephone systems can distort formants. Zhang et al. (2013) reviewed previous research on the effects of telephone transmission on formants, and tested the effects of landline and

mobile telephone transmission on the performance of forensic voice comparison systems.

Once they have made their measurements, some practitioners make tables or plots of the values, and use their training and experience to subjectively assess strength of evidence via examination of those tables or plots. For example, first versus second formant (F1-F2) values could be plotted for realizations of a particular vowel phoneme in both the known- and questioned-speaker recording, and the visual degree of overlap considered. Measurements made on recordings of foil speakers may also be plotted.

Other practitioners use their measurements as input to statistical models that calculate quantifications of strength of evidence. Some practitioners directly report the output of the statistical model as a strength of evidence statement, but others use it as input to a subjective judgment process, which may also include consideration of the results of other analyses such as an auditory analysis.

Descriptions of acoustic-phonetic approaches (non-statistical and statistical) are provided in: Nolan (1997); Rose (2002); Hollein (2002); Nolan (2005); Rose (2006); Jessen (2008); Rose (2013); Drygajlo et al. (2015); Rose (2017).

2.4 Automatic approaches

Human supervised *automatic approaches* evolved from signal-processing engineering, and in particular speech processing. Automatic speaker recognition techniques developed for non-forensic applications have been adapted for forensic application. For most security applications the system has to be fully automatic and make a decision. For example, the system must decide to grant or deny a caller access to bank account information without intervention from a human member of staff. Also, the bank client will initially have been cooperative in enrolling sample recordings of their voice, and the bank client can also be asked to cooperate by calling from a quiet location and be asked to say particular words and phrases. In contrast, in forensic application:

- the questioned speaker is generally not cooperative,
- the questioned speaker is not trying to be identified and they may not even be aware that they are being recorded,
- the recording conditions and speaking styles are much more variable and the quality of the recordings often much poorer, and
- the output of the system is not a binary decision but a quantification of strength of evidence.

Appropriate adaptation of automatic speaker recognition techniques to forensic problems is non-trivial. It requires human supervised systems in which the practitioner carefully selects relevant data for training and testing so that the output of the system is a meaningful quantification of strength of evidence for the case. Inappropriate use of automatic systems is garbage in garbage out (GIGO).

Traditional automatic approaches do not explicitly exploit acoustic-phonetic, segmental, or suprasegmental information. Instead, they usually make spectral measurements at regular intervals, e.g., once every 10 ms within a 20 ms wide window. Such measurements are usually made throughout the recorded speech of the speaker of interest, and the results are pooled irrespective of whether they originated from vowels or consonants or realizations of particular phonemes. The most commonly made measurements are mel frequency cepstral coefficients (*MFCCs*, see Davis & Mermelstein, 1980). At each point in time, the MFCCs characterize the shape of the spectral envelope using a vector of around 14 numbers. *Deltas*, the rate of change of MFCC values over time, and *double deltas*, the rate of change of delta values over time, are usually appended to produce a vector of around 42 numbers (see Furui, 1986). The measurements made by automatic systems, and the derivatives of those measurements, are known as *features*.

The boundary between acoustic-phonetic and automatic approaches is, however, fuzzy. Some automatic approaches use phone recognizers from automatic speech recognition systems to model information related to specific phones or phone classes (e.g., vowels, fricatives, nasal consonants). See Ajili et al. (2016) for an exploration of the effect of excluding particular phone classes on the performance of an automatic forensic voice comparison system. Some systems automatically find voiced segments then automatically measure f0 and formant frequency values at regular intervals throughout those voiced segments. The latter are called long-term formant (LTF) values (see for example Jessen et al., 2014). Some practitioners of acoustic-phonetic approaches mark realizations of particular phonemes, but then make MFCC measurements within those realizations. Using around 14 MFCC values provides more information about the spectrum than f0 plus two or three formant values.

The measurements made in automatic approaches are invariably used as input to statistical models. The statistical modelling approaches used in forensic voice comparison (and automatic speaker recognition more broadly) have evolved over the last 20 years. An approach known as *GMM-UBM* (Gaussian mixture model - universal background model) was introduced around 2000 (Reynolds et al., 2000), and another approach known as *i-vector - PLDA* (identity vector - probabilistic linear discriminant analysis) was introduced around 2010 (Dehak et al., 2011; Prince & Elder, 2007). Current state of the art in automatic speaker recognition research makes use of *DNNs* (deep neural networks; e.g., Richardson et al., 2015), and these are beginning to be adopted for forensic application. GMM-UBM, i-vector PLDA, and DNN-based systems can be considered systems that output *scores*. Scores are similar to likelihood ratios (see Section 3.1) in that they take account of similarity and typicality, but their absolute values are not interpretable. This is not a problem in applications in which a decision is made by comparing a score value to a threshold but is a problem when the purpose is to provide a strength of evidence statement to the court. *Calibration* can be used to convert scores to interpretable likelihood ratios. It was first used in forensic voice comparison around 2007. A standard statistical modelling approach for calibration is *logistic regression* (see: Pigeon et al., 2000; González-Rodríguez et al., 2007; Morrison, 2013).

A great deal of effort in automatic speaker recognition has focused on developing statistical techniques

for dealing with mismatches in speaking styles and recording conditions between known- and questioned-speaker recordings (the statistical techniques are known as *mismatch compensation*, aka channel or session compensation). These techniques include *cepstral mean subtraction* (CMS, see Furui, 1981), *cepstral mean and variance normalization* (CMVN, see Tibrewala & Hermansky, 1997), *feature warping* (Pelecanos & Sridharan, 2001), and *linear discriminant analysis* (LDA). The first three are alternatives applied to features and the latter is usually applied to i-vectors. In automatic speaker recognition, LDA actually usually refers to the use of canonical linear discriminant functions for dimension reduction (see Klecka, 1980) prior to using a different probabilistic classification model.

Some practitioners directly report the output of the statistical model as a strength of evidence statement, but others use it as input to a subjective judgment process, which may also include consideration of the results of other analyses such as auditory and acoustic-phonetic analyses.

Descriptions of automatic approaches are provided in: Ramos Castro (2007); Becker (2012); Enzinger (2016); Marks (2017). For a review of approaches to automatic speaker recognition in general, see Hansen & Hasan (2015).

2.5 Popularity of different approaches

Gold & French (2011) published the results of a survey of practitioners working in a mixture of private, university, and law-enforcement or other government laboratories. In the reported results from 35 respondents:

- 2 (6%) used an auditory only approach.
- Spectrographic approaches were not mentioned.
- 25 (71%) used an auditory-acoustic-phonetic approach.
- 1 (3%) used an acoustic-phonetic-only approach.
- 7 (20%) used a human-supervised automatic approach.

In 2016, INTERPOL published the results of a survey of speaker recognition capabilities of law-enforcement agencies in member countries (Morrison, Sahito, et al., 2016). 44 respondents stated that their agency had speaker recognition capabilities. Of these, many reported using more than one approach, hence the summary statistics below add up to more than 44.

- 15 (25%) used an auditory approach.
- 21 (34%) used a spectrographic or auditory-spectrographic approach.
- 25 (41%) used an auditory-acoustic-phonetic (subjective) approach.

- 15 (25%) used an acoustic-phonetic (statistical) approach.
- 20 (33%) used a human-supervised automatic approach.
- 9 (15%) used a fully-automatic approach. (Assumed to be for investigative rather than forensic application.)

In the results of both surveys, (auditory-)acoustic-phonetic approaches were the most popular, with human-supervised automatic approaches second or a close third. In the results of the INTERPOL survey, (auditory-)spectrographic approaches were also popular. Auditory-only approaches were the least popular.

The two surveys were separated in time but also solicited responses from different groups, hence one cannot conclude that the differences between them are due to changes in practice.

3 Frameworks for evaluation of forensic evidence

In contrast to the term *approaches*, which we use to refer to different ways of extracting information from speech recordings, we use the term *frameworks* to refer to different ways of evaluating strength of evidence based on that information. Frameworks therefore refer to ways of reasoning or ways of drawing inferences. Historical and current practice can be described in terms of a number of different frameworks, including:

- likelihood-ratio
- posterior-probability
- identification / exclusion / inconclusive
- the UK framework.

We describe each of these frameworks below, and then give information about their popularity.

Frameworks are mutually exclusive of each other. Most frameworks can be used in a subjective or a statistical mode. In some frameworks strength of evidence can be expressed verbally or as a numeric value. To some extent, frameworks are independent of approaches, e.g., one could use either an acoustic-phonetic-statistical or an automatic approach in combination with either a posterior-probability or a likelihood-ratio framework. In practice, however, certain combinations are more common than others, e.g., a spectrographic approach combined with a verbal posterior-probability framework, an auditory-acoustic-phonetic approach combined with the UK framework, and an automatic approach combined with a numeric likelihood-ratio framework.

3.1 Likelihood-ratio framework

The likelihood ratio framework is considered the logically correct framework for the evaluation of forensic evidence by many forensic statisticians, forensic scientists, and legal scholars; see for example: Aitken et al. (2011); Willis et al. (2015); Drygajlo et al. (2015); Morrison, Kaye, et al. (2017). General introductions to the likelihood ratio framework include: Aitken et al. (2010); Robertson et al. (2016); Balding & Steele (2015) ch. 1–3 and 11. Introductions in the context of forensic voice comparison include: Rose (2002); Morrison & Thompson (2017); Morrison, Enzinger, & Zhang (2018).

3.1.1 Similarity is not enough

Just considering the degree of similarity of the voices on the known- and questioned-speaker recordings is not enough to quantify strength of evidence. Imagine that we measured the mean fundamental frequency of the voices on two recordings and found that they differed by 5 Hz. Would that indicate that the two recordings are of the same speaker? Would it indicate that it is highly probable that the two recordings are of the same speaker? You will probably answer "no" or "it depends". The two recorded voices are very similar on this metric, there is only a 5 Hz difference between them, but we have to consider whether that 5 Hz difference is more likely to occur because it really is the same speaker or more likely to occur by chance. How do we assess this? Our discussion below focuses on answering this question using *relevant data*, *quantitative measurements*, and *statistical models*.

3.1.2 Histogram models

Imagine that we have multiple recordings of the known speaker (N_k recordings total) and we measure the mean fundamental frequency in each recording. We will designate each of these values an x_{k_i} value, $i \in 1 \ldots N_k$. Let's begin with a simple statistical model based on the proportion of the x_{k_i} values that fall into 10 Hz wide ranges, e.g., what proportion of the x_{k_i} values are greater than 90 Hz and less than or equal to 100 Hz, what proportion are greater than 100 Hz and less than or equal to 110 Hz, etc. We can represent the results as a histogram as in Figure 1 (top panel darker rectangles).[1] We will draw the histogram such that the area of each rectangle represents the proportion of measurements falling within the range of x values which it covers. Since each rectangle represents a proportion of the whole, the sum of the areas of all of the rectangles must be 1. We can now use the histogram as a statistical model. We use the proportions as estimates of probability. We can use the model to estimate the probability that the mean fundamental frequency of a recording would have a value within a specified 10 Hz wide range

[1] All panels in Figure 1 were based on the same simulated data consisting of N_k=100 recordings sampled from the simulated known speaker and N_r=1000 recordings sampled from the simulated relevant population. The data were generated for illustrative purposes only and are not intended to accurately reflect real f0 distributions.

if it were produced by the known speaker. Imagine that the mean fundamental frequency of the voice on the questioned-speaker recording, x_q, is 99 Hz. This falls in the range of greater than 90 Hz and less than or equal to 100 Hz. The area of the rectangle corresponding to this range, and hence the estimated probability of a value falling in this range if it were produced by the known speaker is 0.0360. This gives us a quantification of the *similarity* between the voice on the questioned-speaker recording and the known speaker's voice.

<Figure 1 about here>

We also need to consider the probability of getting an x_q value of 99 Hz if the questioned speaker were not the known speaker but some other speaker selected at random from the relevant population. We will discuss the concept of relevant population in Section 3.1.6 below. For now let us assume that the relevant population is adult males. We obtain recordings of a large number (N_r) of adult male speakers. This is a sample of the relevant population. For the recording of each speaker we measure the mean fundamental frequency, $x_{r_j},\ j \in 1 \ldots N_r$. We then construct a histogram for these data in the same way as we did for the data from the known speaker. This is shown in Figure 1 (top panel lighter rectangles). We use the second histogram as our statistical model to estimate the probability that the x_q value would fall within the range greater than 90 Hz and less than or equal to 100 Hz if it came from a speaker selected at random from the relevant population. In this example, that value is 0.0066. This provides a quantification of the *typicality* of the voice on the questioned-speaker recording with respect to the relevant population.

We are now in a position to answer a question which has two parts:

- What is the probability of getting the measured property of the voice on the questioned-speaker recording if it came from the known speaker? (What is the degree of *similarity*?)

versus

- What is the probability of getting the measured property of the voice on the questioned-speaker recording if it came not from the known speaker but from a speaker selected at random from the relevant population? (What is the degree of *typicality*?)

If we divide the answer to the first part (the similarity part) by the answer to the second part (the typicality part), we get the ratio of the two, and we can say that we estimate that the probability of getting the measured property of the voice on the questioned-speaker recording is 0.0360 / 0.0066 = 5.45 times higher if it came from the known speaker than if it came from some other speaker selected at random from the relevant population.

We can rephrase the two questions above as hypotheses:

- The measured property of the voice on the questioned-speaker recording (the evidence) came from the known speaker.

versus

- The measured property of the voice on the questioned-speaker recording (the evidence) came not from the known speaker but from a speaker selected at random from the relevant population.

For brevity, we will refer to these as the *same-speaker hypothesis* and *different-speaker hypothesis* respectively, and we will use the term *evidence* to refer to the measurement (or measurements) made on the questioned-speaker recording.

Note that using the model described above, we would have gotten the same result had x_q been any value between 90 and 100 Hz, but if the mean of the x_{k_i} were 90 Hz then perhaps the strength of evidence should be lower if x_q is toward the top of the 90–100 Hz range than if it is toward the bottom (further from the known-speaker mean rather than closer). We could address this by making the ranges of the rectangles in our histograms narrower. For example, in Figure 1, the middle and bottom panels have rectangles of widths 5 Hz and 1 Hz respectively. Using these histograms, the corresponding estimates of the probability of getting the measured property of the voice on the questioned-speaker recording if it came from the known speaker versus if it came from some other speaker selected at random from the relevant population are 0.0300⁄0.0084 = 3.57 and 0.0300⁄0.0100 = 3.00 respectively.

If we continue to make the widths of the rectangles narrower and narrower, however, we will eventually run into a problem. Because we are using proportions it does not matter exactly how many known-speaker recordings and how many relevant-population sample recordings we have, and we do not have to have the same number of each, but we do have to have a sufficient number of measurements in each rectangle for the proportions to be reasonable estimates of probability. Assuming we have a limited number of recordings that are a sample of the known speaker's speech, and a limited number of recordings which are a sample of the speech of speakers from the relevant population, as we make the rectangles narrower fewer and fewer x_{k_i} and x_{r_j} values will fall within each rectangle, and the quality of the estimate of the probability for each rectangle will deteriorate. At the extreme, most rectangles would have a zero count and some would have a count of one. The problem is already apparent in the bottom panel of Figure 1.

3.1.3 Parametric statistical models (Gaussian distributions)

A solution to the problem described at the end of the previous subsection is to use a parametric model. If we are willing to assume that the data have a Gaussian distribution (a normal distribution), then we can get relatively good parameter estimates (estimates of the mean and variance) using comparatively little data compared to the amount of data we would need to make a high resolution non-parametric histogram. We calculate the sample means $\hat{\mu}_k$ and $\hat{\mu}_r$ and sample variances $\hat{\sigma}_k^2$ and $\hat{\sigma}_r^2$ from the x_{k_i} and from the x_{r_j} values. This gives us the statistics (the parameter estimates) necessary to define the two Gaussian

distribution models plotted in Figure 2: $y_k = f(x|\hat{\mu}_k, \hat{\sigma}_k^2)$ and $y_r = f(x|\hat{\mu}_r, \hat{\sigma}_r^2)$, where $f(x|\mu, \sigma^2)$ is the Gaussian probability density function (these were trained on the same simulated data used to train the histograms in Figure 1). Note that just as the total area of a histogram is 1, the area under the curve for each Gaussian distribution is 1. If we calculate the area under the curve within a range of values, e.g., greater than 90 Hz and less than or equal to 100 Hz, or greater than 98.5 Hz and less than or equal to 99.5 Hz, then we have a probability estimate of a value falling within that range. But if we pick an exact value, e.g., 99 Hz, then the width of that range is zero and the area is therefore zero. The *y* axis value for the curve at an exact value is not zero, but the *y* value does not represent probability – it represents a quantity known as *probability density* or *likelihood*. At the value of the evidence, x_q, we assess the likelihood of the known-speaker model, $y_{q,k} = f\left(x_q\middle|\hat{\mu}_k, \hat{\sigma}_k^2\right)$ (this quantifies similarity), and the likelihood of the relevant-population model, $y_{q,r} = f\left(x_q\middle|\hat{\mu}_r, \hat{\sigma}_r^2\right)$ (this quantifies typicality), and we divide the former by the latter. The result is a *likelihood ratio*, which is a quantification of the strength of evidence. In this case we estimate that the likelihood of the evidence is $y_{q,k}/y_{q,r} = 0.0279/0.0089 = 3.12$ times higher if it came from the known speaker than if it came from some other speaker selected at random from the relevant population.[2]

<Figure 2 about here>

<Figure 3 about here>

Note that if we keep the degree of similarity the same, but reduce the degree of typicality, then the value of the likelihood ratio increases: see Figure 3 top left panel, in which $\hat{\mu}_r$ has been increased by 10 Hz relative to its value in Figure 2. The value of the likelihood ratio is 6.68.

Vice versa, if typicality increases the value of the likelihood ratio decreases: see Figure 3 top right panel, in which $\hat{\mu}_r$ has been decreased by 10 Hz relative to its value in Figure 2. The value of the likelihood ratio is 1.88.

Also, if we keep the degree of typicality the same, but the degree of similarity increases the value of the likelihood ratio increases: see Figure 3 bottom left panel, in which $\hat{\mu}_k$ has been increased by 10 Hz relative to its value in Figure 2. The value of the likelihood ratio is 4.12. In contrast, if similarity decreases the value of the likelihood ratio decreases: see Figure 3 bottom right panel, in which $\hat{\mu}_k$ has been decreased by 10 Hz relative to its value in Figure 2. The value of the likelihood ratio is 0.99.[3] The value

[2] Technically when describing a likelihood ratio one should talk about *probability of evidence given hypotheses* if the data are discrete, but *likelihood of hypotheses given evidence* if the data are continuous. This, however, is confusing for non-statisticians, and our concern in a forensic or legal context is to avoid inducing the prosecutor's fallacy (Thompson & Schumann, 1987). We may therefore use phrases such as "likelihood of evidence given hypotheses" or "probability of evidence given hypotheses" rather than the technically correct "likelihood of hypotheses given evidence".

[3] If increased similarity corresponds with increased typicality, however, then in some instances the value of the likelihood ratio could decrease as similarity increases.

of the likelihood ratio would also change if the value of x_q changed.

Statistical models are not restricted to univariate measurements and can be applied to multivariate measurements. In addition, statistical models are not restricted to Gaussian distributions, and much more complex distributions can be fitted. In both acoustic-phonetic statistical and automatic approaches, the data are usually multidimensional and have complex distributions. Multivariate Gaussian mixture models are used in both GMM-UBM and i-vector - PLDA.

3.1.4 What does a likelihood ratio mean?

Likelihood ratios can have values that are > 1 or < 1. If the value of the likelihood ratio is > 1, then the evidence is more probable if the same-speaker hypothesis were true than if the different-speaker hypothesis were true. If the value of the likelihood ratio is < 1, then the evidence is more probable if the different-speaker hypothesis were true than if the same-speaker hypothesis were true. But, importantly, what matters is not just whether a likelihood ratio is > or < 1, but how far away it is from 1. The further the value from 1 the greater the strength of evidence.

What does the value of a likelihood ratio mean? From a normative perspective it is the amount by which the trier of fact (the judge or the jury depending on the legal system) should change their belief regarding the relative probabilities of the same- and different-speaker hypotheses.[4]

Before the trier of fact is presented with the forensic voice comparison testimony, they have some belief as to the relative probability that the speaker on the questioned recording will be the defendant versus that the speaker on the questioned recording will be some other speaker. A simplistic example assumes a crime committed on an island of 101 inhabitants. One of the islanders is the defendant and the trier of fact assumes that innocent until proven guilty implies that before considering any evidence the defendant is no more or less likely to be the questioned speaker than any other inhabitant of the island. The prior probability that the defendant is the questioned speaker is therefore 1/101 and the prior probability that someone else is the questioned speaker is 100/101 (1/101 for each inhabitant multiplied by the 100 inhabitants other than the defendant). The ratio of these is (1/101)/(100/101) = 1/100. The ratio of the prior probabilities is called the *prior odds*.

But the trier of fact may have already heard other evidence, and immediately prior to hearing the likelihood ratio from the forensic voice comparison their prior odds may no longer be 1/100. For example, if it is apparent that the voice on the questioned-speaker recording is male, and the defendant is male, and the trier of fact is told that 50% of the other inhabitants of the island are male, then the trier of fact's

[4] This discussion is provided to explain the meaning of a likelihood ratio. It is not intended to instruct triers of fact as to how to reason on legal matters, nor to imply that triers of fact must assign numeric values to evidence which is not the result of a forensic analysis.

prior odds before hearing the forensic voice comparison testimony might be 1/50.

Irrespective of the actual value of the trier of fact's prior odds, normatively the trier of fact should update their belief according to the following formula (the *odds form of Bayes' Theorem*):

(1)

$$\text{prior odds} \times \text{likelihood ratio} = \text{posterior odds}$$

$$\frac{p(H_s)}{p(H_d)} \times \frac{p(E|H_s)}{p(E|H_d)} = \frac{p(H_s|E)}{p(H_d|E)}$$

$$\frac{1}{100} \times 5 = \frac{5}{100} = \frac{1}{20}$$

$$\frac{1}{50} \times 5 = \frac{5}{50} = \frac{1}{10}$$

In which H_s stands for the same-speaker hypothesis, H_d for the different-speaker hypothesis, and E for the evidence. $p(E|H)$ reads as probability of evidence given hypothesis, and $p(H|E)$ reads as probability of hypothesis given evidence. After the trier of fact has heard the strength of evidence expressed as a likelihood ratio, the *posterior odds* are what their belief should be as to the relative probabilities that the defendant is the questioned speaker versus that someone else on the island is the questioned speaker.

Note that in the two examples at the bottom of Equation 1, the differences in the posterior odds are due to differences in the prior odds. The evidence is the same, the strength of evidence is the same, and the likelihood ratio calculated by the forensic practitioner is the same. Both examples use a likelihood ratio of 5. What the trier of fact should do with the likelihood ratio is the same, irrespective of what their prior odds are. A likelihood ratio has an unambiguous meaning. In the context of forensic voice comparison, it is the amount by which (in light of the evidence) the trier of fact should multiply their prior odds in order to update their belief about the relative probabilities of the same- versus the different-speaker hypotheses being true. With suitable changes in wording, an unambiguous definition of a likelihood ratio can be provided addressing strength of evidence questions in other branches of forensic science.

If the value of the likelihood ratio is > 1, the value of the posterior odds will be more than the value of the prior odds, and if the value of the likelihood ratio is < 1, the value of the posterior odds will be less than the value of the prior odds. In this sense, the likelihood ratio framework is symmetrical, it can lead to higher belief in the probability of the same-speaker hypothesis being true over the different-speaker hypothesis being true, or *vice versa*. Likelihood ratios < 1 are crammed into the range 0 to 1, whereas likelihood ratios > 1 are in the range 1 to infinity. A value < 1 can be inverted along with inversion of

the hypotheses, e.g., $p(E|H_s)/p(E|H_d) = 0.001$ is equivalent to $p(E|H_d)/p(E|H_s) = 1000$, and, rather than use a fraction, it is easier to say the probability of the evidence is 1000 times greater if the different-speaker hypothesis were true than if the same-speaker hypothesis were true. For mathematical convenience, log likelihood ratios are often used, e.g., $\log_{10}(0.001) = -3$ and $\log_{10}(1000) = +3$. Likelihood ratios less than 1 convert to log likelihood ratios in the range minus infinity to 0, and likelihood ratios greater than 1 convert to log likelihood ratios in the range 0 to plus infinity. The log-odds version of Bayes' Theorem is additive rather than multiplicative:

(2)

$$\text{log prior odds} + \text{ log likelihood ratio } = \text{ log posterior odds}$$

$$\log\left(\frac{p(H_s)}{p(H_d)}\right) + \log\left(\frac{p(E|H_s)}{p(E|H_d)}\right) = \log\left(\frac{p(H_s|E)}{p(H_d|E)}\right)$$

$$\log\big(p(H_s)\big) - \log\big(p(H_d)\big) + \log\big(p(E|H_s)\big) - \log\big(p(E|H_d)\big) = \log\big(p(H_s|E)\big) - \log\big(p(H_d|E)\big)$$

3.1.5 Subjective likelihood ratios and verbal expression of likelihood ratios

Based on their training and experience, it should be intuitive for a phonetician that an f0 of around 120 Hz is typical for an adult male. Thus, if the voices on the known- and questioned-speaker recordings both have values close to 120 Hz, this does not constitute strong evidence in support of the hypothesis that they were both produced by the same speaker rather than by different speakers. In contrast, an f0 of 70 Hz is atypical. Thus if the voices on the known- and questioned-speaker recordings both have values close to 70 Hz, this does constitute strong evidence in support of the hypothesis that they were both produced by the same speaker rather than by different speakers. Without obtaining an explicit sample of the relevant population and without using a statistical model, it is therefore possible for a phonetician to subjectively assign values to a likelihood ratio. The phonetician could give subjective numeric estimates or could give a verbal expression, e.g., the evidence is much more probable if the same-speaker hypothesis were true than if the different speaker hypothesis were true. Such statements would be consistent with the logic of the likelihood ratio framework.

The 2015 ENFSI guideline on evaluative reporting (Willis et al., 2015) includes examples of verbal expressions of likelihood ratios, such as shown in Table 1. Each expression is associated with a range of numeric likelihood ratio values. Terms such as "much more probable" and "far more probable" are, however, ambiguous. They may be interpreted differently by different people, and even differently by the same person in different contexts.[5]

[5] For more detailed criticism of verbal expressions and ordinal scales, see Marquis et al (2016) and Morrison & Enzinger

<Table 1 about here>

3.1.6 Relevant population

An important issue for calculating a likelihood ratio is: What is the *relevant population*? On listening to the questioned-speaker recording certain things are usually (but not always) obvious, including whether the speaker is male or female, what language they are speaking, and broadly what accent of that language they are speaking. If these are likely to be salient to the trier of fact and unlikely to be disputed by the prosecution or defense, then they can be used for defining the relevant population that the forensic practitioner will adopt. It is important that the forensic practitioner clearly communicate what relevant population they have adopted so that the judge at an admissibility hearing and/or the trier of fact at trial can decide whether it is appropriate, and so that the judge and/or trier of fact can understand what the calculated value of the likelihood ratio means – one cannot understand the answer if one does not understand the question.[6]

In addition, it is important to consider whether the sample is sufficiently representative of the specified relevant population. Part of this consideration is whether the sample is sufficiently large. Also to be considered is whether the sample is biased or is simply of some population other than the specified relevant population. For example, imagine that the questioned speaker and known speaker are adult males with f0 values of approximately 120 Hz, and that the specified relevant population is adult males, but the sample used is actually a sample of adult females. The f0 for the known and questioned speakers would be relatively atypical with respect to the distribution of f0 values in the sample of female speakers. The calculated likelihood ratio value would therefore be large, but this value would be misleading because the question it actually answers involves typicality with respect to the wrong population. The actual question answered would be nonsensical: What is the probability of getting the f0 of the male on the questioned-speaker recording if it were produced by the male known speaker, versus if it were produced by a female speaker?[7] What the forensic practitioner uses as a sample of the relevant population must be clearly communicated to the judge and/or trier of fact so that ultimately the judge and/or trier of fact can decide whether the sample is sufficiently representative of the relevant population.

(2016). For a study of lay people's perception of various verbal and numeric expressions of strength of evidence, see Thompson et al. (2018).

[6] Selection of the relevant population in the context of forensic voice comparison is discussed in Morrison, Ochoa, & Thiruvaran (2012); Gold & Hughes (2014); Hughes & Foulkes (2015); Hicks, Biedermann, et al. (2015, 2017); Morrison, Enzinger, & Zhang (2016, 2017); Hughes & Rhodes (2018).

[7] We note, however, that it is not always obvious whether the questioned speaker is male or female and since the different-speaker hypothesis is that the questioned speaker is not the known speaker, the sex of the known speaker is not relevant for defining the relevant population. In such cases the relevant population could include speakers of both sexes, e.g., males plus females with low pitched voices or females plus males with high pitched voices.

3.2 Posterior-probability framework

Some practitioners express their conclusions as to strength of evidence as posterior probabilities. These could be generated using statistical models or could be subjectively assigned, and they could be expressed numerically or verbally. For example: "There is a 95% probability that the known speaker is the speaker on the questioned-speaker recording." Expressions of certainty are also posterior probability expressions, e.g., "I am 95% certain that it is the same speaker." The American Board of Recorded Evidence protocols for spectrographic approaches require the use of verbal expressions of posterior probability: "identification", "probable identification", "possible identification", "inconclusive", "possible exclusion", "probable exclusion", "exclusion" (American Board of Recorded Evidence, 1999).

The problem with expressing strength of evidence as a posterior probability is that logically in order to calculate a posterior probability one must consider two things: the likelihood ratio and the prior probability, see Section 3.1.4 above.[8] The forensic practitioner must therefore either use some arbitrary prior, or must assign a prior based on the other evidence in the case that the trier of fact has already heard.

Unless the trier of fact tells the forensic practitioner specific priors to use, the forensic practitioner cannot calculate the appropriate posterior probability. Given current legal practice around the world, it is extremely unlikely that the trier of fact would provide specific priors, and in some jurisdictions this is clearly impossible.

The task of the forensic practitioner is to assess the strength of evidence of the particular materials they have been asked to analyze, independent of any other evidence in the case. It is the task of the trier of fact to consider and combine all the evidence. It would be inappropriate for the forensic practitioner to consider the other evidence in the case. Even knowing about the other evidence could bias the forensic practitioner's conclusion. This would mean that the strength of evidence statement that the forensic practitioner presents to the trier of fact would not be (entirely) new independent information for the trier of fact, and the trier of fact would be in danger of double counting the same information. Cognitive bias is a problem of increasing concern in forensic science (see: Risinger et al., 2002; Saks et al., 2003; National Research Council, 2009; Found, 2015; Stoel et al., 2015; National Commission on Forensic Science, 2015; Edmond et al., 2017).

An arbitrary prior is problematic since if one practitioner used a high value for the prior and another practitioner used a low one, and otherwise acted the same, the difference in the priors would make the value of the first practitioner's posterior probability higher and that of the second lower, but this difference would have nothing to do with the materials they were asked to compare. If the value of the arbitrary prior were not revealed, then the reported posterior probability would be misleading. If the value of the arbitrary prior were revealed along with the value of the posterior, then the value of the likelihood

[8] For coherent odds: $o(H) = p(H)/p(\widetilde{H})$, and $p(\widetilde{H}) = 1 - p(H)$. $\widetilde{H}$ means not H. Hence, via substitution and algebraic derivation, the formula to convert from posterior odds to posterior probability is: $p(H) = o(H)/(1 + o(H))$.

ratio could be recovered, but it would have been much simpler just to present the value of the likelihood ratio in the first place.

It may be the case that practitioners who present posterior probabilities are not aware of the logical problems. Jackson (2009) discusses the problems with posterior probabilities and a range of other ways that have been used to express strength of evidence (see also Hicks, Buckleton, et al., 2015)

3.3 Identification / exclusion / inconclusive framework

In an extreme version of the posterior probability framework, the practitioner only reports either "identification", i.e., 100% probability for same speaker, or "exclusion", i.e., 100% probability for different speaker, or declines to express an opinion "inconclusive". In making an "identification" or "exclusion" the forensic practitioner has made the decision as to same speaker or different speaker, which is properly a decision to be made by the trier of fact who also takes other evidence into consideration. Apart from the logical problems associated with a posterior probability framework in general, "identification" or "exclusion" leads to additional problems. Logically, a practitioner who makes an "identification" or an "exclusion" is claiming infallibility – if they acknowledged that they could be wrong then they could not be 100% certain. Also, logically, the practitioner is claiming that no other evidence in the case is relevant – the trier of fact weighs other evidence against the voice evidence, but a posterior probability of 1 equates to a posterior odds of infinity (which could only be obtained if the prior odds or the likelihood ratio were infinite), and no other evidence can have any counter effect against infinitely strong evidence. The trier of fact could, of course, decide to not believe the forensic practitioner.

President Obama's Council of Advisors on Science and Technology stated that forensic practitioners should not be allowed to claim 100% certainty (President's Council of Advisors on Science and Technology, 2016, p. 19).

3.4 UK framework

In 2007, a group of forensic voice comparison practitioners and researchers in the United Kingdom published a position statement that included a framework for evaluation of evidence (French & Harrison, 2007). This became known as the "UK framework". It was explicitly tied to auditory-acoustic-phonetic subjective approaches.

The framework has two stages: "consistency" and "distinctiveness". In the first stage, the practitioner makes a subjective judgment as to "whether the known and questioned samples are compatible, or consistent, with having been produced by the same speaker" (French & Harrison, 2007, p. 141). The choices are "consistent", "not consistent", or "no-decision". If the practitioner decides that the samples are "not consistent", the practitioner may state that they were spoken by different speakers and express

their degree of confidence that this is so (this is a posterior probability). If the practitioner decides that the samples are "consistent", the practitioner then makes a subjective judgment as to whether the known- and questioned-speaker recordings fall into one of five levels of distinctiveness with respect to the relevant population: "exceptionally-distinctive", "highly-distinctive", "distinctive", "moderately-distinctive", or "not-distinctive".

Unlike the numerator and denominator of a likelihood ratio, "consistency" and "distinctiveness" are not measured on the same scale, and there are no explicit algorithms for assigning values to "consistency" or "distinctiveness". The latter are assigned "informally via the analyst's experience and general linguistic knowledge rather than formally and quantitatively" (French et al., 2010, p. 144). Also, the meaning of the conclusion is ambiguous, and there is no normatively correct way to combine it with the strength of other evidence.

The UK framework has been criticized in Rose & Morrison (2009) and Morrison (2009, 2010, 2014). In 2015, the lead authors of the UK position statement abandoned their framework (see French, 2017) in favor of the Association of Forensic Science Providers' standards (Association of Forensic Science Providers, 2009), which require the use of the likelihood ratio framework. French (2017) indicates that they have adopted the use of verbal expressions of likelihood ratios, with the level on the ordinal scale assigned on the basis of subjective judgment.

3.5 Popularity of different frameworks

In the Gold & French (2011) survey, the 35 respondents' use of different frameworks was reported as follows:

- 4 (11%) used numeric likelihood ratios.
- 3 (9%) used verbal likelihood ratios.
- 14 (40%) used posterior probabilities. It was suggested that most or all of these were verbal expressions of posterior probabilities.
- 2 (6%) used identification / exclusion / inconclusive.
- 11 (31%) used the UK framework.
- 1 (3%) was reported as "other".

In the 2016 INTERPOL survey (Morrison, Sahito, et al., 2016), of the 44 respondents who stated that their agency had speaker recognition capabilities, use of different frameworks was reported as follows:

- 10 (23%) used numeric likelihood ratios.

- 9 (20%) used verbal likelihood ratios.
- 3 (7%) used numeric posterior probabilities.
- 4 (9%) used verbal posterior probabilities.
- 22 (50%) used identification / exclusion / inconclusive.
- 3 (7%) used the UK framework.

Some respondents reported using more than one framework, or both a numeric and verbal variant, hence the summary statistics above add up to more than 44 (18, 41%, used numeric and/or verbal likelihood ratios).

4 Empirical validation

The only way to know how well a forensic comparison system works is to test it. US Federal Rule of Evidence 702 - *Daubert*[9] and England & Wales Criminal Practice Directions[10] Section 19A establish demonstration of scientific validity as a key requirement for admissibility. The following organizations also recommend or require empirical validation of forensic methodologies: National Research Council (2009); Forensic Science Regulator of England & Wales (2014, 2017), as part of accreditation; European Network of Forensic Science Institutes (Drygajlo et al., 2015), specifically for forensic voice comparison; President's Council of Advisors on Science and Technology (2016). Morrison (2014) reviewed calls going back to the 1960s for the validity and reliability of forensic voice comparison systems to be empirically validated under casework conditions.

President Obama's Council of Advisors on Science and Technology stated that:

> neither experience, nor judgment, nor good professional practices (such as certification programs and accreditation programs, standardized protocols, proficiency testing, and codes of ethics) can substitute for actual evidence of foundational validity and reliability. The frequency with which a particular pattern or set of features will be observed in different samples, which is an essential element in drawing conclusions, is not a matter of "judgment." It is an empirical matter for which only empirical evidence is relevant. Similarly, an expert's expression of *confidence* based on personal professional experience or expressions of *consensus* among practitioners about the accuracy of their field is no substitute for error rates estimated from relevant studies. For forensic feature-comparison methods, establishing foundational validity based on empirical evidence is thus a *sine qua non*. Nothing can

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

[10] *Criminal Practice Directions* [2015] EWCA Crim 1567

substitute for it.[11]

Below we describe empirical validation of forensic comparison systems within the likelihood ratio framework. A number of different metrics and graphics have been proposed for assessing and reporting the degree of validity and reliability of likelihood ratio systems (see: Morrison, 2011; Meuwly et al., 2017). Below we describe the most popular metric (C_{llr}) and the most popular graphic (Tippett plot).

4.1 Black-box testing

Black-box testing is concerned with how well a system works, not with how it works. Black-box testing therefore treats all systems equally, irrespective of whether they are based on auditory, spectrographic, acoustic-phonetic, or automatic approaches.

The basic procedure for assessing validity using black-box testing is as follows. The tester presents the system with pairs of voice recordings. The tester knows whether each pair is a same-speaker pair or a different-speaker pair, but the system being tested must not know. For each pair, the system outputs a strength of evidence value. For simplicity, let us imagine a system that outputs "same-speaker" or "different-speaker" (this is an identification / exclusion framework). Table 2 shows the possible input and output combinations, and their correctness. If the answer is correct, the tester assigns it a penalty value of 0. If the answer is incorrect (a *miss* or a *false alarm*), the tester assigns it a penalty value of 1. After all the test pairs have been presented, the tester sums the penalty values and divides by the total number of pairs, i.e., calculates the mean penalty value. Usually, the proportion of misses for all same-speaker input and the proportion of false alarms for all different-speaker input are calculated, then the mean of those two proportions is calculated. The resulting value is called *classification error rate* (its inverse is *correct classification rate*).

<Table 2 about here>

4.2 Log likelihood ratio cost (C_{llr})

Classification error rate is not appropriate for assessing the validity of a system that outputs likelihood ratios. Classification error rate requires a same-speaker or different-speaker decision to be made, which logically requires deciding whether the value of a posterior probability exceeds a threshold or not. Also, even if one decided to use the neutral likelihood ratio value of 1 as a threshold, the further a likelihood ratio from 1 the greater the strength of evidence indicated. If, for example, the tester knew that the input was a different-speaker pair and the system returned a likelihood ratio of 1.1, that would be a false alarm and would attract a penalty value of 1. But if the system returned a likelihood ratio of 1000, that would

[11] President's Council of Advisors on Science and Technology (2016) p. 6, emphasis in original.

also be a false alarm and also attract a penalty value of 1, even though a likelihood ratio of 1000 would be much more misleading to the trier of fact than a likelihood ratio of 1.1.

To perform empirical validation compatible with the likelihood ratio framework, instead of assigning penalty values of either 0 or 1 depending on a threshold, the procedure for calculating the *log likelihood ratio cost* (C_{llr}; see: Brümmer & du Preez, 2006; González-Rodríguez et al., 2007; Morrison, 2011) assigns a penalty value according to the magnitude of the likelihood ratio. The function for calculating C_{llr} is given in Equation 3:

(3)

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

Where Λ_s and Λ_d are likelihood ratio outputs corresponding to same- and different-speaker inputs respectively, and N_s and N_d are the number of same- and different-speaker inputs respectively. Figure 4 plots the penalty functions for likelihood ratio outputs corresponding to same- and different-speaker inputs, i.e., the functions within Equation 3's left and right summations respectively.

<Figure 4 about here>

When the input is a same-speaker pair, large positive log likelihood ratios are good and are assigned low penalty values, small positive log likelihood ratios are not as good and are assigned higher penalty values, negative log likelihood ratios are bad and are assigned yet higher penalty values with higher penalty values for larger magnitude negative log likelihood ratios. *Mutatis mutandis* for when the input is a different-speaker pair.

The better the performance of the system, the lower the value of C_{llr}. A "perfect" system would always give infinite likelihood ratios when the input is same-speaker and always give likelihood ratios of zero when the input is different-speaker, and the value of C_{llr} would be 0. Perfect systems do not exist for non-trivial problems, so in practice C_{llr} will never reach 0. A system which always outputs a likelihood ratio of 1 irrespective of the input would provide no useful information to the trier of fact, and would have a C_{llr} value of 1. In practice, because systems are trained and tested on different data, C_{llr} values can be greater than 1. If a system has a C_{llr} value much greater than 1, then it is probably miscalibrated and better performance would be achieved if it were calibrated (see Section 2.4).

4.3 Tippett plots

The first step in drawing a *Tippett plot* (Meuwly, 2001) is the same as the first step in calculating C_{llr}: input same-speaker and different-speaker pairs to the system and get the corresponding output. The next step is to rank all the same-speaker results in ascending order, and all the different-speaker results in

ascending order. Then the cumulative empirical distribution for each group of values is plotted. If, for example, there are N_s=100 same-speaker outputs, Λ_s, the lowest ranked Λ_{s_i} (i=1) is plotted at its log likelihood ratio value on the *x* axis and at its proportional rank on the *y* axis, i.e., at $y = i/N_s = 1/100$. The second lowest ranked Λ_{s_i} (i=2) is plotted at $x = \Lambda_{s_2}$, $y = 2/100$, the third at $x = \Lambda_{s_3}$, $y = 3/100$, etc. The *y* values of the plotted points represent the proportion of likelihood ratios from same-speaker test pairs that have likelihood ratio values less than or equal to the value indicated on the *x* axis. Conventionally, the points are joined by lines and no symbols are drawn at the actual point values. A similar procedure is used for the Λ_d values, but the *y* values of the plotted points represent the proportion of likelihood ratios from different-speaker test pairs that have likelihood ratio values greater than or equal to the value indicated on the *x* axis.

Figure 5 shows two example Tippett plots based on artificial data created for illustrative purposes. The curve rising to the right represents the same-speaker results, and the curve rising to the left represents the different speaker results. Learning to fully appreciate the information in Tippett plots may take some time, but at a basic level the further to the right and the shallower the slope of the same-speaker curve and the further to the left and the shallower the slope of the different-speaker curve the better the performance. A Tippett plot with fewer or less extreme values for misleading test output, i.e., same-speaker pairs resulting in negative log likelihood ratios and different-speaker pairs resulting in positive log likelihood ratios, generally indicates better performance even if the magnitudes of the log likelihood ratios pointing in the correct directions are less extreme. In Figure 5, the Tippett plot in the bottom panel represents a system with better performance than that in the top panel.

The C_{llr} values corresponding to the top and bottom panels in Figure 5 are 0.548 and 0.101 respectively. Tippet plots include all test results and therefore contain much more information than C_{llr}. C_{llr} is a single value summary metric, and is a many to one mapping – multiple different Tippet plots could correspond to the same C_{llr} value.

<Figure 5 about here>

4.4 Appropriate test data

It is important to use test data that represent the relevant population and reflect the speaking styles and recording conditions for the case under investigation. Tests conducted using data that represent other populations and reflect other conditions may be highly misleading with respect to how well the forensic analysis system will perform when used in the case. A system that works well with studio quality audio recordings may work very poorly under casework conditions that include a mismatch between the known- and questioned-speaker recordings and poor quality recording conditions, e.g., due to background noise, reverberation, transmission through communication channels, and being saved in compressed formats. When putting together test pairs, one member of each pair must reflect the

conditions of the known-speaker recording and the other must reflect the conditions of the questioned-speaker recording.

Whether the test data are sufficiently representative of the relevant population and sufficiently reflective of the speaking styles and recording conditions of the case is a judgment that will initially be made by the forensic practitioner, but the forensic practitioner must clearly communicate what they have done so that ultimately the appropriateness of their decision can be considered by the judge at an admissibility hearing and/or the trier of fact at trial.

It is important to test the system that is actually used to compare the known- and questioned-speaker recordings. For example, if an automatic system is used, but the output of the automatic system is used as input to a subjective judgment process which also includes consideration of the results of analyses based on other approaches, it is the output of the final subjective judgment process that must be empirically validated.

For admissibility, the judge first has to consider whether the test data were sufficiently representative of the relevant population and sufficiently reflective of the speaking styles and recording conditions of the case. If the judge decides that the data are sufficient, then the judge can consider whether the demonstrated degree of performance is sufficient to warrant the admission of testimony based on the forensic comparison system.

In addition, for systems based on statistical models, the test data must not be the same data as were used to train the statistical models. Training and testing on the same data gives misleadingly good results compared to when statistical models are tested on new data. In actual application, the known- and questioned-speaker recordings are new data. Therefore it is performance on new data that matters. Test data must therefore come from a completely separate sample of the relevant population, or a procedure known as cross-validation should be used to avoid training and testing on the same data.

5 Legal admissibility and case law

This section briefly discusses legal admissibility and case law in some common-law jurisdictions: United States (particularly Federal cases), Australia (particularly New South Wales), United Kingdom (particularly Northern Ireland and England & Wales), and Canada. We provide somewhat longer summaries of the Canadian cases because, unlike the cases from the other jurisdictions, they have not been previously described in an academic archival venue.

5.1 United States

In the United States, through the 1960s, 70s, and 80s, testimony based on spectrographic / auditory-spectrographic approaches was often proffered in court proceedings. Based on published rulings, the rate

of admission appears to have been somewhat greater than the rate of exclusion. By the 1990s there had been a substantial decline in the number of cases in which it was proffered. In 2003 in *Angleton*,[12] in an admissibility hearing held under Federal Rule of Evidence (FRE) 702 and the criteria established by the Supreme Court in the 1993 case of *Daubert*,[13] the court ruled an auditory-spectrographic approach inadmissible. Among other criteria, FRE 702 - *Daubert* requires consideration of the empirically demonstrated level of performance of the forensic analysis system. The court found that demonstration of an adequate level of performance was lacking. Based on published rulings, no attempt to admit a spectrographic or auditory-spectrographic approach appears to have survived an FRE 702 - *Daubert* challenge since then.

In 2015 in *Ahmed*,[14] testimony was proffered which was in-part based on an automatic approach, but which was combined with auditory and acoustic-phonetic approaches to reach an ultimately subjective assessment of the strength of evidence. An FRE 702 - *Daubert* admissibility hearing was held, but before the judge ruled on the matter the case was resolved via a negotiated plea deal. Thus no decision on admissibility was issued in that case. During the hearing, questions were raised as to whether appropriate data had been used to train the automatic component of the system, whether the system had been empirically tested under conditions reflecting those of the case, and whether the subjective procedure for combining the output of the automatic component and the outputs of the subjective auditory and acoustic-phonetics components was influenced by cognitive bias.

For more detailed discussion of admissibility of forensic voice comparison testimony in the United States see Morrison & Thompson (2017).

5.2 Australia

In New South Wales in 1977 in *Gilmore*[15] the court ruled testimony based on an auditory-spectrographic approach admissible. The decision was based in part on the fact that such testimony had been ruled admissible by a number of courts in the United States in the early to mid 1970s. In a 2012 admissibility hearing in the New South Wales case of *Ly*,[16] the admissibility of testimony based on an auditory-spectrographic approach was challenged. Despite the change that had occurred in the US with respect to admissibility of auditory-spectrographic approaches, the court in *Ly* ruled that *Gilmore* was precedential and that the testimony was therefore admissible (for further discussion, see Enzinger & Morrison, 2017).

---

[12] *United States v Robert N. Angleton*, 269 F.Supp. 2nd 892 (S.D. Tex. 2003)

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

[14] *United States v Ali Ahmed, Madhi Hashi, & Muhamed Yusuf*, No. 12-661 (E.D.N.Y.)

[15] *R v Gilmore*, 1977, 2 NSWLR 935

[16] *R v Ly*, NSW District Court, 2010/295928 (note that this is a reference to an earlier hearing in the case)

Arguable, in *Ly* the practitioner used a mixture of auditory, spectrographic, and acoustic-phonetic subjective approaches, with subjective-judgment used to combine the results. We are aware of several other instances in which auditory-spectrographic-acoustic-phonetic-subjective or auditory-acoustic-phonetic-subjective analyses have been admitted in courts in several Australian jurisdictions.

In 2008 in *Hufnagl*,[17] testimony based on an acoustic-phonetic statistical approach was admitted (see Rose, 2013). In 2017, testimony based on an automatic approach was submitted in a New South Wales case, but the case was resolved by plea deal before going to trial (see Morrison, 2018b, in which questions are raised with respect to whether appropriate data were used for training and whether performance was empirically tested under conditions reflecting those of the case).

5.3 United Kingdom

There are rulings from Northern Ireland and from England & Wales specific to the admissibility of forensic voice comparison testimony. We are not aware of any such rulings from Scotland.

In Northern Ireland in the 2002 case of *O'Doherty*,[18] the appeal court ruled an auditory-only approach inadmissible, but auditory-acoustic-phonetic subjective approaches admissible. It was reported that most practitioners considered auditory-only approaches to be unreliable.

In England & Wales in 1991 in *Robb*,[19] the appeal court ruled an auditory-only approach admissible. In 2008 in *Flynn*,[20] the appeal court opined that *Robb* was still precedential and that courts in England & Wales should not follow the example set in Northern Ireland in *O'Doherty*. The opinion in *Flynn* was echoed in the 2015 appeal court ruling in *Slade*.[21]

The appeal court in *Slade* considered the admissibility of new evidence consisting of forensic voice comparison testimony based in-part on an automatic approach, but which was combined with an auditory-acoustic-phonetic subjective approach to reach an ultimately subjective assessment of the strength of evidence. Some empirical testing of the performance of the automatic component of the analysis was presented, but the court was not satisfied with the quantity and quality of the data used to train and test the automatic system. Nor was it satisfied with the empirically demonstrated level of performance of the system. Also note that what was tested was not the whole auditory-acoustic-phonetic-automatic plus combination-via-subjective-judgment system actually used to assess strength of evidence, but only the automatic component. What needs to be tested is the performance of the whole system, not

[17] *R v Hufnagl*, NSW District Court, 2008

[18] *R v O'Doherty* [2002] NICA 20 / [2003] 1 Cr App R 5

[19] *R v Robb* [1991] 93 Cr App R 161

[20] *R v Flynn and St John* [2008] EWCA Crim 970

[21] *R v Slade et al.* [2015] EWCA Crim 71

just a component of the system (Forensic Science Regulator, 2014, §3.3.1–3.3.2). The appeal court ruled the testimony based in-part on the automatic analysis inadmissible. The ruling was specific to this instance, and did not preclude testimony based on an automatic approach being admissible in future cases. Ironically, testimony based on auditory-only and auditory-acoustic-phonetic subjective approaches had been admitted at trial despite the fact that they had not undergone any empirical testing (admissibility of these approaches does not appear to have been challenged at any point in the proceedings).

Shortly before the appeal in *Slade*, new Criminal Practice Directions on admissibility of expert evidence (CPD 19A) were introduced.[22] The CPD 19A admissibility criteria are similar to those of FRE 702 - *Daubert*. It is not clear whether they had any impact on the decision in *Slade*, but it may be that they will have an impact on future admissibility decisions. For more detailed discussion of admissibility of forensic voice comparison testimony in England & Wales and Northern Ireland see Morrison (2018a).[23]

5.4 Canada

In the 1998 labor arbitration case of *Ontario Hydro v Canadian Union of Public Employees*,[24] testimony was proffered from a university-based academic who taught applied linguistics, speech-language pathology, and phonetics and phonology, and who had conducted research in analysis of normal, pathological, and foreign-accented speech. He performed an auditory-only analysis. The Board of Arbitration considered a number of admissibility criteria identified in rulings made by courts of law. The Board found that the academic was not qualified as an expert in forensic voice comparison because he had no training or experience specifically in that field. The Board found that "his knowledge and experience as a phonetician are not sufficient for this purpose". The Board also found that "there is nothing to indicate that the method used by [the academic] to reach a conclusion regarding voice identification in this case has gained any acceptance and, based on the evidence, we find that it fails to meet a threshold test of reliability." The standards against which the Board compared the academic's auditory-only analysis were, however, those of organizations such as the American Board of Recorded Evidence (1999) which required auditory plus spectrographic analysis, and no mention was made of criticisms of the auditory-spectrographic approach (this case was prior to the US case of *Angleton*).

In 2018 in *R v Dunstan*[25] in a *Charter* Section 8 Application[26] the Ontario Superior Court of Justice considered forensic voice comparison testimony from three practitioners. For this judge-only hearing

[22] *Criminal Practice Directions* [2015] EWCA Crim 1567

[23] For a history of forensic voice comparison in the UK from another perspective, see French (2017).

[24] *Ontario Hydro v Canadian Union of Public Employees* [1998] OLAA No 691

[25] *R v Dunstan* [2018] ONSC 4153

[26] *Canadian Charter of Rights and Freedoms*, s 8, Part I of the Constitution Act, 1982, being Schedule B to the Canada Act 1982 (UK), 1982, c 11. Section 8 provides protection against unreasonable search and seizure.

none of the testimony was ruled admissible or inadmissible *per se*, but in her ruling the judge commented on the appropriateness of the different methodologies used by the practitioners. The recording of interest was of a telephone call made to a police call center six years previously. The recording was noisy and had been saved in a lossy compressed format. The total duration of the speech of the speaker of questioned identity was approximately 10 seconds.

At trial, testimony had been presented by an audio engineer. He applied noise reduction to the questioned-speaker recording, then created multiple compilations in each of which the questioned-speaker recording was either non-synchronously intercalated/concatenated or synchronously mixed with a time-warped recording of either the known speaker, the practitioner, or a third speaker. The judge at the Section 8 hearing concluded that "The comparisons were not fair and reliable." Reasons included the following: Of several known-speaker recordings available, the particular known-speaker recordings used had been selected because they were the ones that sounded the most similar to the questioned speaker. The known-speaker recordings had been time-warped to make them more similar to the recording of the questioned speaker.

One of two practitioners to testify at the Section 8 hearing was a speech-language pathologist who conducted an auditory-acoustic-phonetic subjective analysis. The acoustic-phonetic component was based on acoustic measurements of jitter in vocal fold vibration. No data representative of a relevant population were used (a comparison was made with a recording of the practitioner's own voice), no empirical validation was conducted, and the practitioner's conclusion was reported as a subjective posterior probability. The judge concluded that the practitioner "was not an objective, unbiased witness... He did not express, either in his report or in his testimony, an understanding of his duty to the court to be impartial, independent and unbiased." "[He] said that his task was to find similarities between the two voices, so he did not report dissimilarities that he observed." "[He] was unable to say what proportion of Canadian males have a jitter measurement of 7 per cent ... He did not give evidence from which it can be concluded that the nine test results he provided represent an empirically sound basis from which to draw inferences about voices in the forensic context".

The other practitioner to testify at the Section 8 hearing was the first author of the present chapter. He performed a human-supervised automatic analysis, which included a statistical procedure to reduce the likelihood of overstating the strength of evidence (Morrison & Poh, 2018), and a statistical procedure to take account of the six-year gap between the questioned- and known-speaker recordings even though the recordings available for training and testing were made only hours to days apart.[27] Steps were also taken to reduce the potential for cognitive bias, including that the practitioner did not listen to both the questioned- and known-speaker recordings, he listened to the questioned-speaker recording to prepare it for input into the automatic system but his assistant prepared the known-speaker recording. The automatic system was trained/optimized using multiple recordings of just over 100 speakers specially

[27] This was developed specifically for use in this case building on research reported in Kelly & Hansen (2016).

collected with the intention of representing the relevant population and reflecting the recording conditions in this case – male General Canadian English speakers made multiple mobile telephone calls to the same police call center using the same call recording system as had been used to record the questioned-speaker recording six years previously. Empirical validation was conducted using the same recordings (cross-validation was used to avoid training and testing on the same data) and the validation results were reported. The questioned- and known-speaker recordings were then compared, and the likelihood ratio value output by the system was directly reported as the strength of evidence statement. The calculated likelihood ratio value was close to 1. In theory, a likelihood ratio value of 1 should have no effect on the trier of fact's belief as to the relative probabilities of the same- and different-speaker hypotheses, the posterior odds should be equal to the prior odds, but the practitioner was called to testify in order to contrast his methodology and result with the other practitioners' methodologies and their strong claims about the identity of the questioned speaker. The judge "accept[ed] that [the practitioner] is very well qualified in his field", but had reservations about his application of the human-supervised automatic approach: 1. "it is novel and must be carefully scrutinized." 2. The questioned-speaker speech was only 10 seconds long.[28] 3. With respect to the statistical procedure used to compensate for the six-year time difference between the questioned- and known speaker recording, she was concerned that it had not been adequately validated.[29] 4. With respect to the speakers intended to be representative of the relevant population, who were all police officers, she was concerned that police officers may speak differently from other speakers of General Canadian English.[30]

6 Conclusion

Although the case law and rulings on legal admissibility in common-law jurisdictions remain mixed, we believe that the future of forensic voice comparison lies in the use of human-supervised automatic approaches within the likelihood ratio framework, with empirical testing of system performance under casework conditions and direct reporting of the calculated likelihood ratio value. We argue that this is the most practical way to meet rigorous applications of legal admissibility criteria such as those of FRE

[28] This does not seem to have taken into account that the performance of the forensic voice comparison system was empirically validated under conditions reflecting those of the case under investigation, including the condition that questioned-speaker speech had a duration of 10 seconds. We would argue that a decision on whether to use the output of a system should be made on the basis of consideration of results of empirical testing of that system under conditions reflecting those of the case, rather than directly on what those conditions happen to be.

[29] Published papers including descriptions of the data used to train the statistical procedure were submitted along with the original report, and the practitioner offered to provide a supplemental report he had prepared on the validation of the statistical procedure.

[30] This was an argument advanced by the party who proposed the same-speaker hypothesis. We do not believe that this argument has any merit, but, even if it did, the same-speaker hypothesis they proposed was that the questioned speaker was a particular police officer. Hence there was no basis for their objection that the sample recordings were recordings of police officers and that typicality was therefore assessed with respect to male General Canadian English speakers who were police officers rather than with respect to male General Canadian English speakers in general.

702 - *Daubert* and CPD 19A.

Systems in which the output is directly based on subjective judgment are not transparent, not replicable, and are highly susceptible to cognitive bias. Even if an explanation is given, there is no way to know whether it corresponds to what actually happened in the practitioner's mind or whether it is a *post hoc* rationalization. In contrast, systems based on relevant data, quantitative measurements, and statistical models are transparent and replicable. The procedures can be described in exact detail, and the data and software used can even be provided.

It should be noted that procedures based on relevant data, quantitative measurements, and statistical models do require subjective judgments, but these are judgments about relevant populations and relevant data which are far removed from the output of the system. The appropriateness of such judgments should be debated before the judge at an admissibility hearing and/or the trier of fact at trial. After these initial judgments, the remainder of the system is objective. Systems based on relevant data, quantitative measurements, and statistical models are therefore much more resistant to the potential influence of cognitive bias than are systems in which the output is directly based on subjective judgment.

In addition, procedures based directly on subjective judgment generally require considerable human time to perform each test trial. Thus they are practically difficult to test compared to an automatic procedure that can run thousands of test trials in seconds. If a subjective procedure were found to outperform an automatic procedure, it would be preferred, but performance would have to be empirically demonstrated under relevant conditions.

We prefer automatic approaches over acoustic-phonetic statistical approaches because we have found the former to outperform the latter, especially under forensically realistic conditions or conditions approaching forensically realistic conditions (see: Enzinger et al., 2012; Zhang et al., 2013; Zhang & Enzinger, 2013; Enzinger, 2014; Enzinger & Kasess, 2014; Jessen et al., 2014; Enzinger & Morrison, 2017).[31] We also prefer automatic approaches because acoustic-phonetic approaches generally require much greater investment of human time and therefore take longer and are more expensive.

In evaluation of strength of evidence for presentation in court, it would be inappropriate for a system to be fully automatic, it should be human supervised. The forensic practitioner is responsible for determining an appropriate question to ask and selecting appropriate data and statistical models in order to answer that question. An automatic system is a tool, and inappropriate use of the tool will lead to inappropriate results. A potential danger of automatic systems is that they could be too easy to use, and therefore too easy to misuse. Appropriate training and knowledge is therefore essential.

[31] These are primarily papers written by ourselves. Our backgrounds are in acoustic phonetics, and we thus had an interest in empirically assessing the performance of acoustic-phonetic-statistical approaches under casework conditions. Other than our work, there is very little published work in which acoustic-phonetic-statistical and automatic systems have both been empirically tested on the same test data under forensically realistic or close to forensically realistic conditions.

7 References

Aitken, C.G.G., Berger, C.E.H., Buckleton, J.S., Champod, C., Curran. J.M., Dawid, A.P., Evett, I.W., Gill, P., González-Rodríguez, J., Jackson, G., Kloosterman, A., Lovelock, T., Lucy, D., Margot, P., McKenna, L., Meuwly, D., Neumann, C., Nic Daéid, N., Nordgaard, A., Puch-Solis, R., Rasmusson, B., Redmayne, M., Roberts, P., Robertson, B., Roux, C., Sjerps, M.J., Taroni, F., Tjin-A-Tsoi, T., Vignaux, G.A., Willis, S.M. and Zadora, G. (2011). Expressing evaluative opinions: A position statement. *Science & Justice*, 51, 1–2. Available at: http://dx.doi.org/10.1016/j.scijus.2011.01.002

Aitken, C.G.G., Roberts, P. and Jackson G. (2010). *Fundamentals of Probability and Statistical Evidence in Criminal Proceedings: Guidance for Judges, Lawyers, Forensic Scientists and Expert Witnesses*. Royal Statistical Society, London, UK. Available at: http://bit.ly/1WnoXRx [Accessed 1 February 2017]

Ajili, M. (2017). *Reliability of voice comparison for forensic applications*. PhD. University of Avignon.

Ajili, M., Bonastre, J.F., Ben Kheder W., Rossato S. and Kahn J. (2016). Phonetic content impact on forensic voice comparison. In *Proceedings of the 2016 IEEE Spoken Language Technology Workshop (SLT)*. pp. 201–217. Available at: http://dx.doi.org/10.1109/SLT.2016.7846267

American Board of Recorded Evidence (1999). *Voice comparison standards*. Available at: http://www.tapeexpert.com/pdf/abrevoiceid.pdf [Accessed 1 February 2010].

Association of Forensic Science Providers (2009). Standards for the formulation of evaluative forensic science expert opinion. *Science & Justice*, 49, 161–164. Available at: https://doi.org/10.1016/j.scijus.2009.07.004

Balding, D.J. and Steele, C. (2015). *Weight-of-evidence for forensic DNA profiles*. 2nd ed. Chichester, UK: Wiley. Available at: https://doi.org/10.1002/9781118814512

Becker T. (2012). *Automatischer forensischer Stimmenvergleich* [Automatic forensic voice comparison]. PhD. University of Trier.

Brümmer N. and du Preez J. (2006). Application independent evaluation of speaker detection. *Computer Speech and Language*, 20, 230–275. Available at: https://doi.org/10.1016/j.csl.2005.08.001

Cambier-Langeveld T., van Rossum M. and Vermeulen J. (2014). Whose voice is that? Challenges in forensic phonetics. In van Heuven V. and Caspers J., eds., *Above and Beyond the Segments: Experimental Linguistics and Phonetics*. Amsterdam: John Benjamins, pp. 14–27.

Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4), 357–366. Available at: https://doi.org/10.1109/TASSP.1980.1163420

Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P. and Ouellet, P. (2011). Front-end factor analysis for speaker verification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 19(4), 788–798. Available at: https://doi.org/10.1109/TASL.2010.2064307

Drygajlo, A., Jessen, M., Gfroerer, S., Wagner, I., Vermeulen, J. and Niemi T. (2015). *Methodological guidelines for best practice in forensic semiautomatic and automatic speaker recognition, including guidance on the conduct of proficiency testing and collaborative exercises*. European

Network of Forensic Science Institutes. Available at: http://enfsi.eu/wp-content/uploads/2016/09/guidelines_fasr_and_fsasr_0.pdf [Accessed 1 January 2018].

Edmond, G., Towler, A., Growns, B., Ribeiro, G., Found, B., White, D., Ballantyne, K., Searston, R.A., Thompson, M.B., Tangen, J.M., Kemp, R.I. and Martire K. (2017). Thinking forensics: Cognitive science for forensic practitioners. *Science & Justice*, 57, 144–154. Available at: http://dx.doi.org/10.1016/j.scijus.2016.11.005

Enzinger, E. (2014). A first attempt at compensating for effects due to recording-condition mismatch in formant-trajectory-based forensic voice comparison. In *Proceedings of the 15th Australasian International Conference on Speech Science and Technology*. Australasian Speech Science and Technology Association. pp. 133–136 Available at: http://www.assta.org/sst/SST-14/6.A.%20 FORENSICS%202/1.%20ENZINGER.pdf [Accessed 1 February 2017]

Enzinger, E. (2016). *Implementation of forensic voice comparison within the new paradigm for the evaluation of forensic evidence*. PhD. University of New South Wales. Available at: http://handle.unsw.edu.au/1959.4/55772 [Accessed 1 February 2017]

Enzinger, E., and Kasess, C.H. (2014). Bayesian vocal tract model estimates of nasal stops for speaker verification. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2014)*. pp. 1685–1689. Available at: http://dx.doi.org/10.1109/ ICASSP.2014.6853885

Enzinger, E., Morrison, G.S. and Ochoa, F. (2016). A demonstration of the application of the new paradigm for the evaluation of forensic evidence under conditions reflecting those of a real forensic-voice-comparison case. *Science & Justice*, 56, 42–57. Available at: http://dx.doi.org/ 10.1016/j.scijus.2015.06.005

Enzinger, E. and Morrison, G.S. (2017). Empirical test of the performance of an acoustic-phonetic approach to forensic voice comparison under conditions similar to those of a real case. *Forensic Science International*, 277, 30–40. Available at: http://dx.doi.org/10.1016/j.forsciint.2017.05.007

Enzinger, E., Zhang, C. and Morrison G.S. (2012). Voice source features for forensic voice comparison – an evaluation of the Glottex® software package. In *Proceedings of Odyssey 2012, The Language and Speaker Recognition Workshop*. pp. 78–85. Available at: http://isca-speech.org/archive/ odyssey_2012/od12_078.html [Accessed 1 February 2017] [Errata and addenda available at: https://box.entn.at/pdfs/enzinger2012_odyssey_vsferradd.pdf Accessed 1 February 2017]

Forensic Science Regulator (2014). *Guidance on validation (FSR-G-201 Issue 1)*. Birmingham, UK: Forensic Science Regulator. Available at: https://www.gov.uk/government/publications/forensic-science-providers-validation [Accessed 19 March 2017]

Forensic Science Regulator (2017). *Codes of practice and conduct for forensic science providers and practitioners in the criminal justice system (version 4.0)*. Birmingham, UK: Forensic Science Regulator. Available at: https://www.gov.uk/government/publications/forensic-science-providers-codes-of-practice-and-conduct-2017 [Accessed 13 June 2018]

Found, B. (2015). Deciphering the human condition: The rise of cognitive forensics. *Australian Journal of Forensic Sciences*, 47, 386–401. Available at: http://dx.doi.org/10.1080/00450618.2014.965204

French, J.P. (2017). A developmental history of forensic speaker comparison in the UK. *English Phonetics*, 21, 271 – 286.

French J.P. and Harrison P. (2007). Position statement concerning use of impressionistic likelihood terms in forensic speaker comparison cases. *International Journal of Speech, Language and the Law*, 14, 137–144. Available at: https://doi.org/10.1558/ijsll.v14i1.137

French, J.P., Nolan, F., Foulkes, P., Harrison P. and McDougall, K. (2010). The UK position statement on forensic speaker comparison: A rejoinder to Rose and Morrison. *International Journal of Speech, Language and the Law*, 17, 143–152. Available at: https://doi.org/10.1558/ijsll.v17i1.143

French, J.P. and Stevens L. (2013). Forensic speech science. In Jones M.J. and Knight R.A., editors, *The Bloomsbury Companion to Phonetics*. London, UK: Bloomsbury, pp. 183–197. Available at: http://dx.doi.org/10.5040/9781472541895.ch-012

Furui, S. (1981). Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(2), 254–272. Available at: https://doi.org/10.1109/TASSP.1981.1163530

Furui, S. (1986). Speaker-independent isolated word recognition using dynamic features of speech spectrum. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34, 52–59. Available at: https://doi.org/10.1109/TASSP.1986.1164788

Gold, E. and French J.P. (2011). International practices in forensic speaker comparison. *International Journal of Speech, Language and the Law*, 18, 143–152. Available at: http://dx.doi.org/10.1558/ijsll.v18i2.293

Gold E. and Hughes V. (2014). Issues and opportunities: The application of the numerical likelihood ratio framework to forensic speaker comparison. *Science & Justice*, 54, 292–299. Available at: http://dx.doi.org/10.1016/j.scijus.2014.04.003

González-Rodríguez, J., Rose, P., Ramos, D., Toledano, D.T. and Ortega-García J. (2007). Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15, 2104–2115. Available at: https://doi.org/10.1109/TASL.2007.902747

Gruber, J.S. and Poza F. (1995). Voicegram Identification Evidence. *American Jurisprudence Trials*, volume 54. Westlaw.

Hansen, J.H.L. and Hasan, T. (2015). Speaker recognition by machines and humans: A tutorial review. *IEEE Signal Processing Magazine*, November 74–99. Available at: http://dx.doi.org/10.1109/MSP.2015.2462851

Hicks, T., Biedermann, A., de Koeijer, J.A., Taroni, F., Champod, C. and Evett I.W. (2015). The importance of distinguishing information from evidence-observations when formulating propositions. *Science & Justice*, 55, 520–525. Available at: http://dx.doi.org/10.1016/j.scijus.2015.06.008

Hicks, T., Biedermann, A., de Koeijer, J.A., Taroni, F., Champod, C. and Evett I.W. (2017). Reply to Morrison et al. (2016) Refining the relevant population in forensic voice comparison – A response to Hicks et alii (2015) The importance of distinguishing information from evidence/observations when formulating propositions, *Science & Justice*, 57, 401–402. Available at: http://dx.doi.org/10.1016/j.scijus.2017.04.005

Hicks, T., Buckleton, J.S., Bright, J.A. and Taylor D. (2015). A Framework for interpreting evidence. In: Buckleton, J.S., Bright, J.A. and Taylor, D., eds., *Forensic DNA Evidence Interpretation*. 2nd ed. Boca Raton, FL: CRC, pp. 37–86.

Hollien, H. (2002). *Forensic voice identification*. San Diego, CA: Academic Press.

Hollien, H. (2016). An approach to speaker identification. *Journal of Forensic Sciences*, 61, 334–344. Available at: http://dx.doi.org/10.1111/1556-4029.13034

Hollien, H., Didla, G., Harnsberger, J.D. and Hollien, K.A. (2016). The case for aural perceptual speaker identification. *Forensic Science International*, 269, 5–20. Available at: http://dx.doi.org/10.1016/j.forsciint.2016.08.007

Hughes, V. and Foulkes, P. (2015). The relevant population in forensic voice comparison: Effects of varying delimitations of social class and age. *Speech Communication*, 66, 218–230. Available at: http://dx.doi.org/10.1016/j.specom.2014.10.006

Hughes, V. and Rhodes, R. (2018). Questions, propositions and assessing different levels of evidence: Forensic voice comparison in practice. *Science & Justice*, 58, 250–257. Available at: https://doi.org/10.1016/j.scijus.2018.03.007

Jackson, G. (2009). Understanding forensic science opinions. In: Fraser, J. and Williams, R., eds., *Handbook of Forensic Science*. Cullompton, UK: Willan, pp 419–445.

Jessen, M. (2008). Forensic phonetics. *Language and Linguistics Compass*, 2, 671–711. Available at: https://doi.org/10.1111/j.1749-818x.2008.00066.x

Jessen, M., Alexander, A. and Forth O. (2014). Forensic voice comparisons in German with phonetic and automatic features using Vocalise software. In: *Proceedings of the 54th Audio Engineering Society (AES) Forensics Conference*, pp. 28–35.

Kelly F. and Hansen J.H.L. (2016). Score-aging calibration for speaker verification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24, 2414–2424. Available at: http://dx.doi.org/10.1109/TASLP.2016.2602542

Kersta, L.G. (1962). Voiceprint identification. *Nature*, 196, 1253–1257. Available at: https://doi.org/10.1038/1961253a0

Nolan, F. (1997). Speaker recognition and forensic phonetics. In: Hardcastle, W.J. and Laver, J., eds., *The handbook of phonetic sciences*. Oxford, UK: Blackwell.

Marks, D.B. (2017). *A framework for performing forensic and investigatory speaker comparisons using automated methods*. MSc. University of Colorado Denver.

Marquis, R., Biedermann, A., Cadola, L., Champod, C., Gueissaz, L., Massonnet, G., Mazzella, W.D., Taroni, F. and Hicks, T.N. (2016). Discussion on how to implement a verbal scale in a forensic laboratory: Benefits, pitfalls and suggestions to avoid misunderstandings. *Science & Justice*, 56, 364–370. Available at: http://dx.doi.org/10.1016/j.scijus.2016.05.009

Meuwly, D. (2001). *Reconnaissance de locuteurs en sciences forensiques: l'apport d'une approche automatique* [Speaker recognition in forensic science: The contribution of an automatic approach]. PhD. University of Lausanne.

Meuwly, D. (2003a). Le mythe de l'empreinte vocale I [The myth of voiceprinting I]. *Revue Internationale de Criminologie et Police Technique*, 56, 219–236.

Meuwly D., (2003b). Le mythe de l'empreinte vocale II [The myth of voiceprinting II]. *Revue Internationale de Criminologie et Police Technique*, 56, 361–374.

Meuwly, D., Ramos, D. and Haraksim R. (2017). A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation, *Forensic Science International*, 276, 142–153. Available at: http://dx.doi.org/10.1016/j.forsciint.2016.03.048

Morrison, G.S. (2009). Forensic voice comparison and the paradigm shift. *Science & Justice*, 49, 298–308. Available at: https://doi.org/10.1016/j.scijus.2009.09.002

Morrison, G.S. (2010). Forensic voice comparison. In Freckelton I. and Selby, H., eds., *Expert Evidence*. Sydney, Australia: Thomson Reuters, ch. 99.

Morrison, G.S. (2011). Measuring the validity and reliability of forensic likelihood-ratio systems. *Science & Justice*, 51, 91–98. Available at: http://dx.doi.org/10.1016/j.scijus.2011.03.002

Morrison, G.S. (2013). Tutorial on logistic-regression calibration and fusion: Converting a score to a likelihood ratio. *Australian Journal of Forensic Sciences*, 45, 173–197. Available at: http://dx.doi.org/10.1080/00450618.2012.733025

Morrison, G.S., (2014). Distinguishing between forensic science and forensic pseudoscience: Testing of validity and reliability, and approaches to forensic voice comparison. *Science & Justice*, 54, 245–256. Available at: http://dx.doi.org/10.1016/j.scijus.2013.07.004

Morrison, G.S. (2018a). Admissibility of forensic voice comparison testimony in England and Wales. *Criminal Law Review*, (1), 20–33.

Morrison, G.S. (2018b). The impact in forensic voice comparison of lack of calibration and of mismatched conditions between the known-speaker recording and the relevant-population sample recordings. *Forensic Science International*, 283, e1–e7. Available at: http://dx.doi.org/10.1016/j.forsciint.2017.12.024

Morrison, G.S. and Enzinger, E. (2016). What should a forensic practitioner's likelihood ratio be? *Science & Justice*, 56, 374–379. Available at: http://dx.doi.org/10.1016/j.scijus.2016.05.007

Morrison, G.S., Enzinger E. and Zhang C. (2016). Refining the relevant population in forensic voice comparison – A response to Hicks et alii (2015) The importance of distinguishing information from evidence/observations when formulating propositions. *Science & Justice*, 56, 492–497. Available at: http://dx.doi.org/10.1016/j.scijus.2016.07.002

Morrison, G.S., Enzinger, E., Zhang, C. (2017). Reply to Hicks et alii (2017) Reply to Morrison et alii (2016) Refining the relevant population in forensic voice comparison - A response to Hicks et alii (2015) The importance of distinguishing information from evidence/observations when formulating propositions. Available at: http://arxiv.org/abs/1704.07639

Morrison, G.S., Enzinger E. and Zhang C. (2018). Forensic speech science. In Freckelton I. and Selby, H., eds., *Expert Evidence*. Sydney, Australia: Thomson Reuters, ch. 99.

Morrison, G.S., Kaye D.H., Balding D.J., Taylor D., Dawid P., Aitken C.G.G., Gittelson S., Zadora G., Robertson B., Willis S.M., Pope S., Neil M., Martire K.A., Hepler A., Gill R.D., Jamieson A., de Zoete J., Ostrum R.B. and Caliebe A. (2016). A comment on the PCAST report: Skip the "match"/"non-match" stage. *Forensic Science International*, 272, e7–e9. Available at: http://dx.doi.org/10.1016/j.forsciint.2016.10.018

Morrison, G.S., Ochoa F. and Thiruvaran T. (2012). Database selection for forensic voice comparison. In: *Proceedings of Odyssey 2012: The Language and Speaker Recognition Workshop*. International Speech Communication Association, pp. 62–77.

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

Morrison, G.S., Sahito, F.H., Jardine, G., Djokic, D., Clavet, S., Berghs, S., and Goemans Dorny, C. (2017). INTERPOL survey of the use of speaker identification by law enforcement agencies. *Forensic Science International*, 263, 92–100. Available at: http://dx.doi.org/10.1016/j.forsciint.2016.03.044

Morrison, G.S. and 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.

National Commission on Forensic Science (2015). *Ensuring that forensic analysis is based upon task-relevant information*. Available at: https://www.justice.gov/ncfs/file/818196/download [Accessed: 1 January 2018].

National Research Council (1979). *On the theory and practice of voice identification*. Washington, DC: National Academies Press.

National Research Council (2009). *Strengthening forensic science in the United States: A path forward*. Washington, DC: National Academies Press.

Nolan, F. (1997). Speaker recognition and forensic phonetics. In: Hardcastle, W.J. and Laver, J., eds., *The handbook of phonetic sciences*. Oxford, UK: Blackwell, pp. 744–767.

Nolan, F. (2005). Forensic speaker identification and the phonetic description of voice quality. In: Hardcastle, W.J. and Mackenzie Beck, J., eds., *A figure of speech: A festschrift for John Laver*. Mahwah, NJ: Erlbaum, pp. 385–411.

Pelecanos, J. and Sridharan, S. (2001). Feature warping for robust speaker verification. In: *Proceedings of Odyssey 2001: The Speaker Recognition Workshop*, pp. 213–218.

Pigeon, S., Druyts, P. and Verlinde, P. (2000). Applying logistic regression to the fusion of the NIST'99 1-speaker submissions, *Digital Signal Processing*, 10, 237–248. Available at: http://dx.doi.org/10.1006/dspr.1999.0358

Poza, F. and Begault, D.R. (2005). Voice identification and elimination using aural-spectrographic protocols. In: *Proceedings of the Audio Engineering Society 26th International Conference: Audio Forensics in the Digital Age*, paper number 1-1.

President's Council of Advisors on Science and Technology (2016). *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*. Available at: https://obamawhitehouse.archives.gov/administration/eop/ostp/pcast/docsreports/ [Accessed: 6 February 2017]

Prince, S.J.D. and Elder, J.H. (2007). Probabilistic linear discriminant analysis for inferences about identity. In *Proceedings of the IEEE 11th International Conference on Computer Vision (ICCV)*, pp. 1–8. Available at: https://doi.org/10.1109/ICCV.2007.4409052

Ramos Castro, D. (2007). *Forensic evaluation of the evidence using automatic speaker recognition systems*. PhD. Autonomous University of Madrid.

Reynolds, D.A., Quatieri, T.F. and Dunn, R.B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10, 19–41. Available at: https://doi.org/10.1006/dspr.1999.0361

Richardson, F., Reynolds, D.A. and Dehak, N. (2015). Deep neural network approaches to speaker and language recognition. IEEE Signal Processing Letters, 22, 1671–1675. Available at: https://doi.org/10.1109/LSP.2015.2420092

Risinger, D.M., Saks, M.J., Thompson, W.C. and Rosenthal R. (2002). The Daubert/Kumho implications of observer effects in forensic science: Hidden problems of expectation and suggestion. *California Law Review*, 90, 1–56. Available at: http://www.jstor.org/stable/3481305

Robertson, B., Vignaux G.A., and Berger C.E.H. (2016). *Interpreting Evidence: Evaluating Forensic Science in the Courtroom*. 2nd ed. Chichester, UK: Wiley. Available at: https://doi.org/10.1002/9781118492475

Rose, P. (2002). *Forensic speaker identification*. London, UK: Taylor and Francis.

Rose, P. (2006). Technical forensic speaker recognition. *Computer Speech and Language*, 20: 159–191. https://doi.org/10.1016/j.csl.2005.07.003

Rose, P., (2013). Where the science ends and the law begins- likelihood ratio-based forensic voice comparison in a $150 million telephone fraud. *International Journal of Speech, Language and the Law*, 20, 227–324. Available at: http://dx.doi.org/10.1558/ijsll.v20i2.277

Rose, P. (2017). Likelihood ratio-based forensic voice comparison with higher level features: Research and reality. *Computer Speech & Language*, 45, 475–502 Available at: http://dx.doi.org/10.1016/j.csl.2017.03.003

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, 139–163. Available at: http://dx.doi.org/10.1558/ijsll.v16i1.139

Saks, M.J., Risinger D.M., Rosenthal, R. and Thompson, W.C. (2003). Context effects in forensic science: a review and application of the science of science to crime laboratory practice in the United States, *Science & Justice*, 43, 77–90. Available at: http://dx.doi.org/10.1016/S1355-0306(03)71747-X

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

Stoel, R.D., Berger, C.E.H., Kerkhoff, W., Mattijssen, E.J.A.T. and Dror E.I. (2015). Minimizing contextual bias in forensic casework. In: Strom K.J. and Hickman M.J., eds., *Forensic Science and the Administration of Justice: Critical Issues and Directions*. Thousand Oaks, CA: Sage, pp. 67–86. Available at: http://dx.doi.org/10.4135/9781483368740.n5

Thompson, W.C., Hofstein Grady, R., Lai, E. and Stern, H.S. (2018). Perceived strength of reporting statements about source conclusions. *Law, Probability and Risk*, 17, 133–155. Available at: http://dx.doi.org/10.1093/lpr/mgy012

Tibrewala, S. and Hermansky, H. (1997). Multi-band and adaptation approaches to robust speech recognition. In: *Proceedings of Eurospeech*, pp. 2619–2622.

Tosi, O. (1979). *Voice Identification: Theory and Legal Applications*. Baltimore, MD: University Park Press.

Willis, S.M., McKenna, L., McDermott, S., O'Donell, G., Barrett, A., Rasmusson, A., Nordgaard, A., Berger, C.E.H., Sjerps, M.J., Lucena-Molina, J.J., Zadora, G., Aitken, C.G.G., Lunt, L., Champod, C., Biedermann, A., Hicks, T.N. and Taroni, F. (2015). *ENFSI guideline for evaluative reporting*

in forensic science. European Network of Forensic Science Institutes. Available at: http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf [Accessed: 1 January 2018]

Zhang, C. and Enzinger E. (2013). Fusion of multiple formant-trajectory- and fundamental-frequency-based forensic-voice-comparison systems: Chinese /ei1/, /ai2/, and /iau1/. In: *Proceedings of the 21st International Congress on Acoustics (ICA), Proceedings of Meetings on Acoustics*, volume 19, paper 060044. Available at: http://dx.doi.org/10.1121/1.4798793

Zhang, C., Morrison, G.S., Enzinger, E. and Ochoa F. (2013). Effects of telephone transmission on the performance of formant-trajectory-based forensic voice comparison – female voices. *Speech Communication*, 55, 796–813. Available at: http://dx.doi.org/10.1016/j.specom.2013.01.011

Table 1. Examples of verbal expressions of likelihood ratios and corresponding ranges of numeric likelihood ratio values in the 2015 ENFSI guideline on evaluative reporting.

Verbal Expression			Range of Values
The forensic findings are	slightly more	probable given one proposition relative to the other.	2 – 10
	more		10 – 100
	appreciably more		100 – 1000
	much more		1000 – 10,000
	far more		10,000 – 1 million
	exceedingly more		1 million +

Table 2. List of input and output possibilities and corresponding correctness for a system which outputs either "same-speaker" or "different-speaker".

		output	
		same-speaker	different-speaker
input	same-speaker	✓ hit	✗ miss
	different-speaker	✗ false alarm	✓ correct rejection

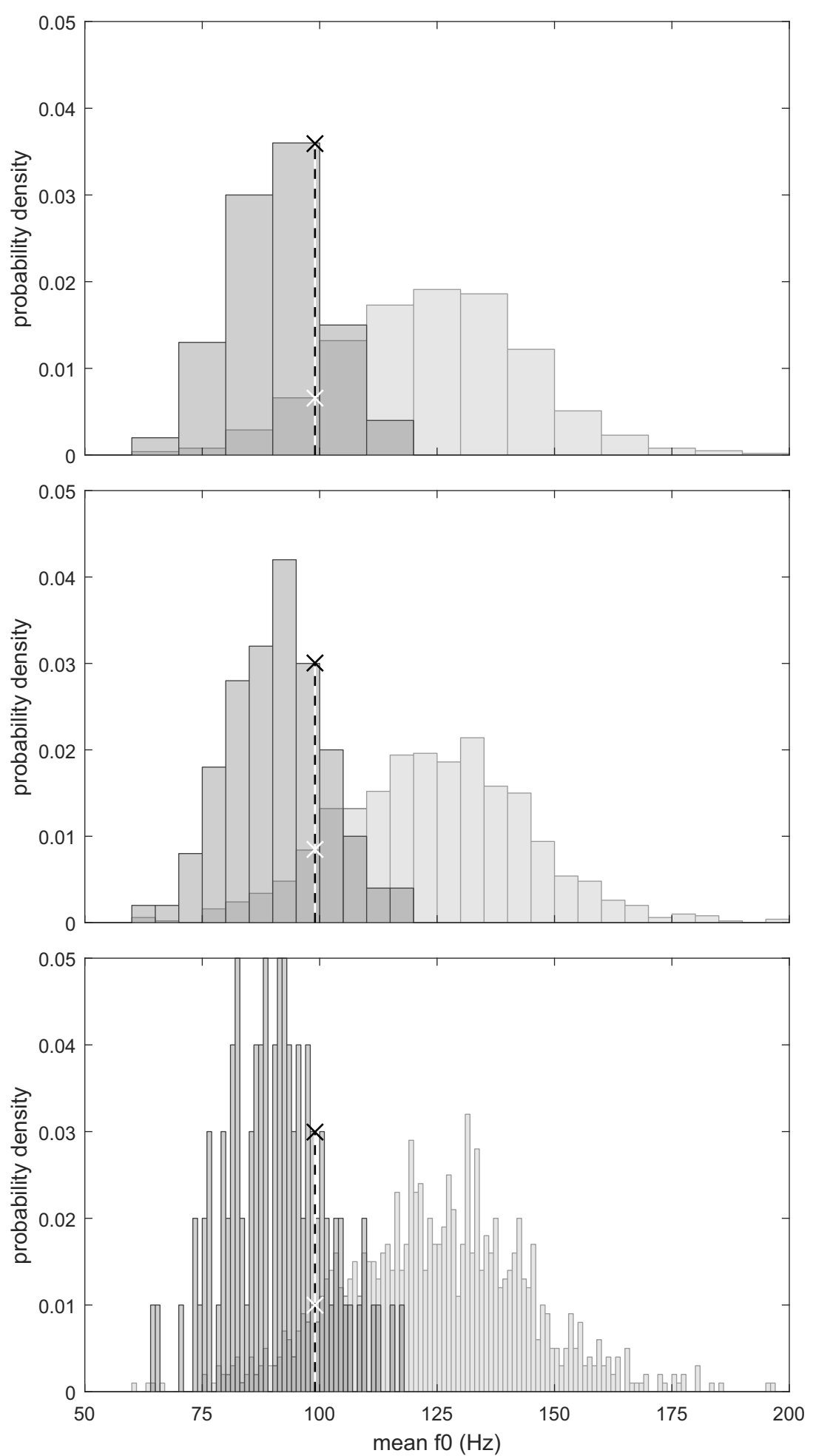
Figure 1. Histogram models for likelihood ratio calculation. The same data are modeled using histograms with different rectangle widths, i.e., from top to bottom: 10 Hz, 5 Hz, 1 Hz.

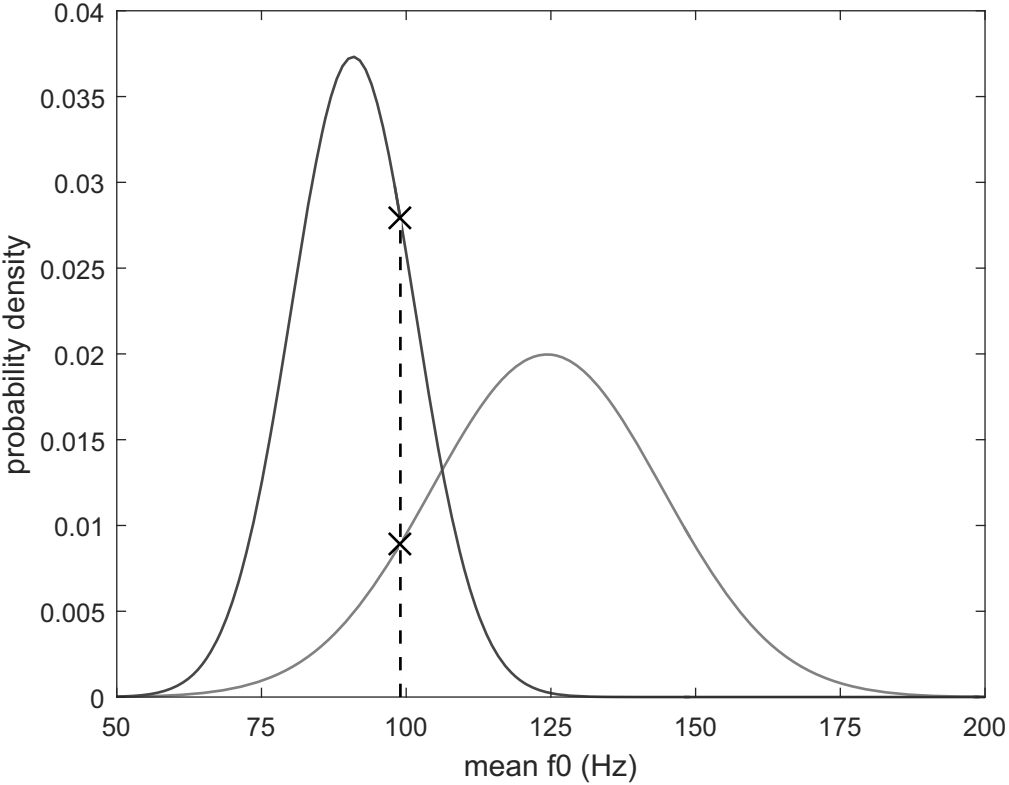
Figure 2. Gaussian distributions for likelihood ratio calculation. The distributions were fitted to the same data as represented by the histograms in Figure 1.

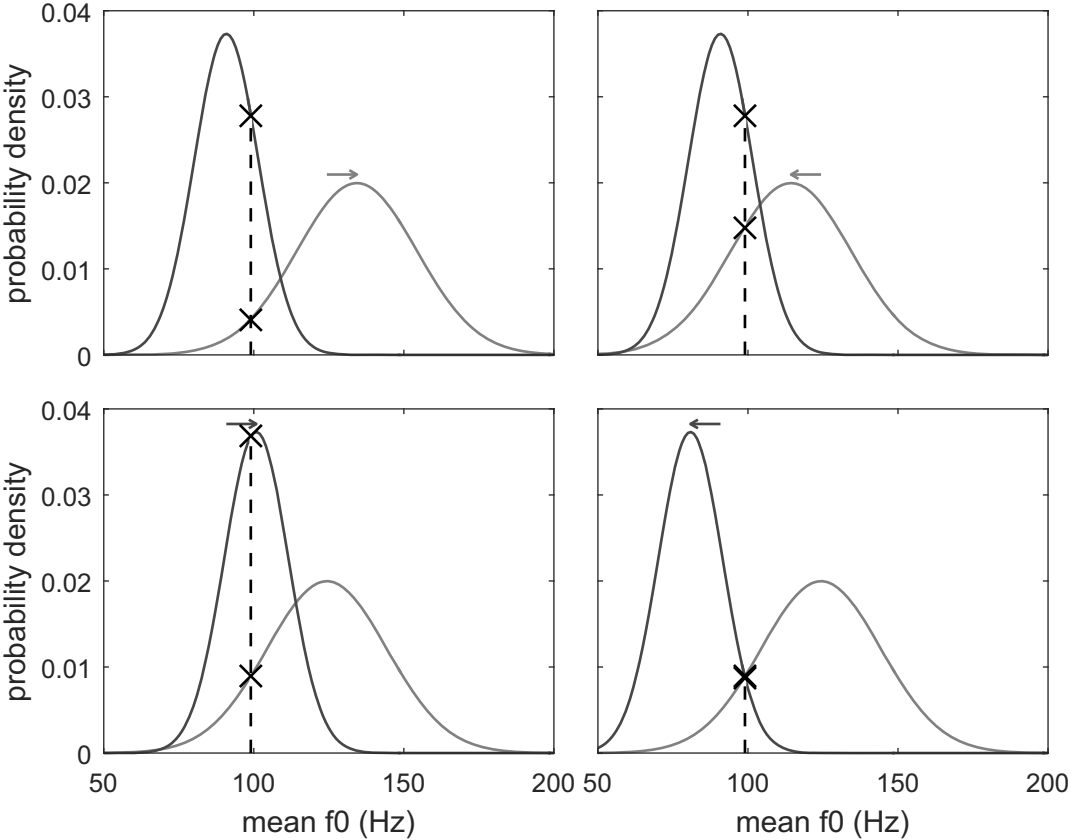
Figure 3. Gaussian distributions for likelihood ratio calculation. The distributions are shifted relative to Figure 2 in order to represent different degrees of typicality (top row) and different degrees of similarity (bottom row).

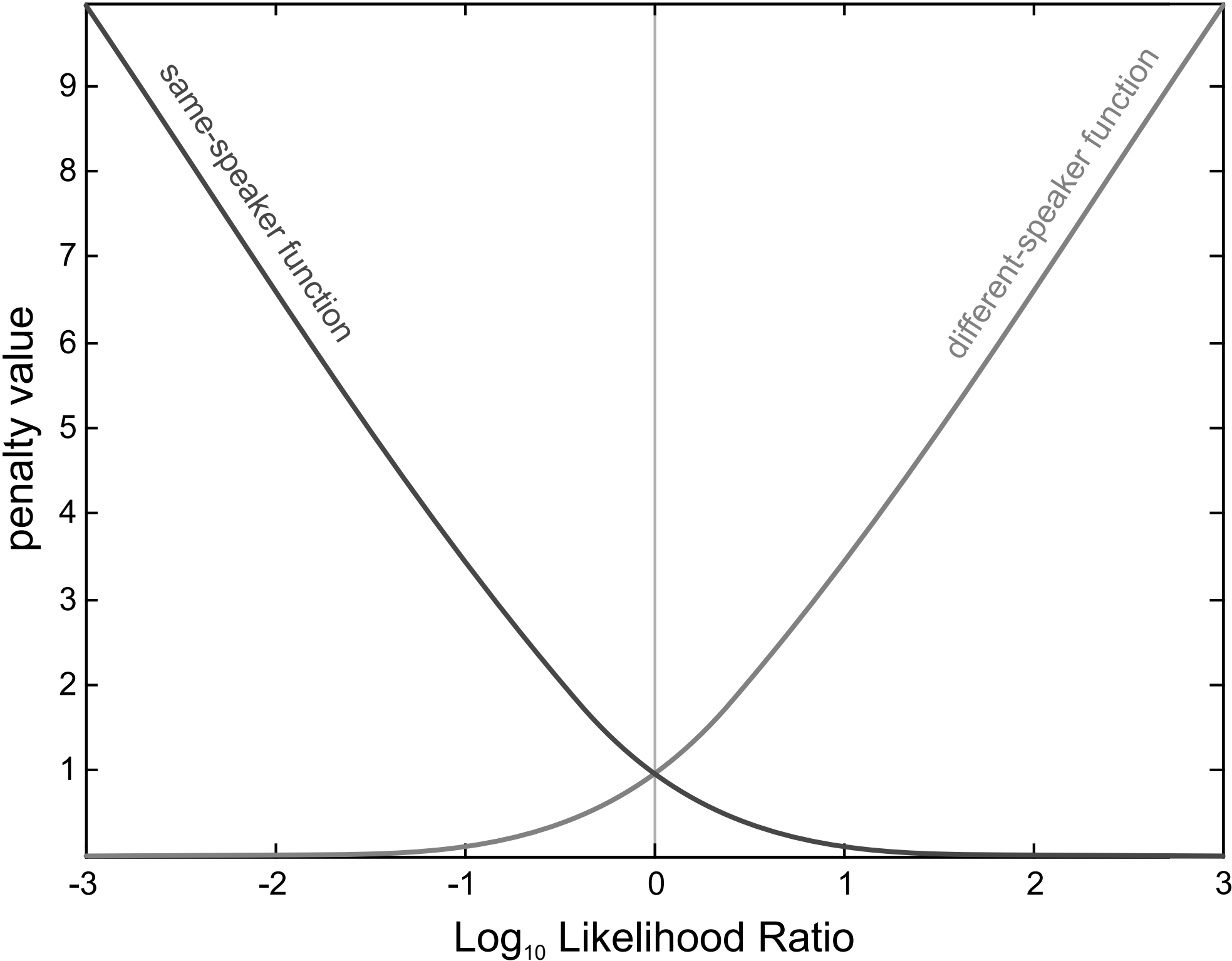
Figure 4. C_{llr} penalty functions.

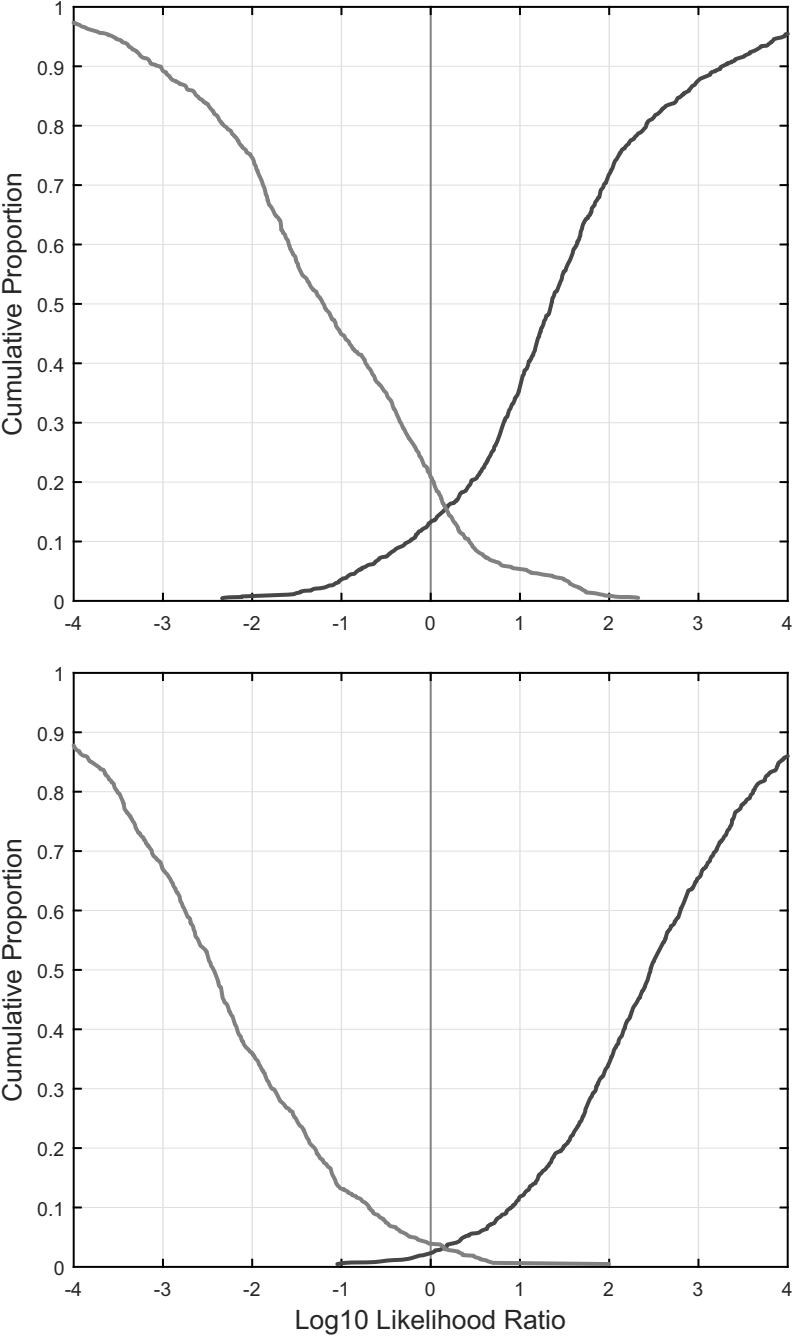
Figure 5. Example Tippett plots.

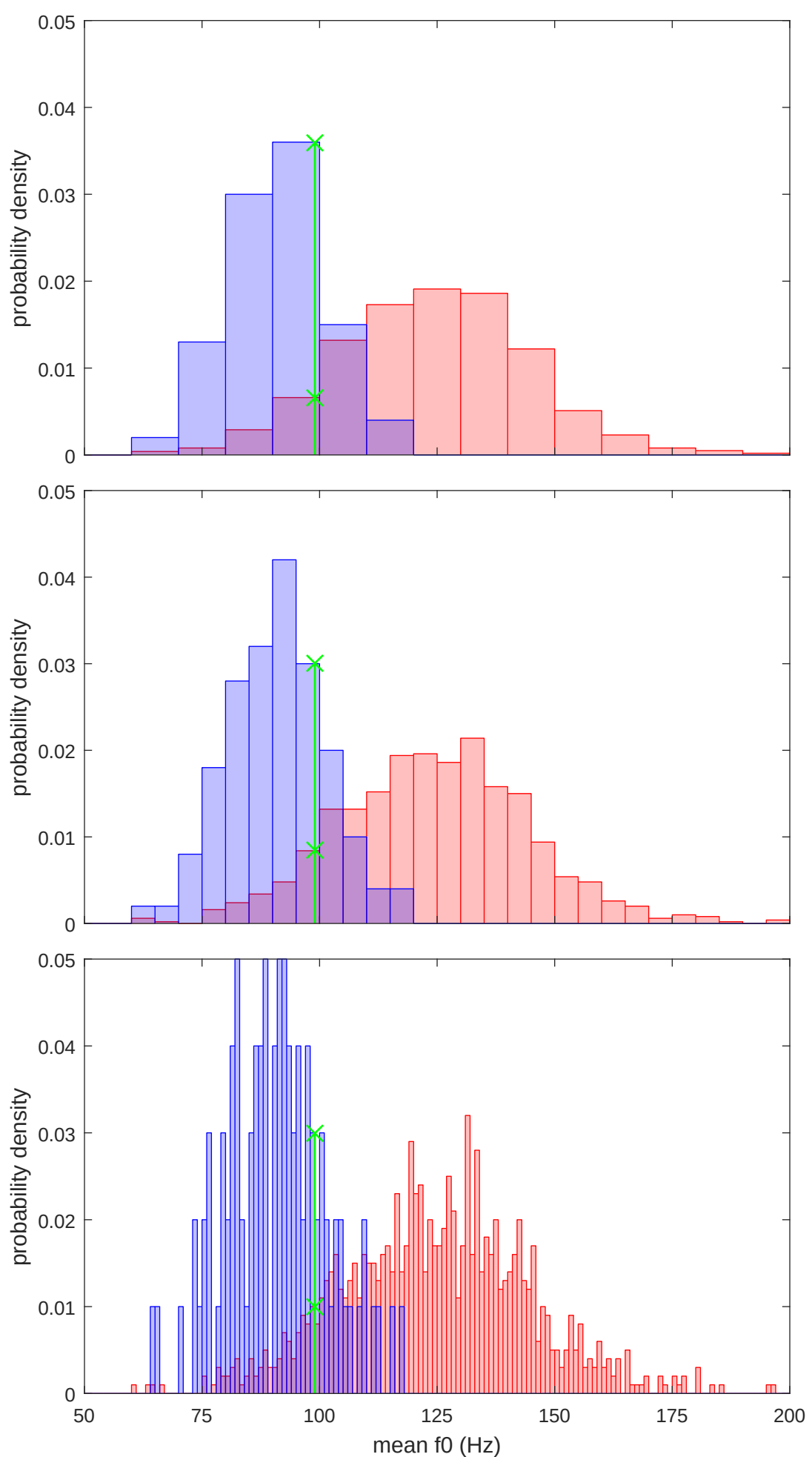

probability density
0.05
0.04
0.03
0.02
0.01
0
probability density
0.05
0.04
0.03
0.02
0.01
0
probability density
0.05
0.04
0.03
0.02
0.01
0
50
75
100
125
150
175
200
mean f0 (Hz)

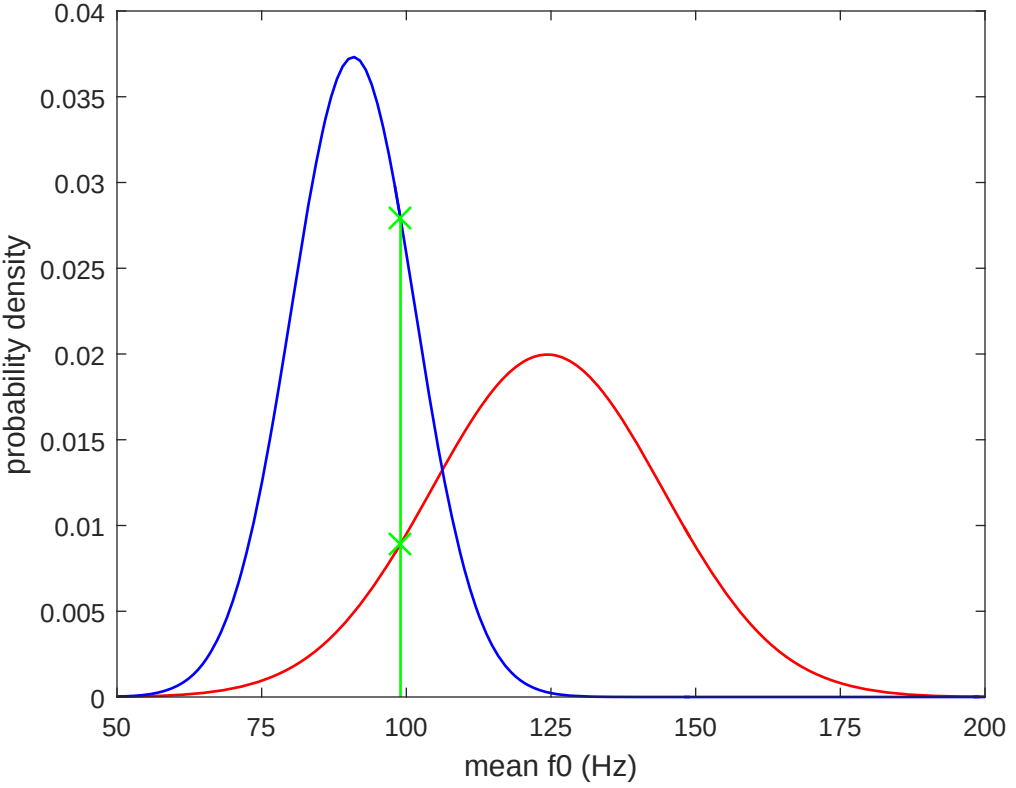

0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
probability density
50
75
100
125
150
175
200
mean f0 (Hz)

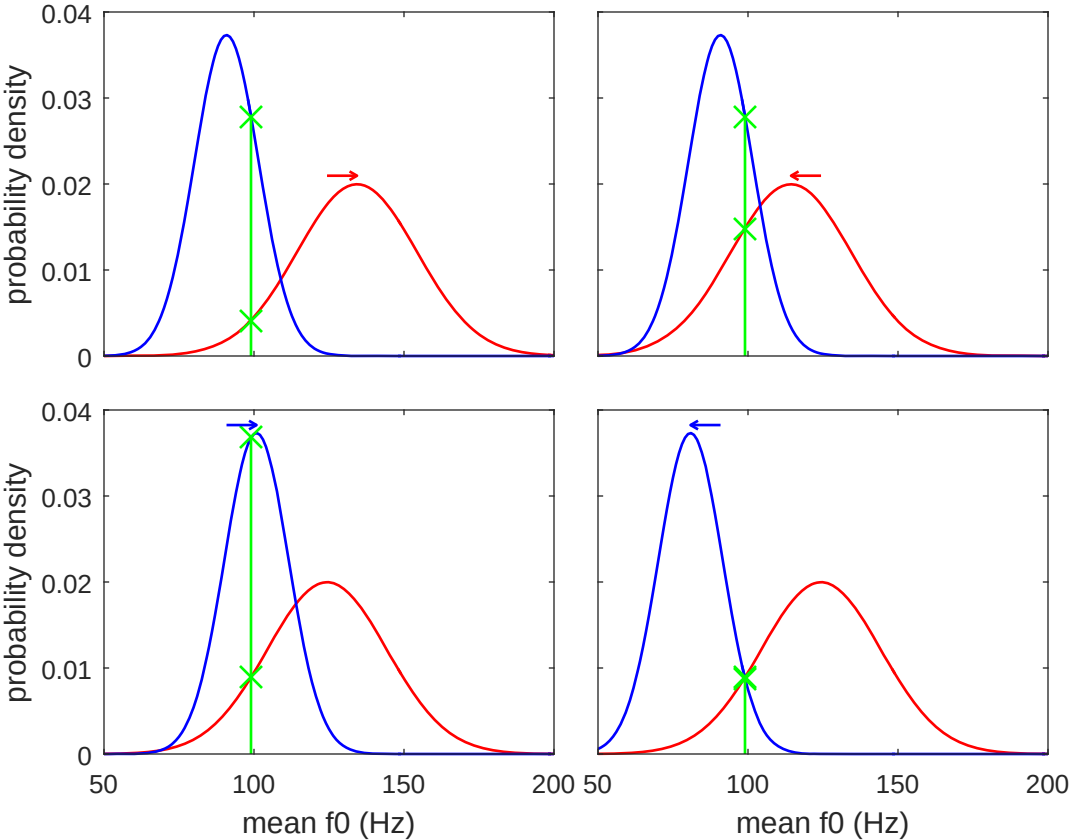

probability density
0.04
0.03
0.02
0.01
0
probability density
0.04
0.03
0.02
0.01
0
50
100
150
200
mean f0 (Hz)
50
100
150
200
mean f0 (Hz)

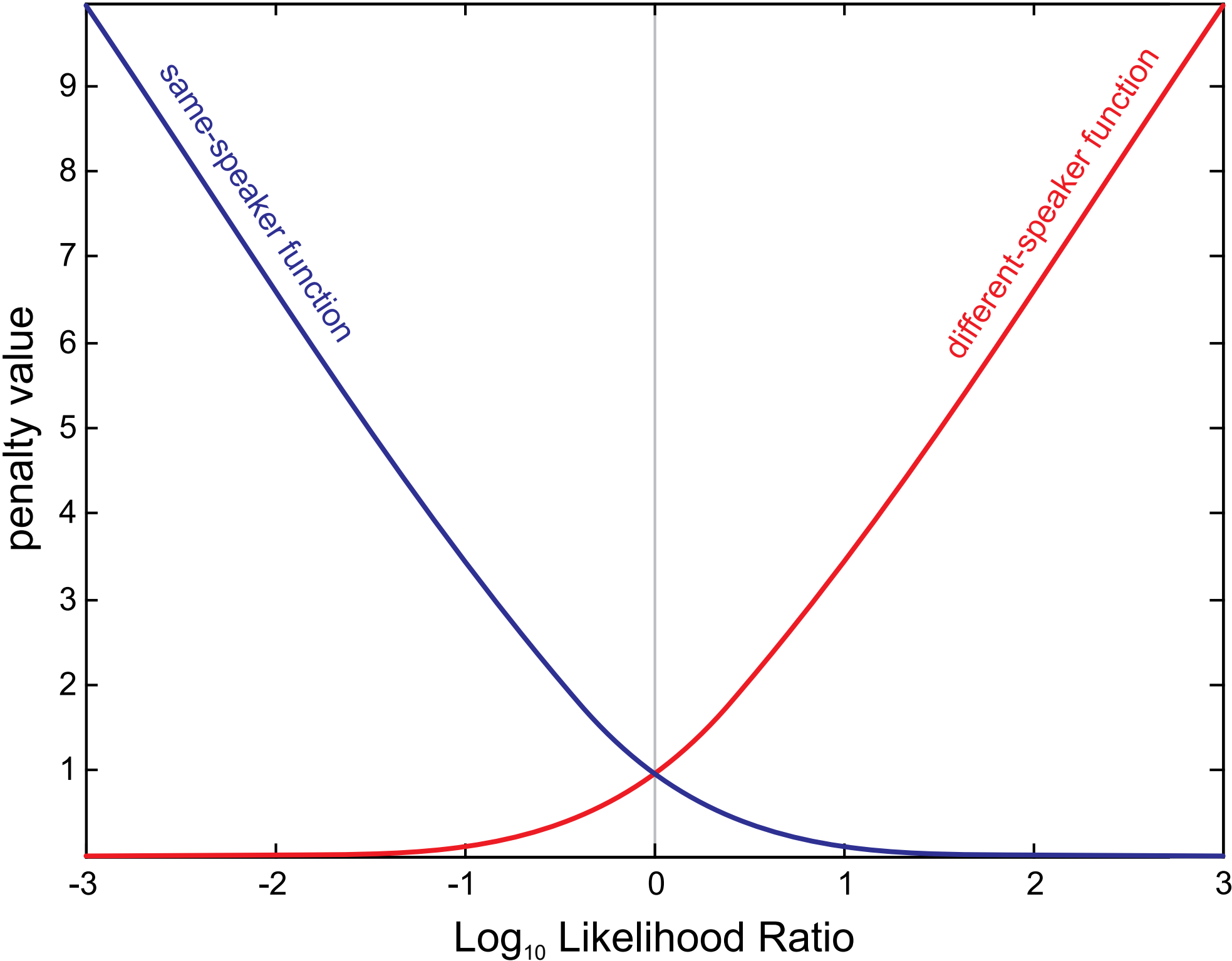

same-speaker function
different-speaker function
penalty value
9
8
7
6
5
4
3
2
1
-3
-2
-1
0
1
2
3
Log10 Likelihood Ratio

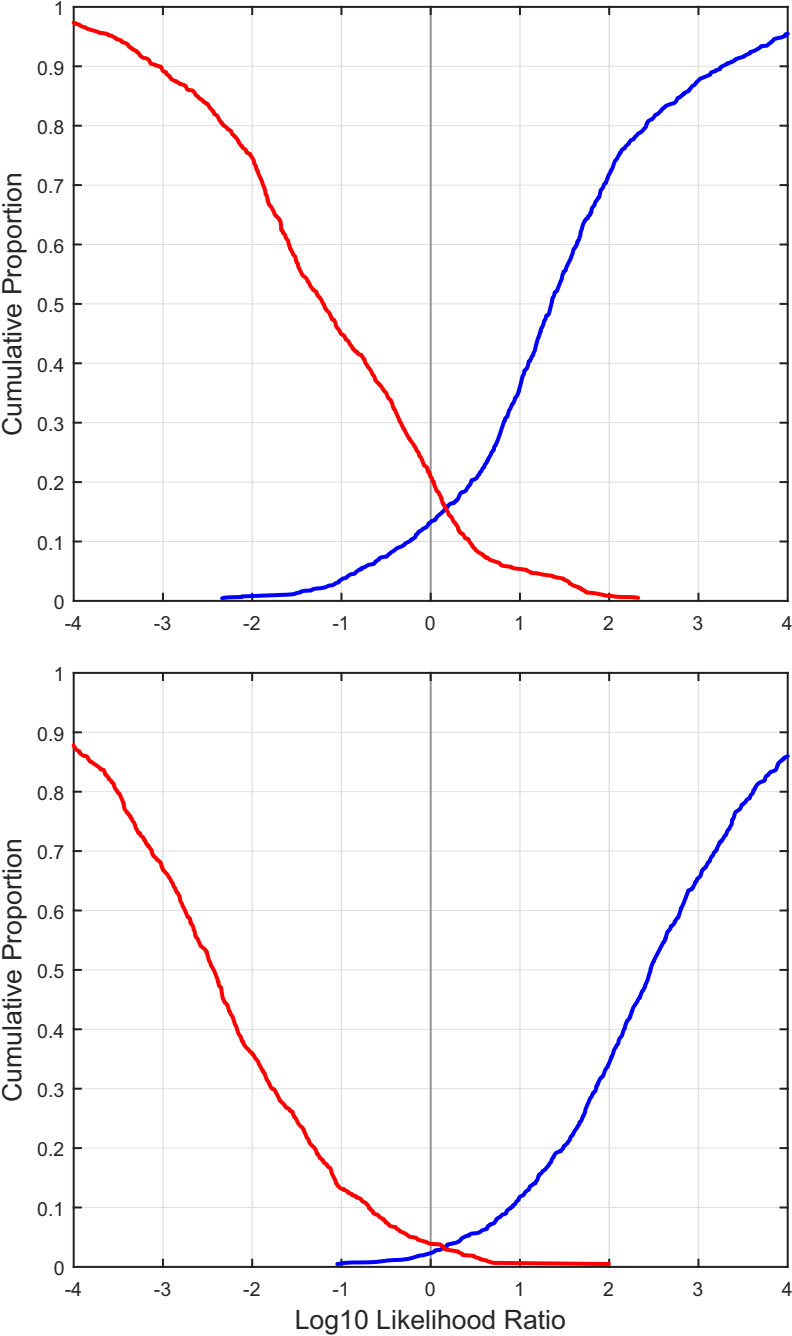

Cumulative Proportion
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
-4
-3
-2
-1
0
1
2
3
4
Cumulative Proportion
Log10 Likelihood Ratio


probability density
0.05
0.04
0.03
0.02
0.01
0
probability density
0.05
0.04
0.03
0.02
0.01
0
probability density
0.05
0.04
0.03
0.02
0.01
0
50
75
100
125
150
175
200
mean f0 (Hz)


0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
probability density
50
75
100
125
150
175
200
mean f0 (Hz)


probability density
0.04
0.03
0.02
0.01
0
probability density
0.04
0.03
0.02
0.01
0
50
100
150
200
mean f0 (Hz)
50
100
150
200
mean f0 (Hz)


same-speaker function
different-speaker function
penalty value
9
8
7
6
5
4
3
2
1
-3
-2
-1
0
1
2
3
Log10 Likelihood Ratio


Cumulative Proportion
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
-4
-3
-2
-1
0
1
2
3
4
Cumulative Proportion
Log10 Likelihood Ratio