This is a preprint of:
Morrison G.S., Zhang C. (2023). Forensic voice comparison: Overview. In Houck M., Wilson L., Eldridge H., Lewis S., Lothridge K., Reedy P. (Eds.), Encyclopedia of Forensic Sciences (3rd Ed.), vol. 2, pp. 737–750. Elsevier. https://doi.org/10.1016/B978-0-12-823677-2.00130-6
Preprint at http://forensic-voice-comparison.net/encyclopedia/

Forensic voice comparison – Overview

Authors

Geoffrey Stewart Morrison

Forensic Data Science Laboratory, Aston University
Forensic Evaluation Ltd
geoff-morrison@forensic-evaluation.net

Cuiling Zhang

Forensic Data Science Laboratory, Aston University
School of Criminal Investigation, Southwest University of Political Science and Law
cuiling-zhang@forensic-voice-comparison.net

Keywords

[10–15 keywords, listed alphabetically]

admissibility

automatic speaker recognition

forensic phonetics

forensic speaker comparison

forensic speaker identification

forensic speaker recognition

forensic speech science

forensic voice comparison

likelihood ratio

validation

voicegram

voiceprint

Abstract

[50–100 words]

In forensic voice comparison, a forensic practitioner analyzes a recording of a speaker of questioned identity and one or more recordings of a speaker of known identity, and compares the analytical results in order to draw an inference that will assist a legal-decision maker to decide whether the recordings are of the same speaker or of different speakers. This entry provides an overview of analytical approaches (including auditory, spectrographic, acoustic-phonetic, and automatic) and interpretive frameworks (including the likelihood-ratio framework) that have been used in forensic voice comparison. It also briefly discusses legal admissibility and validation of forensic voice comparison.

Key points

[short bulleted list of key points]

- analytical approaches for extracting information:

    - auditory-acoustic-phonetic
    - auditory-spectrographic
    - acoustic-phonetic-statistical
    - human-supervised-automatic
- interpretive frameworks for drawing inferences from the analytical results:
    - categorical-opinion
    - posterior-probability
    - likelihood-ratio
    - UK
- legal admissibility
- validation

Acknowledgements

The writing of this entry was supported by Research England's Expanding Excellence in England Fund as part of funding for the Aston Institute for Forensic Linguistics 2019–2023.

1 Introduction

In forensic voice comparison, a forensic practitioner analyzes a recording of a speaker of questioned identity and one or more recordings of a speaker of known identity, and compares the analytical results in order to draw an inference that will assist a legal-decision maker to make a decision as to whether the recordings are of the same speaker or of different speakers.

Forensic voice comparison is challenging because, although there is variability in the properties of voices between speakers, individual speakers also have natural within-speaker variability, recording quality is often poor, there are often mismatches in speaking style and recording conditions between the questioned-speaker recording and the known-speaker recording, and the speaking styles and conditions vary from case to case. Common speaking styles include normal vocal effort and raised vocal effort. Common conditions include different types and volumes of background noise, and processing of the recordings using different lossy codecs to reduce the amount of data transmitted over communications systems or saved to storage media. Hansen & Bořil (2018) provide a taxonomy of sources of speaker-intrinsic and speaker-extrinsic variability that can potentially affect recordings used in forensic voice comparison. Differences between the properties of the voices on different recordings could be due to within-speaker variability or between-speaker variability. The forensic practitioner has to assess the relative probabilities of obtaining the observed properties of the voices on the questioned-speaker and known-speaker recordings if they were produced by the same speaker versus if they were produced by different speakers.

The process of evaluation of strength of forensic evidence consists of analysis (i.e., extraction of information from items of interest) and interpretation (i.e., drawing inferences with respect to the meaning of the information extracted by the analysis). To some extent, analytical methods and interpretive methods are independent of one another, but there are often correlations or even causations, with particular interpretive frameworks being more commonly used with particular analytical approaches.

Below, different analytical approaches and different interpretive frameworks that have been used for forensic voice comparison are described. Also covered are the popularity of different approaches and frameworks, legal admissibility, and validation of forensic voice comparison. Some material has been adapted from Morrison & Thompson (2017), Morrison, Enzinger, Zhang (2018), and Morrison & Enzinger (2019a). The latter publications provide more detailed coverage of some topics and also cover other topics related to forensic voice comparison and to forensic speech science more broadly.

2 Analytical approaches

There are four basic analytical approaches used for forensic voice comparison: auditory, spectrographic, acoustic phonetic, and human-supervised automatic. Certain combinations of approaches are common. In the following subsections, the following basic and commonly-combined approaches are described:

- auditory and auditory-acoustic-phonetic approaches
- spectrographic and auditory-spectrographic approaches
- acoustic-phonetic-statistical approach
- human-supervised-automatic approach

2.1 Auditory and auditory-acoustic-phonetic approaches

In the auditory approach, the practitioner listens to the questioned-speaker and known-speaker recordings. They listen for similarities which they would expect to hear if the two recordings consisted of speech from the same speaker, but which they would not expect to hear if the recordings consisted of speech from different speakers. They also listen for differences which they would expect to hear if the two recordings consisted of speech from different speakers, but which they would not expect to hear if the recordings consisted of speech from the same speaker. They may listen to any or all of the pronunciation of particular vowel sounds and of particular consonant sounds, the pronunciation of particular words or phrases, and other more global properties such as intonation patterns.

Practitioners of the auditory approach will usually make use of software tools which allow them to listen to short sections of speech from each recording, one immediately after the other. They will also usually have training in auditory phonetics, including training in using a phonetic alphabet to transcribe the speech sounds they hear. A phonetic transcription allows a practitioner to document the details of what they hear.

To document the results of their analyses of whole recordings, practitioners of the auditory approach may use tables of presence or absence, or degree, of particular speech properties that they hear, or of counts of occurrences of particular realizations of speech sounds that they hear (Hollien et al., 2016; San Segundo et al., 2019; Gurlekian et al., 2022).

Most practitioners of auditory approaches listen only to the questioned-speaker and known-speaker recordings. Some practitioners, however, 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. In an approach known as blind grouping (Cambier-Langeveld et al., 2014), one practitioner prepares recordings of the questioned speaker, the known speaker, and several foil speakers. This may involve cutting each original recording into multiple shorter recordings. The first practitioner must take care in selecting foil-speaker recordings so that the questioned-speaker and known-speaker recordings do not stand out relative to the foil-speaker recordings because of speaking style, linguistic content, or recording conditions. The first practitioner presents the recordings to a second practitioner without telling the second practitioner the origin of each recording or how many speakers there are in total. The second practitioner then attempts to group the recordings by speaker. Whether the second practitioner groups the questioned-speaker and known-speaker recordings together constitutes the result of the analysis. The correctness of the grouping of foil speakers serves as a test of performance.

The auditory approach is commonly combined with the acoustic-phonetic approach, leading to the auditory-acoustic-phonetic approach. In the acoustic-phonetic approach the practitioner uses software tools to make quantitative measurements of acoustic properties of parts of the voice recordings. Measurements may be made on particular speech sounds that occur in both the questioned-speaker and known-speaker recordings. The particular sections of the recording containing the speech sounds of interest are usually manually selected. The types of measurements made are generally the same as the types of measurements made in acoustic phonetics, an area of research which studies the transmission of human speech through the air between the speaker's vocal tract and the listener's ear. Practitioners of the acoustic-phonetic approach usually have training in acoustic phonetics.

An example of properties commonly measured in the acoustic-phonetic approach are vowel formants, which are the resonances of the vocal tract. Individuals with longer vocal tracts have lower resonances than those with shorter vocal tracts. The length of the vocal tract can vary from person to person, but when speaking a person changes the length and shape of their vocal tract to produce a range of different resonance frequencies. The differences between vowel sounds such as "ee," "oo," and "ah" are due to different resonances resulting from the speaker moving their tongue, jaw, lips, etc. to make different vocal-tract shapes. How speakers pronounce vowel sounds, and hence the formant frequencies produced, can vary from speaker to speaker, but how an individual speaker pronounces vowel sounds can also vary from instance to instance. Another commonly made measurement is fundamental frequency, which is the acoustic correlate of what listeners perceive as pitch. Whereas formants are related to the length and shape of the vocal tract, fundamental frequency is related to the size of a speaker's vocal folds and the configuration in which they hold and put tension on their vocal folds. Fundamental frequency varies both between and within speakers.

In the auditory-acoustic-phonetic approach, the acoustic measurements that practitioners make are usually collected in tables or used to make plots, e.g., the measured formant frequencies of instances of vowels are plotted in 2D scatterplots with the frequencies of the first formant on one axis and second formant on the other axis, see Figure 1. (In Figure 1 the vowels were clearly spoken, were in the same phonetic context, and were recorded in a sound booth – greater within-speaker variability would be expected in forensically realistic conditions.)

<Figure 1 near here>

Figure 1. Example 2D scatterplot of mean first- and second-formant frequencies (F1 and F2) of 10 instances of the vowel in the word "beep" spoken by each of two different male speakers of Western-Canadian English (upward and downward pointing triangles), plus 10 instances from each of 46 other Western-Canadian-English speakers (circles).

French et al. (2010) pp. 146–147 provided a list of features considered in the auditory-acoustic-phonetic approach:

> 1. Vocal setting and voice quality ... with up to 38 individual elements to be considered.
>
> 2. Intonation ...
>
> 3. Pitch, measured as average and variation in fundamental frequency.
>
> 4. Articulation rate.
>
> 5. Rhythmical features.
>
> 6. Connected speech processes such as patterns of assimilation and elision.
>
> 7. A large set of consonantal features, including energy loci of fricatives and

plosive bursts, durations of nasals, liquids, and fricatives in specific phonological environments, voice onset time of plosives, presence/absence of (pre-)voicing in lenis plosives, and discrete sociolinguistic variables.

8. A large set of vowel features, including acoustic patterns such as formant configurations, centre frequencies, densities, and bandwidths, and auditory qualities of sociolinguistic variables.

9. Higher-level linguistic information including use and patterning of discourse markers, lexical choices, morphological and syntactic variants, pragmatic behaviour such as turn-taking and telephone call opening habits, aspects of multilingual behaviour such as codeswitching.

10. Evidence of speech impediment, voice and language pathology.

11. Non-linguistic features characteristic of the speaker, for example patterns of audible breathing, throat-clearing, tongue clicking, and both filled and silent hesitation phenomena.

Practitioners of the auditory-acoustic-phonetic approach rely on their training and experience to make judgments as to whether the properties of the speech they hear and acoustic-phonetic measurements they read in their tables or see in their plots would be more likely to occur if the questioned-speaker and known-speaker recordings were recordings of the same speaker or if they were recordings of different speakers. French et al. (2010) p. 144:

> forensic phoneticians ... judge the distinctiveness of the features found in the criminal and suspect samples, and ... [make a] comparison with a broader population, ... informally via the analyst's experience and general linguistic knowledge rather than formally and quantitatively.

Further descriptions of auditory and auditory-acoustic-phonetic approaches can be found in Jessen (2018, 2021) and in Hudson et al. (2021). The European Network of

Forensic Science Institutes (ENFSI) is preparing a guideline on the use of auditory-acoustic-phonetic approaches (Wagner et al., 2021). Criticisms of auditory-acoustic-phonetic practices can be found in Morrison (2014, 2018a).

2.2 Spectrographic and auditory-spectrographic approaches

In the spectrographic approach, the practitioner takes parts of the audio recordings (typically words or phrases) and converts them into pictures. These pictures are called spectrograms. Spectrograms represent time on the *x* axis, frequency on the *y* axis, and intensity as the darkness of a monochrome scale or as the color in a colormap scale. Examples of monochrome spectrograms are provided in Figure 2. The practitioner looks at spectrograms derived from the questioned-speaker recording and spectrograms derived from the known-speaker recording, and may also look at spectrograms derived from recordings of foil speakers. They look at the spectrograms in search of similarities which they would expect to see if the two recordings were of the same speaker but not if they were of different speakers, and also in search of differences they would expect to see if the two recordings were of different speakers but not if they were of the same speaker.

<Figure 2 near here>

Figure 2. Examples of monochrome spectrograms showing the word “beifu” (a proper name) spoken by two different male speakers of Standard Chinese. Left panels: two different instances of “beifu” spoken by the first speaker. Right panels: two different instances of “beifu” spoken by the second speaker. Top panels: recordings made using a digital recorder in a quiet room. Bottom panels: recordings of telephone calls made between landline and mobile telephones.

In contrast to other approaches, in which practitioners usually use existing recordings of the known speaker (e.g., recordings of police interviews), practitioners of the spectrographic approach usually make new recordings of the known speaker in which the known speaker is required to say the same words as appear on the questioned-speaker recording and in the same manner as they were said on the questioned-speaker recording. This is required by published protocols.

In the United States, protocols for performing auditory-spectrographic forensic voice comparison were developed by the Federal Bureau of Investigation (FBI), and by the International Association of Voice Identification (IAVI), which later became part of the International Association for Identification (IAI), and later split to become the American Board of Recorded Evidence (ABRE), which was part of the American College of Forensic Examiners Institute (ACFEI). The FBI ceased using the auditory-spectrographic approach in 2011, in favor of the human-supervised-automatic approach (Archer, 2012). The IAI no longer promulgates forensic-voice-comparison protocols, and ACFEI is no longer in operation (Balko, 2017).

ABRE (1999) §7.1.5 required the examiner to visually compare on spectrograms:

> a. General formant shaping and positioning ...
>
> b. Pitch striations ...
>
> c. Energy distribution ...
>
> d. Word length ...
>
> e. Coupling ...
>
> f. Other. Plosives, fricatives, and inter-formant features ... inhalation noise, repetitious throat clearing, or utterances like “um” and “uh” ...

And to auditorily compare:

> a. Pitch ...
>
> b. Intonation ...

c. Stress/Emphasis ...

d. Rate ...

e. Disguise ...

f. Mode ...

g. Psychological state ...

h. Speech defects. ...

i. Vocal quality ...

j. Other ... long-term fluctuations of pitch (vibrato), vocal fry (extremely low pitching), pitch breaks, and stuttering.

The IAI and ABRE protocols required the practitioner to state their conclusions as one of “Identification, Probable Identification, Possible Identification, Inconclusive, Possible Elimination, Probable Elimination, or Elimination”. For each of the levels on this conclusion scale, the protocols specified criteria such as the number of words that had to be examined and the number that had to “match” (Gruber & Poza, 1995 §60; ABRE, 1999 §7.3).

In contrast to the IAI and ABRE protocols, Poza & Begault (2005) recommended a gestalt approach. Gruber & Poza (1995) §65–67 and Poza & Begault (2005) recommended the use of recordings and spectrograms of foil speakers, which was not required by the IAI or ABRE protocols.

In the early 1970s, there was a debate about whether a visual only or a visual plus auditory approach was better, which ended with a preference for the latter, i.e., the auditory-spectrographic approach. From the late 1960s, the spectrographic and auditory-spectrographic approaches were highly controversial. By the end of the 1970s, their popularity in the United States began to decline, but they continued to be used into the early 2000s. Following a Daubert admissibility hearing, they were declared inadmissible in U.S. v. Angleton, 269 F.Supp 2nd 892 (S.D. Tex. 2003). Gruber & Poza (1995) §6 summarized the objections to the auditory-spectrographic approach as follows:

> (1) there is simply no adequate theoretical foundation to justify the procedures used in forensic voicegram identification; (2) the competency of forensic examiners, both in absolute terms and relative to laypersons who just listen to voices, is largely unknown; (3) the so called Tosi "Extrapolation," which turned the tide in favor of admissibility by generalizing from laboratory to real-world scenarios, is unproven and highly questionable; and (4) that to assert that the individual examiner's experience, combined with his competence and talent, should, in the end, override any concerns about the problems associated with subjective decision making is to make a very questionable assumption.

Further descriptions of spectrographic and auditory-spectrographic approaches can be found in Tosi (1979) and in National Research Council (1979). Criticisms of spectrographic and auditory-spectrographic approaches can be found in Gruber & Poza (1995) and in Morrison (2014).

2.3 Acoustic-phonetic-statistical approach

As previously mentioned, in the acoustic-phonetic approach the practitioner uses software tools to make quantitative measurements of acoustic properties of parts of the voice recordings. The difference between the auditory-acoustic-phonetic approach and the acoustic-phonetic-statistical approach is that, whereas in the former the values resulting from the measurements are interpreted by the subjective judgment of the practitioner, in the latter those values are input to statistical models.

The acoustic-phonetic approach has been combined with statistical models that calculate likelihood ratios. These require extracting acoustic-phonetic measurements not only from the questioned-speaker and known-speaker recordings, but also from recordings of speakers sampled from the relevant population (see the discussion of the likelihood-ratio framework below).

The acoustic-phonetic-statistical approach usually requires manual selection of the portions of the recordings to be measured, and often requires manual intervention in the process of making the measurements. In contrast, in the human-supervised-automatic approach (discussed below) these tasks are performed automatically. The acoustic-phonetic-statistical approach therefore usually has much higher human-labor costs than the human-supervised-automatic approach. Some methods use measurements of the type normally used in acoustic phonetics, but do so fully automatically. These could be classed as either acoustic-phonetic-statistical methods or as human-supervised-automatic methods. Here, we will group them with the former.

Speech properties that have been used in acoustic-phonetic-statistical systems include fundamental frequency and formant trajectories. The latter are the patterns of change in formant values over the time-course of individual vowels. Figure 3 provides examples of formant trajectories – each speaker said the Chinese word “iao” (the number one) multiple times as part of a task that required exchange of technical information over the telephone. The recordings however, were direct-microphone recordings made in sound booths. Parametric functions such as polynomials or discrete cosine transform (DCTs) can be fitted to each format trajectory and the coefficient values from the functions used as features in a statistical model.

<Figure 3 near here>

Figure 3. Examples of formant trajectories of multiple instances of the word “iao” (the number one) spoken by two different female speakers of Standard Chinese. Left panels: instances of “iao” spoken by the first speaker. Right panels: instances of “iao” spoken by the second speaker. Top panels: raw formant measurements. Bottom panels: smoothed time-normalized trajectories based on zeroth through second DCT

coefficients. Different colors indicate instances of "iao" spoken during different recording sessions that were approximately two weeks apart.

In direct comparisons under forensically-realistic conditions (or conditions approaching forensically-realistic conditions), human-supervised-automatic systems have been found to perform much better than acoustic-phonetic-statistical systems, and when the two are combined the improvement in performance over a human-supervised-automatic system alone is at-best negligible (Zhang et al. 2013; Enzinger & Morrison, 2017). An exception is presented in Franco-Pedroso & González-Rodríguez (2016), which achieved an increase in performance when combining an acoustic-phonetic system with an automatic system. Franco-Pedroso & González-Rodríguez (2016) extracted formant trajectories automatically, and, compared to the other studies, extracted them from more speech sounds in a much larger database. They also used a later generation of automatic-speaker-recognition technology.

Further descriptions of the acoustic-phonetic-statistical approach can be found in Rose (2017) and Jessen (2018). The European Network of Forensic Science Institutes (ENFSI) has published a guideline on the use of acoustic-phonetic-statistical and human-supervised-automatic approaches (Drygajlo et al., 2015).

2.4 Human-supervised-automatic approach

The human-supervised-automatic approach makes use of automatic-speaker-recognition technology, which also has non-forensic applications. Software tools are used to make measurements of the acoustic properties of the questioned-speaker and known-speaker recordings, and of recordings of other speakers sampled from the relevant population for the case. Acoustic measurements are usually made over the whole of the speech of the speaker of interest in a recording, and there is usually no focus on individual speech sounds, words, or phrases. The types of measurements made are the same as those used in speech processing in general, which includes applications such as automatic speech recognition. An example of a common type of measurement is a vector of mel-frequency cepstral coefficients (MFCCs). MFCC vectors consist of a set of numbers, e.g., 14 numbers, which describe the frequency components (the spectrum) of the speech during a short interval of time, e.g., 20 ms. MFCC vectors are usually extracted every few milliseconds, e.g., every 10 ms with a 50% overlap between adjacent 20 ms long intervals. An MFCC vector provides more detailed measurements of the acoustic spectrum of a speech signal than do acoustic-phonetic measurements such as fundamental frequency plus two or three formants.

The measurements made in the human-supervised-automatic approach are automatically entered into statistical models. The combination of the human-supervised-automatic approach and the likelihood-ratio framework is common, so these models usually calculate likelihood ratios. Human-supervised-automatic systems based on state-of-the-art automatic-speaker-recognition technology use deep neural networks (DNNs). DNNs are trained using data from at least tens of thousands of recordings of at least thousands of speakers. The speakers are diverse and the recordings of each speaker are diverse in speaking styles and recordings conditions. The DNN therefore learns about both within-speaker and between-speaker variability. The MFCC vectors (or other measurements) from a recording are presented to the DNN, and vectors of numbers known as x-vectors (or more generically, DNN embeddings), are extracted from the DNN – they are the activation levels of the nodes in a pre-final layer of the DNN. One x-vector is extracted for each recording. x-vectors are the same length irrespective of the length of the recording. Different recordings of the same speaker tend to cluster together in the x-vector space, whereas recordings of different speakers tend to be separated from each other in the x-vector space. x-vectors are extracted from the questioned-speaker and known-speaker recordings, and these are input to backend models that calculate likelihood ratios. The x-vectors used to train or adapt the backend models must come from recordings of speakers who are representative of the relevant population for the case, and those recordings must reflect the conditions of the questioned-speaker and known-speaker recordings in the case, including any mismatch between the conditions of the questioned-speaker and known-speaker recordings. Compared to earlier generations of automatic-speaker-recognition technology, x-vector systems perform substantially better under forensically realistic conditions (Morrison & Enzinger, 2019b).

Further descriptions of the human-supervised-automatic approach can be found in Morrison et al. (current volume) and in Morrison et al. (2020). A bibliography of essential scientific literature for human-supervised-automatic approaches has been published by the Speaker Recognition Subcommittee of the Organization of Scientific Area Committees for Forensic Science (OSAC SR, 2021). For criticisms of misuses of the human-supervised-automatic approach, see Morrison & Thompson (2017) and Morrison (2018a, 2018b).

3 Interpretive frameworks

Interpretive frameworks that have been used for forensic voice comparison include:

- categorical-conclusion framework
- posterior-probability framework
- likelihood-ratio framework
- UK framework

Each is discussed in its own subsection below.

3.1 Categorical-conclusion framework

Some practitioners state categorical conclusions that the voices on the questioned-speaker and known-speaker recordings either come from the same speaker or that they come from different speakers. Categorical conclusions are reached on the basis of subjective judgment, and require the practitioner to have reached a sufficient degree of certainty in their choice of conclusion. There is no objective threshold for what constitutes a sufficient degree of certainty, it is a subjective judgment on the part of the practitioner.

Categorical conclusions are often expressed as “identification” or “exclusion”, or, if the practitioner has not reached a sufficient degree of certainty, as “inconclusive”. “Identification” or “same speaker” and “exclusion” or “different speaker” are extreme cases of verbal posterior probabilities corresponding to 1 and 0 respectively. Logically, if the probability is 1 then no other evidence such as the testimony of an alibi witness could outweigh it. If the probability is 0 then no other evidence such as the testimony of an eyewitness to the crime could outweigh it. If no other evidence could outweigh the evidence presented by the forensic practitioner, then logically the forensic practitioner would have made a definitive decision on the issue of identity. That is a decision which should be made by the legal-decision maker after weighing all the evidence presented to them, it should not be made by a forensic practitioner. The task of the forensic practitioner is to evaluate and express an opinion associated with the one piece of evidence that they were asked to evaluate.

3.2 Posterior-probability framework

Some practitioners state conclusions as numeric posterior probabilities with respect to a single hypothesis, e.g., there is a 95% probability that the voice on the questioned-speaker recording was produced by the known speaker. A numeric posterior probability could be output by a statistical model, but could also be the result of a subjective judgment made by a practitioner. As already mentioned, the IAI and ABRE protocols for the auditory-spectrographic approach required the practitioner to state their conclusion as one of “Identification, Probable Identification, Possible Identification, Inconclusive, Possible Elimination, Probable Elimination, or Elimination” (Gruber & Poza, 1995 §60; ABRE, 1999 §7.3). These are verbal expressions of posterior probabilities.

As a matter of logic, posterior probabilities cannot be derived solely via comparison of the properties of the questioned-speaker and known-speaker recordings. Logically, according to Bayes' Theorem, a posterior probability must be the result of combining a likelihood ratio with a prior probability. Even if they are not aware of it, a forensic practitioner who presents a posterior probability must have at least implicitly used a prior probability. The prior probability will be either arbitrary or based on other information. The choice of prior probability will affect the value of the posterior probability, but this will be hidden from the legal-decision maker if only the posterior probability is presented. A practitioner who arbitrarily chooses a low prior probability will present a lower posterior probability, and a practitioner who arbitrarily chooses a high prior probability will present a higher posterior probability, but the legal-decision maker will be mislead into thinking that the difference is related to a difference in the evidence, the properties of the questioned-speaker and known-speaker recordings. The task of the forensic practitioner is to evaluate and express an opinion associated with the one piece of evidence that they were asked to evaluate. They should not consider other information. Considering all the information presented during a trial is the task of the legal-decision maker. If a forensic practitioner's conclusion took account of other information that the legal-decision maker had already taken account of, but the legal-decision maker thought that the forensic practitioner was presenting new independent information, then the legal-decision maker would double count that information.

3.3 Likelihood-ratio framework

The likelihood-ratio framework is advocated as the logically correct framework for evaluation of evidence by the vast majority of experts in forensic inference and statistics, including Aitken et al. (2011) and Morrison et al. (2017), with 31 and 19 signatories respectively. Its use is also advocated by key organizations including:

- Association of Forensic Science Providers of the United Kingdom and of the Republic of Ireland (2009)
- Royal Statistical Society (Aitken et al., 2010)

- European Network of Forensic Science Institutes (Willis et al., 2015)
- National Institute of Forensic Science of the Australia New Zealand Policing Advisory Agency (Ballantyne et al., 2017)
- American Statistical Association (Kafadar et al., 2019)
- Forensic Science Regulator for England & Wales (2021)

The likelihood-ratio framework requires assessment of the probability of obtaining the evidence, E, if one hypothesis, H_1, were true versus the probability of obtaining the evidence, E, if an alternative hypothesis, H_2, were true, see Equation (1), in which Λ is the likelihood ratio.

(1)

$$\Lambda = \frac{p(E|H_1)}{p(E|H_2)}$$

The two hypotheses must be mutually exclusive. One hypothesis should represent the position of the prosecution in the case, and the other the position of the defense, for example:

$\boldsymbol{H_1}$**:** the speaker on the questioned-speaker recording is the known speaker

versus

$\boldsymbol{H_2}$**:** the speaker on the questioned-speaker recording is not the known speaker but some other speaker selected at random from the relevant population

or

$\boldsymbol{H_1}$**:** the speakers on the questioned-speaker and the known-speaker recordings are the same speaker

versus

$\boldsymbol{H_2}$**:** the speakers on the questioned-speaker and the known-speaker recordings are not the same speaker but two different speakers each selected at random from the relevant population

The first example is of hypotheses for a specific-source likelihood ratio and the second example is of hypotheses for a common-source likelihood ratio (Ommen & Saunders, 2021). In both examples, the numerator of the likelihood ratio quantifies the similarity between the questioned-speaker and known-speaker recordings. In the specific-source example the denominator quantifies the typicality of the questioned-speaker recording with respect to the relevant population. In the common-source example, the denominator quantifies the typicality of the questioned-speaker and known-speaker recordings with respect to the relevant population. The relevant population is the population from which the questioned speaker could potentially have come if they were not the known speaker. The relevant population can usually be restricted to either male or female speakers who speak a particular language with a particular accent (Morrison, Enzinger, Zhang, 2016).

For continuously-valued data, such as acoustic measurements made on voice recordings, likelihood ratios are calculated as the ratio of two probability-density functions rather than the ratio of probabilities per se. For a specific-source likelihood ratio, the evidence, E, is the measurements made on the questioned-speaker recording, and the measurements made on one or more known-speaker recordings are used to build a specific-source probability-density model for the known speaker. For a common-source likelihood ratio, the evidence, E, is the measurements made on both the questioned-speaker recording and one or more known-speaker recordings. Human-supervised-automatic forensic-voice-comparison systems using state-of-the-art automatic-speaker-recognition technology calculate common-source likelihood ratios.

Many authors advocate using the logic of the likelihood-ratio framework even if likelihood-ratio values are not calculated using quantitative measurements and statistical models, but are instead assigned on the basis of the forensic practitioner's subjective judgment. In order to ensure that the practitioner transparently follows the logic of the likelihood-ratio framework, Willis et al. (2015) recommends that the practitioner separately state a subjectively assigned numeric value for the numerator of the likelihood ratio and a subjectively assigned numeric value for the denominator of the likelihood ratio. The practitioner can then potentially be asked to justify their assignment of each value.

Many authors advocate using verbal expressions of likelihood ratios either in addition to or in place of numeric likelihood-ratio values. Such verbal expressions are usually associated with ranges of numeric likelihood-ratio values, although the association is by fiat – there is no intrinsic relationship between the numeric ranges and the verbal expressions. An example of a verbal opinion scale, based on Willis et al. (2015), is provided in Table 1. There may be practitioners who do not actually follow the logic of the likelihood-ratio framework, but make a subjective posterior-probability judgment and then for reporting purposes pick a level from a verbal-likelihood-ratio scale.

Table 1. Examples of verbal expressions of likelihood ratios intended to correspond to ranges of numeric likelihood-ratio values. If the likelihood-ratio value is less than 1, the same expressions can be used with the ratio inverted and the order of $\boldsymbol{H_1}$ and $\boldsymbol{H_2}$ reversed.

<Table 1 near here>

Further descriptions of the likelihood-ratio framework in general can be found in [Authors (current volume)], Robertson et al. (2016), and Aitken et al. (2021). Introductions in the context of forensic voice comparison are provided in Morrison & Thompson (2017), Morrison, Enzinger, Zhang (2018), and Morrison & Enzinger

(2019a). For criticisms of the subjective assignment of likelihood ratios, see Risinger (2013), Martire et al. (2017), Morrison (2017), and Morrison, Ballantyne, Geoghegan (2018). For criticisms of the use of verbal-likelihood-ratio scales, see Marquis et al. (2016) and Morrison & Enzinger (2016).

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 for use in conjunction with the auditory-acoustic-phonetic approach (French & Harrison, 2007). This became known as the UK framework. 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 questioned-speaker and known-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". If the task is closed-set (the size of the relevant population is small and data are available from each member of the population), the practitioner can make a categorical statement of identification.

Unlike the numerator and denominator of a likelihood ratio, consistency and distinctiveness are not values 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).

The UK framework was criticized in Rose & Morrison (2009) and Morrison (2009, 2010, 2014) as not logically tenable, as suffering from cliff-edge effects, and for failing to consider empirical validation. As reported in French (2017), in 2015 the lead authors of the UK position statement abandoned their framework in favor of the Association of Forensic Science Providers' (AFSP, 2019) standard. The latter requires the use of the likelihood-ratio framework, but allows for subjective assignment of likelihood-ratio values. French (2017) indicated that they adopted the use of the verbal-expression scale from AFSP (2019), with the level on the scale assigned on the basis of subjective judgment. The AFSP (2019) verbal-expression scale (reproduced in Table 2) does not actually contain expressions of likelihood ratios – the expressions only mention one hypothesis and they state the level of support for that hypothesis rather than the probability of obtaining the evidence if the hypothesis were true. These expressions have been called support statements.

Table 2. Verbal expressions in AFSP (2019) intended to correspond to ranges of numeric likelihood ratio values, but that are not themselves expressions of likelihood ratios.

<Table 2 near here>

4 Popularity of analytical approaches and interpretive frameworks

Gold & French (2011) published the results of a survey of forensic-voice-comparison practitioners working in a mixture of private, university, and law-enforcement or other government laboratories. They reported results from 35 respondents. Morrison, Sahito, et al. (2016), published the results of a survey of forensic-voice-comparison capabilities of law-enforcement agencies in INTERPOL member countries. They reported results from 44 respondents who stated that their agency had forensic-voice-comparison capabilities. Gold & French (2019) published the results of a second survey that had 39 respondents, who had some overlap with the respondents in the earlier Gold & French survey. Summaries of the results are provided in Figure 4 and Figure 5. Respondents often reported using more than one approach or framework, hence the summary statistics in the figures add up to more than the 100% of the number of respondents. With respect to approaches, Gold & French (2019) only reported results for the human-supervised-automatic approach, hence in Figure 4 no Gold & French (2019) values are entered for the other approaches. With respect to frameworks, the Gold & French surveys did not provide breakdowns for verbal versus numeric expressions of posterior probabilities, but Gold & French (2011) suggested that all or most were verbal, hence in Figure 5 no Gold & French values are entered for numeric posterior probabilities.

<Figure 4 near here>

Figure 4. Popularity of different analytical approaches for forensic voice comparison according to published surveys.

<Figure 5 near here>

Figure 5. Popularity of different interpretive frameworks for forensic voice comparison according to published surveys.

Because of differences among the surveys in terms of their design and respondent populations, one should be cautious about drawing inferences about chronological trends. Over the time period covered, however, there appears to have been an increase in the popularity of the human-supervised-automatic approach and of the likelihood-ratio framework. The popularity of the UK framework also appears to have decreased, along with an increase in the popularity of support statements.

5 Legal admissibility

In common-law jurisdictions, courts had tended to focus on the admissibility of particular approaches to forensic voice comparison, although, from a scientific perspective, what should inform decisions as to whether the output of a forensic-voice-comparison system is good enough to be used in court is empirical validation of its performance under conditions reflecting those of the case under investigation. President’s Council of Advisors on Science and Technology (2016) p. 46:

> Without appropriate estimates of accuracy, an examiner’s statement that two samples are similar—or even indistinguishable—is scientifically meaningless: it has no probative value, and considerable potential for prejudicial impact. Nothing—not training, personal experience nor professional practices—can substitute for adequate empirical demonstration of accuracy.

Any approach could, in principle, be validated under conditions reflecting those of the case, and a decision could then be made as to whether the demonstrated degree of performance is sufficient. Courts in Australia, England & Wales, and Northern Ireland have admitted testimony based on auditory-spectrographic and auditory-acoustic-phonetic approaches without empirical validation of performance under conditions reflecting those of the case, or even under any conditions.

Few common-law jurisdictions have clear rules or precedents banning particular approaches to forensic voice comparison, so, pending challenges, all could potentially be admitted. Exceptions include that, as previously mentioned, the spectrographic and auditory-spectrographic approaches have been ruled inadmissible in US Federal Court, U.S. v. Angleton, 269 F.Supp 2nd 892 (S.D. Tex. 2003), and the auditory approach (but not the auditory-acoustic-phonetic approach) has been ruled inadmissible in Northern Ireland, R v O’Doherty [2002] NICA 20 / [2003] 1 Cr App R 5.

Morrison & Thompson (2017) presents a review of legal admissibility of different analytical approaches and interpretive frameworks in the United States. French (2017) and Morrison (2018a) present reviews covering England & Wales and Northern

Ireland. Briefer reviews covering Australia and Canada are included in Morrison & Enzinger (2019a).

6 Validation

A consensus on validation of forensic voice comparison, with 13 authors and 7 additional supporters, has been published as Morrison et al. (2021). The consensus provides guidance on how to validate forensic voice comparison systems. The scope of the consensus is systems that output numeric likelihood ratios for potential evidential use. In practice, these will tend to be human-supervised-automatic systems. Key points, listed in §2.12 of Morrison et al. (2021), are:

2.12.1 The forensic practitioner should communicate to the court what propositions the forensic practitioner has adopted for the case, including what they have adopted as the relevant population.

2.12.2 The forensic practitioner should communicate to the court what the forensic practitioner understands the conditions of the questioned-speaker and known-speaker recordings to be.

2.12.3 The forensic-voice-comparison system should be well calibrated.

2.12.4 Validation data should be representative of the relevant population for the case, and reflective of the conditions of the questioned-speaker and known-speaker recordings in the case.

2.12.5 The forensic practitioner's decision as to whether the validation data are sufficiently representative of the relevant population for the case, and sufficiently reflective of the conditions of the questioned-speaker and known-speaker recordings in the case, will be a subjective judgment.

2.12.6 Validation results should be presented as a Tippett plot and a C_{llr} value. These should be examined for signs of miscalibration.

2.12.7 The validation threshold (acceptance criterion) for C_{llr} should be 1. As long as C_{llr} is less than 1, the system is providing useful information.

2.12.8 To decide whether the likelihood-ratio value calculated for the comparison of the questioned-speaker and known-speaker recordings is supported by the validation results, it should be compared with the values shown in the Tippett plot.

An appendix in Morrison et al. (2021) provides details of how to calculate and interpret C_{llr} values, and how to draw and interpret Tippett plots. A tutorial on calibration is provided in Morrison (2013).

Published validation studies include those in a virtual special issue of the journal *Speech Communication*, in which multiple human-supervised-automatic systems were tested using the same data that reflected a set of forensically realistic conditions. A summary of the results is presented in Morrison & Enzinger (2019b).

7 Conclusion

In forensic voice comparison, popular analytical approaches have included auditory-acoustic-phonetic, auditory-spectrographic, acoustic-phonetic-statistical, and human-supervised-automatic. The human-supervised-automatic approach appears to be increasing in popularity. Compared to other approaches, the human-supervised-automatic approach is more objective and practically easier to validate. It outperforms the acoustic-phonetic-statistical approach and is less costly in human labor. Popular interpretive frameworks have included categorical-conclusion, posterior-probability, likelihood-ratio, and UK. The likelihood-ratio framework appears to be increasing in popularity. It is the logically correct framework for evaluation of evidence, and in combination with the human-supervised-automatic analytical approach outputs numeric likelihood-ratio values. Given its multiple advantages, an increase in the popularity of this combination is expected, alongside a gradual decline in the popularity of other approaches and frameworks. Morrison et al. (current volume) describes in more detail the human-supervised-automatic approach in combination with the likelihood-ratio framework.

8 References

[Authors (current volume) entry / entries on the likelihood-ratio framework]

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. (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*. London, UK: Royal Statistical Society. (https://rss.org.uk/news-publication/publications/law-guides/)

Aitken, C.G.G., Taroni, F. and Bozza, S. (2021). *Statistics and evaluation of evidence for forensic scientists* (3rd edn.). Chichester, UK: Wiley. (https://doi.org/10.1002/9781119245438)

American Board of Recorded Evidence (1999). *Voice comparison standards*. American Board of Recorded Evidence.

Archer, C. (2012, July 11). Voice recognition capabilities at the FBI – from the 1960s to the present. *Homeland Security News Wire*. (https://www.homelandsecuritynewswire.com/bull20120711-voice-recognition-capabilities-at-the-fbi-from-the-1960s-to-the-present)

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

Balko, R. (2017, August 31). The emperor of junk science forensics has died. *Washington Post*. (https://www.washingtonpost.com/news/the-watch/wp/2017/08/31/the-emperor-of-junk-science-forensics-has-died/)

Ballantyne, K., Bunford, J., Found, B., Neville, D., Taylor, D., Wevers, G. and Catoggio, D. (2017). *An introductory guide to evaluative reporting*. National Institute of Forensic Science of the Australia New Zealand Policing Advisory Agency. (http://www.anzpaa.org.au/forensic-science/our-work/projects/evaluative-reporting)

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*, pp.14–27. Amsterdam: John Benjamins.

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. (http://enfsi.eu/wp-content/uploads/2016/09/guidelines_fasr_and_fsasr_0.pdf)

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. (http://dx.doi.org/10.1016/j.forsciint.2017.05.007)

Forensic Science Regulator (2021). *Codes of practice and conduct: development of evaluative opinions* (FSR-C-118 Issue 1). (https://www.gov.uk/government/publications/development-of-evaluative-opinions)

Franco-Pedroso, J. and González-Rodríguez, J. (2016). Linguistically-constrained formant-based i-vectors for automatic speaker recognition. *Speech Communication* **76**, 61–81. (http://dx.doi.org/10.1016/j.specom.2015.11.002)

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. (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. (https://doi.org/10.1558/ijsll.v17i1.143)

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

Gold, E. and French, J.P. (2019). International practices in forensic speaker comparison: Second survey. *International Journal of Speech, Language and the Law* **26**, 1–20. (https://doi.org/10.1558/ijsll.38028)

Gruber, J.S. and Poza, F. (1995). Voicegram identification evidence. *American Jurisprudence Trials* **54**.

Gurlekian, J.A., Suligoy, S., Univaso, P., Torres, H., Masessa, E., Molina, N. (2022). Determining the likelihood ratio from perceptual attributes of voice. *Journal of Voice*. (https://doi.org/10.1016/j.jvoice.2022.01.022)

Hansen, J.H.L. and Bořil, H. (2018). On the issues of intra-speaker variability and realism in speech, speaker, and language recognition tasks. *Speech Communication* **101**, 94–108. (https://doi.org/10.1016/j.specom.2018.05.004)

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

Hudson, T., McDougall, K. and Hughes, V. (2021). Forensic phonetics. In Knight, R., Setter, J. (eds.), *The Cambridge handbook of phonetics*, pp. 631–656. Cambridge, UK: Cambridge University Press. (https://doi.org/10.1017/9781108644198.026)

Jessen, M. (2018). Forensic voice comparison. In Visconti, J. (ed.). *Handbook of communication in the legal sphere*, pp. 219–255. Berlin: De Gruyter. (https://doi.org/10.1515/9781614514664-012)

Jessen, M. (2021). Speaker profiling and forensic voice comparison: The auditory-acoustic approach. In Coulthard, M., May, A. and Sousa-Silva, R. (eds.) *The Routledge Handbook of Forensic Linguistics* (2nd edn.), pp. 382–399. London, UK: Routledge. (https://doi.org/10.4324/9780429030581)

Kafadar, K., Stern, H., Cuellar, M., Curran, J., Lancaster, M., Neumann, C., Saunders, C., Weir, B. and Zabell, S. (2019). *American Statistical Association position on statistical statements for forensic evidence*. American Statistical Association. (https://www.amstat.org/asa/files/pdfs/POL-ForensicScience.pdf)

Marquis, R., Biedermann, A., Cadola L., Champod, C., Gueissaz , L., Massonnet, G., Mazzella W.D., Taroni, F. and Hicks T. (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. (http://dx.doi.org/10.1016/j.scijus.2016.05.009)

Martire, K.A., Edmond, G., Navarro, D.J., Newell, B.R. (2017). On the likelihood of "encapsulating all uncertainty", *Science & Justice* **57**, 76–79. (http://dx.doi.org/10.1016/j.scijus.2016.10.004)

Morrison, G.S. (2009). Forensic voice comparison and the paradigm shift. *Science & Justice* **49**, 298–308. (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. (preprint: http://expert-evidence.forensic-voice-comparison.net/)

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. (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. (http://dx.doi.org/10.1016/j.scijus.2013.07.004)

Morrison, G.S. (2017). What should a forensic practitioner's likelihood ratio be? II. *Science & Justice* **57**, 472–476. (http://dx.doi.org/10.1016/j.scijus.2017.08.004)

Morrison, G.S. (2018a). Admissibility of forensic voice comparison testimony in England and Wales. *Criminal Law Review* **2018** issue 1, 20–33. (preprint: http://geoff-morrison.net/#Admissibility_EW_2018)

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. (http://dx.doi.org/10.1016/j.forsciint.2017.12.024)

Morrison, G.S., Ballantyne, K. and Geoghegan, P.H. (2018). A response to Marquis et al (2017). What is the error margin of your signature analysis? *Forensic Science International* **287**, e11–e12. (https://doi.org/10.1016/j.forsciint.2018.03.009)

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

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

Morrison, G.S. and Enzinger, E. (2019b). Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (forensic_eval_01) - Conclusion. *Speech Communication* **112**, 37–39. (https://doi.org/10.1016/j.specom.2019.06.007)

Morrison, G.S., Enzinger, E., Hughes, V., Jessen, M., Meuwly, D., Neumann, C., Planting, S., Thompson, W.C., van der Vloed, D., Ypma, R.J.F., Zhang, C., Anonymous, A. and Anonymous, B. (2021). Consensus on validation of forensic voice comparison. *Science & Justice* **61**, 229–309. (https://doi.org/10.1016/j.scijus.2021.02.002)

Morrison, G.S., Enzinger, E., Ramos, D., González-Rodríguez, J. and Lozano-Díez, A. (2020). Statistical models in forensic voice comparison. In Banks, D.L., Kafadar, K., Kaye, D.H. and Tackett, M. (eds.) *Handbook of Forensic Statistics*, pp. 451–497. Boca Raton, FL: CRC. (https://doi.org/10.1201/9780367527709)

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. (http://dx.doi.org/10.1016/j.scijus.2016.07.002)

Morrison, G.S., Enzinger, E. and Zhang, C. (2018). Forensic speech science. In Freckelton, I., Selby, H., (eds.) *Expert Evidence*. Sydney, Australia: Thomson Reuters, ch. 99. (preprint: http://expert-evidence.forensic-voice-comparison.net/)

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. (2017). A comment on the PCAST report: Skip the "match"/"non-match" stage. *Forensic Science International* **272**, e7–e9. (http://dx.doi.org/10.1016/j.forsciint.2016.10.018)

Morrison, G.S., Sahito F.H., Jardine G., Djokic D., Clavet S., Berghs S. and Goemans Dorny C. (2016). INTERPOL survey of the use of speaker identification by law enforcement agencies. *Forensic Science International* **263**, 92–100. (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. (https://doi.org/10.7916/stlr.v18i2)

Morrison, G.S., Weber, P., Enzinger, E., Labrador-Serrano, B., Lozano-Díez, A., Ramos, D. and González-Rodríguez, J. (current volume). Forensic voice comparison – Human-supervised-automatic approach. In Houck M., Wilson L., Lewis S., Eldridge H., Reedy P., Lothridge K. (Eds.), *Encyclopedia of Forensic Sciences* (3rd Ed.). Elsevier.

National Research Council (1979). *On the theory and practice of voice identification*.

Washington, DC: National Academies Press.

Ommen, D.M. and Saunders, C.P. (2021). A problem in forensic science highlighting the differences between the Bayes factor and likelihood ratio. *Statistical Science* **36**, 344–359. (https://doi.org/10.1214/20-STS805)

Poza, F. and Begault, D.R. (2005). Voice identification and elimination using aural-spectrographic protocols. *Proceedings of the Audio Engineering Society 26th International Conference: Audio Forensics in the Digital Age*, pp. 1–8.

President's Council of Advisors on Science and Technology (2016). *Forensic science in criminal courts: ensuring scientific validity of feature-comparison methods*. (https://obamawhitehouse.archives.gov/administration/eop/ostp/pcast/docsreports/)

Risinger, D.M. (2013). Reservations about likelihood ratios (and some other aspects of forensic 'Bayesianism'). *Law, Probability and Risk* **12**, 63–73, (http://dx.doi.org/10.1093/lpr/mgs011)

Robertson, B., Vignaux, G.A.and Berger, C.E.H. (2016). *Interpreting evidence: evaluating forensic science in the courtroom* (2nd edn.). Chichester, UK: Wiley. (http://dx.doi.org/10.1002/9781118492475)

Rose, P. (2017). Likelihood ratio-based forensic voice comparison with higher level features: research and reality. *Computer Speech & Language* **45**, 475–502 (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. (http://dx.doi.org/10.1558/ijsll.v16i1.139)

San Segundo, E., Foulkes, P., French, P., Harrison, P., Hughes, V. and Kavanagh, C. (2019). The use of the Vocal Profile Analysis for speaker characterization:

methodological proposals. *Journal of the International Phonetic Association* **49**, 353–380. (https://doi.org/10.1017/S0025100318000130)

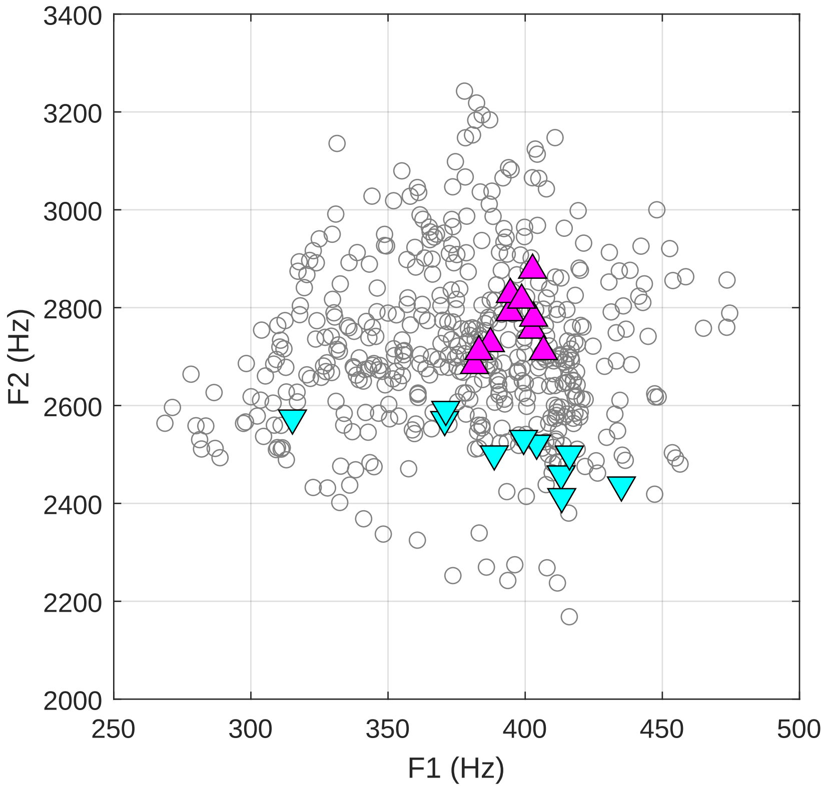
Speaker Recognition Subcommittee of the Organization of Scientific Area Committees for Forensic Science (2021). *Essential scientific literature for human-supervised automatic approaches to forensic speaker recognition*. Organization of Scientific Area Committees for Forensic Science. (https://www.nist.gov/document/essentialscientificliteratureforhuman)

Tosi, O. (1979). *Voice identification: Theory and legal applications*. Baltimore, MD: University Park Press.

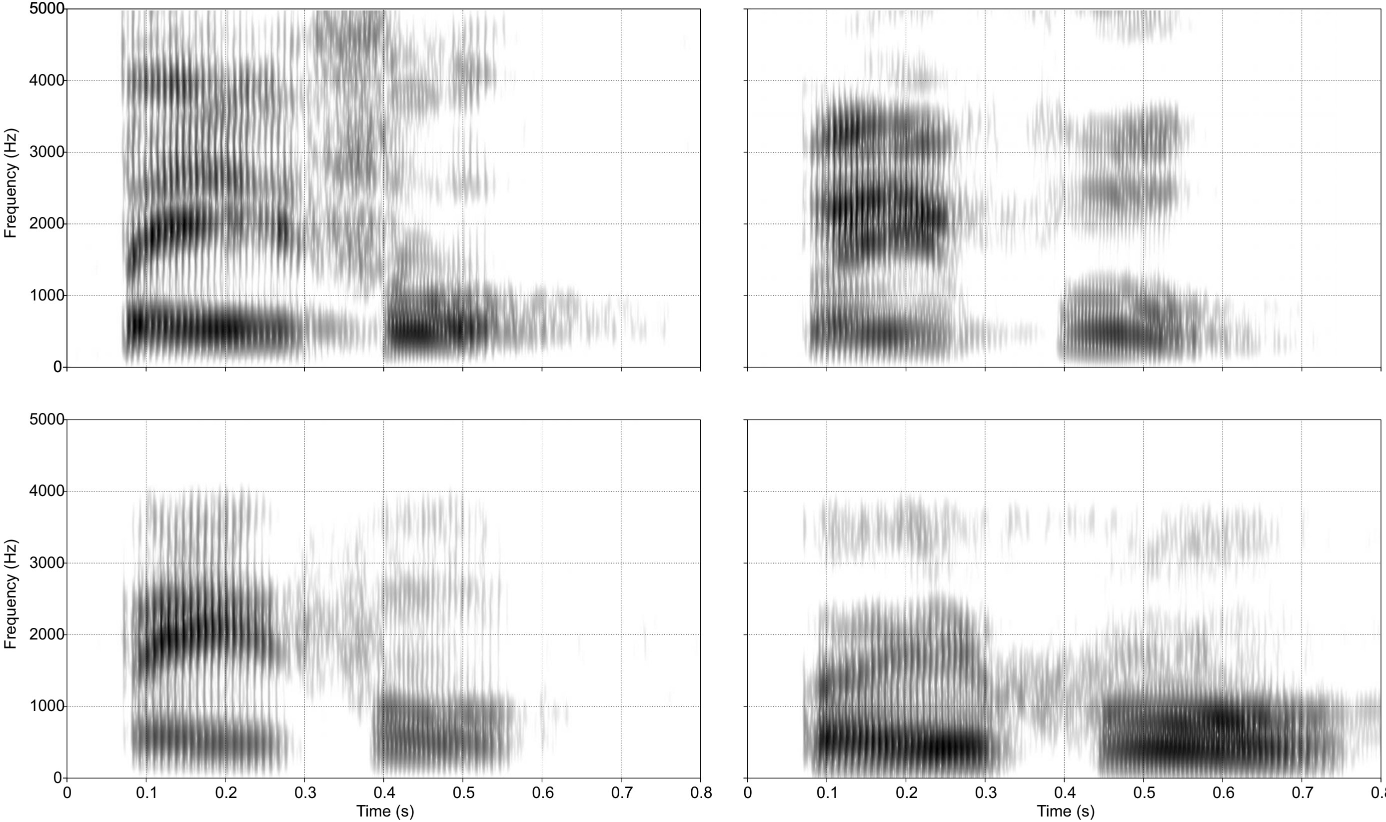
Wagner, I., Boss, D., Svirava, T., Siparov, I., Hughes, V. and Rolfes, M. (2021). *Best practice manual for the methodology of forensic speaker comparison*. European Network of Forensic Science Institutes. (https://enfsi.eu/wp-content/uploads/2021/07/2021-07-07-final-draft-BPM-SPEAKER-COMPARISON.pdf)

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. (http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf)

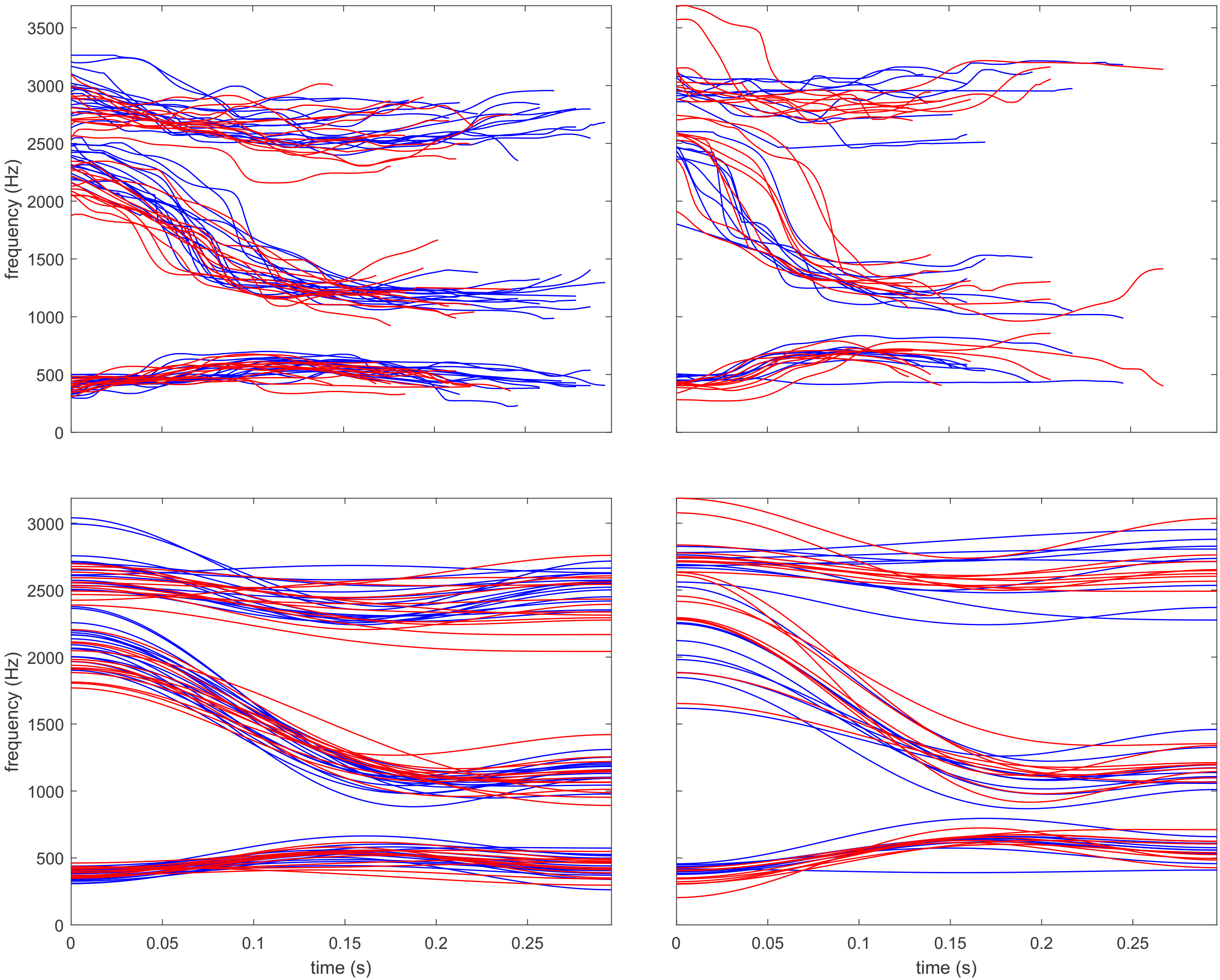
Zhang, C., Morrison, G.S., Enzinger, E. and Ochoa, F. (2013). Effects of telephone transmission on the performance of formant-trajectory-based forensic voice comparison – female voices. *Speech Communication* **55**, 796–813. (http://dx.doi.org/10.1016/j.specom.2013.01.011)

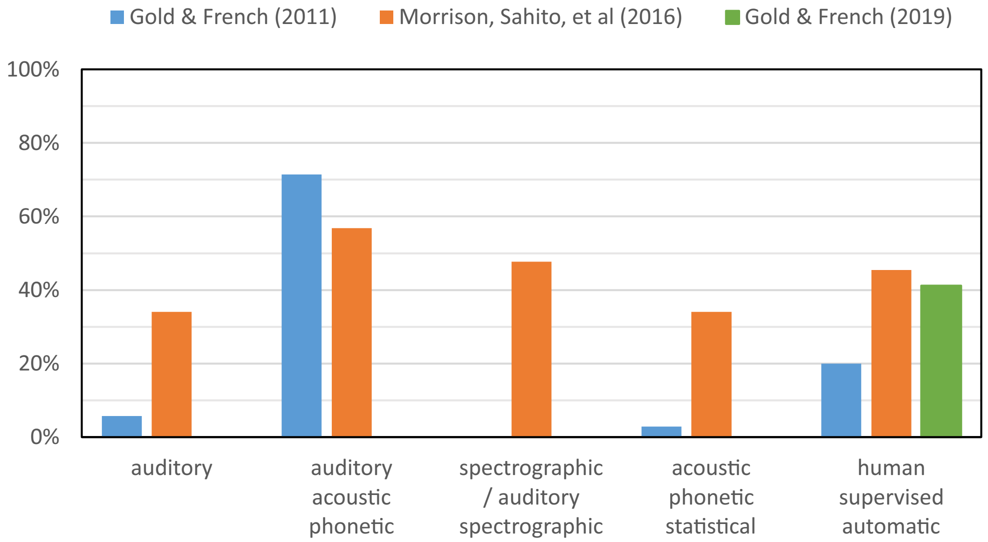


3400
3200
3000
2800
2600
2400
2200
2000
F2 (Hz)
250
300
350
400
450
500
F1 (Hz)



Frequency (Hz)
Frequency (Hz)
Time (s)
Time (s)


frequency (Hz)
frequency (Hz)
time (s)
time (s)


Gold & French (2011)
Morrison, Sahito, et al (2016)
Gold & French (2019)
100%
80%
60%
40%
20%
0%
auditory
auditory acoustic phonetic
spectrographic / auditory spectrographic
acoustic phonetic statistical
human supervised automatic

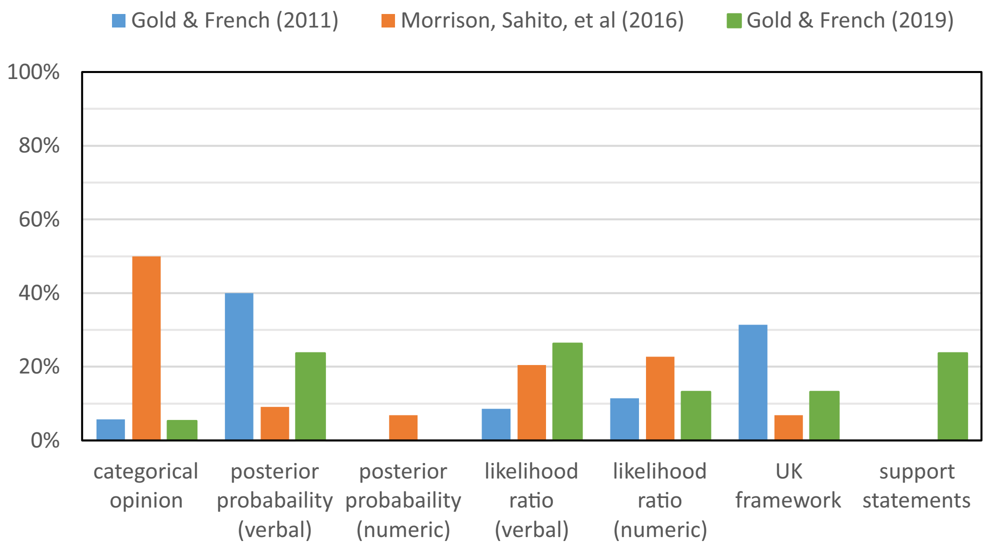

Gold & French (2011)
Morrison, Sahito, et al (2016)
Gold & French (2019)
100%
80%
60%
40%
20%
0%
categorical opinion
posterior probabaility (verbal)
posterior probabaility (numeric)
likelihood ratio (verbal)
likelihood ratio (numeric)
UK framework
support statements

Table 1. Examples of verbal expressions of likelihood ratios intended to correspond to ranges of numeric likelihood-ratio values. If the likelihood-ratio value is less than 1, the same expressions can be used with the ratio inverted and the order of H_1 and H_2 reversed.

Ranges of numeric likelihood ratios	**Verbal expressions of likelihood ratios**
1 – 2	The observations are *approximately equally probable* irrespective of whether H_1 were true or whether H_2 were true.
2 – 10	The observations are *slightly more probable* if H_1 were true than if H_2 were true.
10 – 100	The observations are *more probable* if H_1 were true than if H_2 were true.
100 – 1,000	The observations are *appreciably more probable* if H_1 were true than if H_2 were true.
1,000 – 10,000	The observations are *much more probable* if H_1 were true than if H_2 were true.
10,000 – 1,000,000	The observations are *far more probable* if H_1 were true than if H_2 were true.
1,000,000 or more	The observations are *exceedingly more probable* if H_1 were true than if H_2 were true.

Table 2. Verbal expressions in AFSP (2019) intended to correspond to ranges of numeric likelihood ratio values, but that are not themselves expressions of likelihood ratios.

Ranges of numeric likelihood ratios	**Verbal expressions (support statements)**
>1 – 10	Weak support for hypothesis
10 – 100	Moderate support
100 – 1,000	Moderately strong support
1,000 – 10,000	Strong support
10,000 – 1,000,000	Very strong
>1,000,000	Extremely strong