

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22

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

23 forensic speech science

24 forensic voice comparison

25 likelihood ratio

26 validation

27 voicegram

28 voiceprint

29

30 **Abstract**

31 [50–100 words]

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

41

42 **Key points**

43 [short bulleted list of key points]

44 • analytical approaches for extracting information:

- 45 ○ auditory-acoustic-phonetic
- 46 ○ auditory-spectrographic
- 47 ○ acoustic-phonetic-statistical
- 48 ○ human-supervised-automatic
- 49 • interpretive frameworks for drawing inferences from the analytical results:
 - 50 ○ categorical-opinion
 - 51 ○ posterior-probability
 - 52 ○ likelihood-ratio
 - 53 ○ UK
- 54 • legal admissibility
- 55 • validation

56

57 **Acknowledgements**

58 The writing of this entry was supported by Research England's Expanding Excellence
59 in England Fund as part of funding for the Aston Institute for Forensic Linguistics
60 2019–2023.

61 **1 Introduction**

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

67 Forensic voice comparison is challenging because, although there is variability in the
68 properties of voices between speakers, individual speakers also have natural within-
69 speaker variability, recording quality is often poor, there are often mismatches in
70 speaking style and recording conditions between the questioned-speaker recording and
71 the known-speaker recording, and the speaking styles and conditions vary from case to
72 case. Common speaking styles include normal vocal effort and raised vocal effort.
73 Common conditions include different types and volumes of background noise, and
74 processing of the recordings using different lossy codecs to reduce the amount of data
75 transmitted over communications systems or saved to storage media. Hansen & Bořil
76 (2018) provide a taxonomy of sources of speaker-intrinsic and speaker-extrinsic
77 variability that can potentially affect recordings used in forensic voice comparison.
78 Differences between the properties of the voices on different recordings could be due
79 to within-speaker variability or between-speaker variability. The forensic practitioner
80 has to assess the relative probabilities of obtaining the observed properties of the voices
81 on the questioned-speaker and known-speaker recordings if they were produced by the
82 same speaker versus if they were produced by different speakers.

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

89 Below, different analytical approaches and different interpretive frameworks that have
90 been used for forensic voice comparison are described. Also covered are the popularity
91 of different approaches and frameworks, legal admissibility, and validation of forensic
92 voice comparison. Some material has been adapted from Morrison & Thompson
93 (2017), Morrison, Enzinger, Zhang (2018), and Morrison & Enzinger (2019a). The
94 latter publications provide more detailed coverage of some topics and also cover other
95 topics related to forensic voice comparison and to forensic speech science more
96 broadly.

97 **2 Analytical approaches**

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

- 102 • auditory and auditory-acoustic-phonetic approaches
- 103 • spectrographic and auditory-spectrographic approaches
- 104 • acoustic-phonetic-statistical approach
- 105 • human-supervised-automatic approach

106 **2.1 Auditory and auditory-acoustic-phonetic approaches**

107 In the auditory approach, the practitioner listens to the questioned-speaker and known-
108 speaker recordings. They listen for similarities which they would expect to hear if the
109 two recordings consisted of speech from the same speaker, but which they would not
110 expect to hear if the recordings consisted of speech from different speakers. They also
111 listen for differences which they would expect to hear if the two recordings consisted
112 of speech from different speakers, but which they would not expect to hear if the
113 recordings consisted of speech from the same speaker. They may listen to any or all of

114 the pronunciation of particular vowel sounds and of particular consonant sounds, the
115 pronunciation of particular words or phrases, and other more global properties such as
116 intonation patterns.

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

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

127 Most practitioners of auditory approaches listen only to the questioned-speaker and
128 known-speaker recordings. Some practitioners, however, also listen to a set of foil
129 speakers, i.e., speakers who sound broadly similar to the known and questioned
130 speakers (including same sex, language spoken, and accent spoken), speaking in a
131 similar speaking style and under similar recording conditions. In an approach known
132 as blind grouping (Cambier-Langeveld et al., 2014), one practitioner prepares
133 recordings of the questioned speaker, the known speaker, and several foil speakers.
134 This may involve cutting each original recording into multiple shorter recordings. The
135 first practitioner must take care in selecting foil-speaker recordings so that the
136 questioned-speaker and known-speaker recordings do not stand out relative to the foil-
137 speaker recordings because of speaking style, linguistic content, or recording
138 conditions. The first practitioner presents the recordings to a second practitioner
139 without telling the second practitioner the origin of each recording or how many
140 speakers there are in total. The second practitioner then attempts to group the
141 recordings by speaker. Whether the second practitioner groups the questioned-speaker

142 and known-speaker recordings together constitutes the result of the analysis. The
143 correctness of the grouping of foil speakers serves as a test of performance.

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

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

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

177 <Figure 1 near here>

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

183

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

186 1. Vocal setting and voice quality ... with up to 38 individual elements to be
187 considered.

188 2. Intonation ...

189 3. Pitch, measured as average and variation in fundamental frequency.

190 4. Articulation rate.

191 5. Rhythmical features.

192 6. Connected speech processes such as patterns of assimilation and elision.

193 7. A large set of consonantal features, including energy loci of fricatives and

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

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

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

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

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

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

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

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

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

223 **2.2 Spectrographic and auditory-spectrographic approaches**

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

236 <Figure 2 near here>

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

243

244 In contrast to other approaches, in which practitioners usually use existing recordings
245 of the known speaker (e.g., recordings of police interviews), practitioners of the

246 spectrographic approach usually make new recordings of the known speaker in which
247 the known speaker is required to say the same words as appear on the questioned-
248 speaker recording and in the same manner as they were said on the questioned-speaker
249 recording. This is required by published protocols.

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

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

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

267 And to auditorily compare:

- 268 a. Pitch ...
- 269 b. Intonation ...

- 270 c. Stress/Emphasis ...
- 271 d. Rate ...
- 272 e. Disguise ...
- 273 f. Mode ...
- 274 g. Psychological state ...
- 275 h. Speech defects. ...
- 276 i. Vocal quality ...
- 277 j. Other ... long-term fluctuations of pitch (vibrato), vocal fry (extremely low
278 pitching), pitch breaks, and stuttering.

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

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

289 In the early 1970s, there was a debate about whether a visual only or a visual plus
290 auditory approach was better, which ended with a preference for the latter, i.e., the
291 auditory-spectrographic approach. From the late 1960s, the spectrographic and
292 auditory-spectrographic approaches were highly controversial. By the end of the 1970s,
293 their popularity in the United States began to decline, but they continued to be used

294 into the early 2000s. Following a Daubert admissibility hearing, they were declared
295 inadmissible in *U.S. v. Angleton*, 269 F.Supp 2nd 892 (S.D. Tex. 2003). Gruber &
296 Poza (1995) §6 summarized the objections to the auditory-spectrographic approach as
297 follows:

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

307 Further descriptions of spectrographic and auditory-spectrographic approaches can be
308 found in Tosi (1979) and in National Research Council (1979). Criticisms of
309 spectrographic and auditory-spectrographic approaches can be found in Gruber & Poza
310 (1995) and in Morrison (2014).

311 **2.3 Acoustic-phonetic-statistical approach**

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

318 The acoustic-phonetic approach has been combined with statistical models that
319 calculate likelihood ratios. These require extracting acoustic-phonetic measurements
320 not only from the questioned-speaker and known-speaker recordings, but also from

321 recordings of speakers sampled from the relevant population (see the discussion of the
322 likelihood-ratio framework below).

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

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

341

342 <Figure 3 near here>

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

348 coefficients. Different colors indicate instances of “iao” spoken during different
349 recording sessions that were approximately two weeks apart.

350

351 In direct comparisons under forensically-realistic conditions (or conditions
352 approaching forensically-realistic conditions), human-supervised-automatic systems
353 have been found to perform much better than acoustic-phonetic-statistical systems, and
354 when the two are combined the improvement in performance over a human-supervised-
355 automatic system alone is at-best negligible (Zhang et al. 2013; Enzinger & Morrison,
356 2017). An exception is presented in Franco-Pedroso & González-Rodríguez (2016),
357 which achieved an increase in performance when combining an acoustic-phonetic
358 system with an automatic system. Franco-Pedroso & González-Rodríguez (2016)
359 extracted formant trajectories automatically, and, compared to the other studies,
360 extracted them from more speech sounds in a much larger database. They also used a
361 later generation of automatic-speaker-recognition technology.

362 Further descriptions of the acoustic-phonetic-statistical approach can be found in Rose
363 (2017) and Jessen (2018). The European Network of Forensic Science Institutes
364 (ENFSI) has published a guideline on the use of acoustic-phonetic-statistical and
365 human-supervised-automatic approaches (Drygajlo et al., 2015).

366 **2.4 Human-supervised-automatic approach**

367 The human-supervised-automatic approach makes use of automatic-speaker-
368 recognition technology, which also has non-forensic applications. Software tools are
369 used to make measurements of the acoustic properties of the questioned-speaker and
370 known-speaker recordings, and of recordings of other speakers sampled from the
371 relevant population for the case. Acoustic measurements are usually made over the
372 whole of the speech of the speaker of interest in a recording, and there is usually no
373 focus on individual speech sounds, words, or phrases. The types of measurements made
374 are the same as those used in speech processing in general, which includes applications

375 such as automatic speech recognition. An example of a common type of measurement
376 is a vector of mel-frequency cepstral coefficients (MFCCs). MFCC vectors consist of
377 a set of numbers, e.g., 14 numbers, which describe the frequency components (the
378 spectrum) of the speech during a short interval of time, e.g., 20 ms. MFCC vectors are
379 usually extracted every few milliseconds, e.g., every 10 ms with a 50% overlap
380 between adjacent 20 ms long intervals. An MFCC vector provides more detailed
381 measurements of the acoustic spectrum of a speech signal than do acoustic-phonetic
382 measurements such as fundamental frequency plus two or three formants.

383 The measurements made in the human-supervised-automatic approach are
384 automatically entered into statistical models. The combination of the human-
385 supervised-automatic approach and the likelihood-ratio framework is common, so
386 these models usually calculate likelihood ratios. Human-supervised-automatic systems
387 based on state-of-the-art automatic-speaker-recognition technology use deep neural
388 networks (DNNs). DNNs are trained using data from at least tens of thousands of
389 recordings of at least thousands of speakers. The speakers are diverse and the
390 recordings of each speaker are diverse in speaking styles and recordings conditions.
391 The DNN therefore learns about both within-speaker and between-speaker variability.
392 The MFCC vectors (or other measurements) from a recording are presented to the
393 DNN, and vectors of numbers known as x-vectors (or more generically, DNN
394 embeddings), are extracted from the DNN – they are the activation levels of the nodes
395 in a pre-final layer of the DNN. One x-vector is extracted for each recording. x-vectors
396 are the same length irrespective of the length of the recording. Different recordings of
397 the same speaker tend to cluster together in the x-vector space, whereas recordings of
398 different speakers tend to be separated from each other in the x-vector space. x-vectors
399 are extracted from the questioned-speaker and known-speaker recordings, and these
400 are input to backend models that calculate likelihood ratios. The x-vectors used to train
401 or adapt the backend models must come from recordings of speakers who are
402 representative of the relevant population for the case, and those recordings must reflect
403 the conditions of the questioned-speaker and known-speaker recordings in the case,

404 including any mismatch between the conditions of the questioned-speaker and known-
405 speaker recordings. Compared to earlier generations of automatic-speaker-recognition
406 technology, x-vector systems perform substantially better under forensically realistic
407 conditions (Morrison & Enzinger, 2019b).

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

415 **3 Interpretive frameworks**

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

- 417 • categorical-conclusion framework
- 418 • posterior-probability framework
- 419 • likelihood-ratio framework
- 420 • UK framework

421 Each is discussed in its own subsection below.

422 **3.1 Categorical-conclusion framework**

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

428 constitutes a sufficient degree of certainty, it is a subjective judgment on the part of the
429 practitioner.

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

443 **3.2 Posterior-probability framework**

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

454 As a matter of logic, posterior probabilities cannot be derived solely via comparison of

455 the properties of the questioned-speaker and known-speaker recordings. Logically,
456 according to Bayes' Theorem, a posterior probability must be the result of combining
457 a likelihood ratio with a prior probability. Even if they are not aware of it, a forensic
458 practitioner who presents a posterior probability must have at least implicitly used a
459 prior probability. The prior probability will be either arbitrary or based on other
460 information. The choice of prior probability will affect the value of the posterior
461 probability, but this will be hidden from the legal-decision maker if only the posterior
462 probability is presented. A practitioner who arbitrarily chooses a low prior probability
463 will present a lower posterior probability, and a practitioner who arbitrarily chooses a
464 high prior probability will present a higher posterior probability, but the legal-decision
465 maker will be misled into thinking that the difference is related to a difference in the
466 evidence, the properties of the questioned-speaker and known-speaker recordings. The
467 task of the forensic practitioner is to evaluate and express an opinion associated with
468 the one piece of evidence that they were asked to evaluate. They should not consider
469 other information. Considering all the information presented during a trial is the task
470 of the legal-decision maker. If a forensic practitioner's conclusion took account of other
471 information that the legal-decision maker had already taken account of, but the legal-
472 decision maker thought that the forensic practitioner was presenting new independent
473 information, then the legal-decision maker would double count that information.

474 **3.3 Likelihood-ratio framework**

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

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

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

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

491 (1)

492
$$\Lambda = \frac{p(E|H_1)}{p(E|H_2)}$$

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

496 **H_1** : the speaker on the questioned-speaker recording is the known speaker

497 versus

498 **H_2** : the speaker on the questioned-speaker recording is not the known speaker but
499 some other speaker selected at random from the relevant population

500 or

501 **H_1** : the speakers on the questioned-speaker and the known-speaker recordings are
502 the same speaker

503 versus

504 **H_2** : the speakers on the questioned-speaker and the known-speaker recordings are
505 not the same speaker but two different speakers each selected at random from the
506 relevant population

507 The first example is of hypotheses for a specific-source likelihood ratio and the second
508 example is of hypotheses for a common-source likelihood ratio (Ommen & Saunders,
509 2021). In both examples, the numerator of the likelihood ratio quantifies the similarity
510 between the questioned-speaker and known-speaker recordings. In the specific-source
511 example the denominator quantifies the typicality of the questioned-speaker recording
512 with respect to the relevant population. In the common-source example, the
513 denominator quantifies the typicality of the questioned-speaker and known-speaker
514 recordings with respect to the relevant population. The relevant population is the
515 population from which the questioned speaker could potentially have come if they were
516 not the known speaker. The relevant population can usually be restricted to either male
517 or female speakers who speak a particular language with a particular accent (Morrison,
518 Enzinger, Zhang, 2016).

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

529 Many authors advocate using the logic of the likelihood-ratio framework even if
530 likelihood-ratio values are not calculated using quantitative measurements and
531 statistical models, but are instead assigned on the basis of the forensic practitioner's

532 subjective judgment. In order to ensure that the practitioner transparently follows the
533 logic of the likelihood-ratio framework, Willis et al. (2015) recommends that the
534 practitioner separately state a subjectively assigned numeric value for the numerator of
535 the likelihood ratio and a subjectively assigned numeric value for the denominator of
536 the likelihood ratio. The practitioner can then potentially be asked to justify their
537 assignment of each value.

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

547

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

552 <Table 1 near here>

553

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

558 (2019a). For criticisms of the subjective assignment of likelihood ratios, see Risinger
559 (2013), Martire et al. (2017), Morrison (2017), and Morrison, Ballantyne, Geoghegan
560 (2018). For criticisms of the use of verbal-likelihood-ratio scales, see Marquis et al.
561 (2016) and Morrison & Enzinger (2016).

562 **3.4 UK framework**

563 In 2007, a group of forensic-voice-comparison practitioners and researchers in the
564 United Kingdom published a position statement that included a framework for
565 evaluation of evidence for use in conjunction with the auditory-acoustic-phonetic
566 approach (French & Harrison, 2007). This became known as the UK framework. The
567 framework has two stages: “consistency” and “distinctiveness”. In the first stage, the
568 practitioner makes a subjective judgment as to “whether the known and questioned
569 samples are compatible, or consistent, with having been produced by the same speaker”
570 (French & Harrison, 2007, p. 141). The choices are “consistent”, “not consistent”, or
571 “no-decision”. If the practitioner decides that the samples are “not consistent”, the
572 practitioner may state that they were spoken by different speakers and express their
573 degree of confidence that this is so (this is a posterior probability). If the practitioner
574 decides that the samples are “consistent”, the practitioner then makes a subjective
575 judgment as to whether the questioned-speaker and known-speaker recordings fall into
576 one of five levels of distinctiveness with respect to the relevant population:
577 “exceptionally-distinctive”, “highly-distinctive”, “distinctive”, “moderately-
578 distinctive”, or “not-distinctive”. If the task is closed-set (the size of the relevant
579 population is small and data are available from each member of the population), the
580 practitioner can make a categorical statement of identification.

581 Unlike the numerator and denominator of a likelihood ratio, consistency and
582 distinctiveness are not values on the same scale, and there are no explicit algorithms
583 for assigning values to consistency or distinctiveness. The latter are assigned
584 “informally via the analyst’s experience and general linguistic knowledge rather than
585 formally and quantitatively” (French et al., 2010, p. 144).

586 The UK framework was criticized in Rose & Morrison (2009) and Morrison (2009,
587 2010, 2014) as not logically tenable, as suffering from cliff-edge effects, and for failing
588 to consider empirical validation. As reported in French (2017), in 2015 the lead authors
589 of the UK position statement abandoned their framework in favor of the Association
590 of Forensic Science Providers' (AFSP, 2019) standard. The latter requires the use of
591 the likelihood-ratio framework, but allows for subjective assignment of likelihood-ratio
592 values. French (2017) indicated that they adopted the use of the verbal-expression scale
593 from AFSP (2019), with the level on the scale assigned on the basis of subjective
594 judgment. The AFSP (2019) verbal-expression scale (reproduced in Table 2) does not
595 actually contain expressions of likelihood ratios – the expressions only mention one
596 hypothesis and they state the level of support for that hypothesis rather than the
597 probability of obtaining the evidence if the hypothesis were true. These expressions
598 have been called support statements.

599

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

603 <Table 2 near here>

604

605 **4 Popularity of analytical approaches and interpretive frameworks**

606 Gold & French (2011) published the results of a survey of forensic-voice-comparison
607 practitioners working in a mixture of private, university, and law-enforcement or other
608 government laboratories. They reported results from 35 respondents. Morrison, Sahito,
609 et al. (2016), published the results of a survey of forensic-voice-comparison
610 capabilities of law-enforcement agencies in INTERPOL member countries. They
611 reported results from 44 respondents who stated that their agency had forensic-voice-

612 comparison capabilities. Gold & French (2019) published the results of a second survey
613 that had 39 respondents, who had some overlap with the respondents in the earlier Gold
614 & French survey. Summaries of the results are provided in Figure 4 and Figure 5.
615 Respondents often reported using more than one approach or framework, hence the
616 summary statistics in the figures add up to more than the 100% of the number of
617 respondents. With respect to approaches, Gold & French (2019) only reported results
618 for the human-supervised-automatic approach, hence in Figure 4 no Gold & French
619 (2019) values are entered for the other approaches. With respect to frameworks, the
620 Gold & French surveys did not provide breakdowns for verbal versus numeric
621 expressions of posterior probabilities, but Gold & French (2011) suggested that all or
622 most were verbal, hence in Figure 5 no Gold & French values are entered for numeric
623 posterior probabilities.

624 <Figure 4 near here>

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

627 <Figure 5 near here>

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

630

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

637 **5 Legal admissibility**

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

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

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

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

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

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

667 **6 Validation**

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

674 2.12.1 The forensic practitioner should communicate to the court what propositions
675 the forensic practitioner has adopted for the case, including what they have
676 adopted as the relevant population.

677 2.12.2 The forensic practitioner should communicate to the court what the forensic
678 practitioner understands the conditions of the questioned-speaker and
679 known-speaker recordings to be.

680 2.12.3 The forensic-voice-comparison system should be well calibrated.

681 2.12.4 Validation data should be representative of the relevant population for the
682 case, and reflective of the conditions of the questioned-speaker and known-
683 speaker recordings in the case.

684 2.12.5 The forensic practitioner's decision as to whether the validation data are
685 sufficiently representative of the relevant population for the case, and
686 sufficiently reflective of the conditions of the questioned-speaker and
687 known-speaker recordings in the case, will be a subjective judgment.

688 2.12.6 Validation results should be presented as a Tippett plot and a C_{llr} value. These
689 should be examined for signs of miscalibration.

690 2.12.7 The validation threshold (acceptance criterion) for C_{lr} should be 1. As long
691 as C_{lr} is less than 1, the system is providing useful information.

692 2.12.8 To decide whether the likelihood-ratio value calculated for the comparison
693 of the questioned-speaker and known-speaker recordings is supported by the
694 validation results, it should be compared with the values shown in the Tippett
695 plot.

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

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

703 7 Conclusion

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

717 more detail the human-supervised-automatic approach in combination with the
718 likelihood-ratio framework.

719

720 **8 References**

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

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

730 Aitken, C.G.G., Roberts, P. and Jackson, G. (2010). *Fundamentals of probability and*
731 *statistical evidence in criminal proceedings: Guidance for judges, lawyers,*
732 *forensic scientists and expert witnesses*. London, UK: Royal Statistical Society.
733 (<https://rss.org.uk/news-publication/publications/law-guides/>)

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

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

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

- 742 capabilities-at-the-fbi-from-the-1960s-to-the-present)
- 743 Association of Forensic Science Providers (2009). Standards for the formulation of
744 evaluative forensic science expert opinion. *Science & Justice* **49**, 161–164.
745 (<http://dx.doi.org/10.1016/j.scijus.2009.07.004>)
- 746 Balko, R. (2017, August 31). The emperor of junk science forensics has died.
747 *Washington Post*. ([https://www.washingtonpost.com/news/the-
748 watch/wp/2017/08/31/the-emperor-of-junk-science-forensics-has-died/](https://www.washingtonpost.com/news/the-watch/wp/2017/08/31/the-emperor-of-junk-science-forensics-has-died/))
- 749 Ballantyne, K., Bunford, J., Found, B., Neville, D., Taylor, D., Wevers, G. and
750 Catoggio, D. (2017). *An introductory guide to evaluative reporting*. National
751 Institute of Forensic Science of the Australia New Zealand Policing Advisory
752 Agency. ([http://www.anzpaa.org.au/forensic-science/our-
753 work/projects/evaluative-reporting](http://www.anzpaa.org.au/forensic-science/our-work/projects/evaluative-reporting))
- 754 Cambier-Langeveld T., van Rossum M. and Vermeulen J. (2014). Whose voice is
755 that? Challenges in forensic phonetics. In van Heuven V. and Caspers J. (eds.)
756 *Above and Beyond the Segments: Experimental Linguistics and Phonetics*,
757 pp.14–27. Amsterdam: John Benjamins.
- 758 Drygajlo, A., Jessen, M., Gfroerer, S., Wagner, I., Vermeulen, J. and Niemi T.
759 (2015). *Methodological guidelines for best practice in forensic semiautomatic
760 and automatic speaker recognition, including guidance on the conduct of
761 proficiency testing and collaborative exercises*. European Network of Forensic
762 Science Institutes. ([http://enfsi.eu/wp-
763 content/uploads/2016/09/guidelines_fasr_and_fsasr_0.pdf](http://enfsi.eu/wp-content/uploads/2016/09/guidelines_fasr_and_fsasr_0.pdf))
- 764 Enzinger, E. and Morrison, G.S. (2017). Empirical test of the performance of an
765 acoustic-phonetic approach to forensic voice comparison under conditions
766 similar to those of a real case. *Forensic Science International* **277**, 30–40.
767 (<http://dx.doi.org/10.1016/j.forsciint.2017.05.007>)

- 768 Forensic Science Regulator (2021). *Codes of practice and conduct: development of*
769 *evaluative opinions* (FSR-C-118 Issue 1).
770 (<https://www.gov.uk/government/publications/development-of-evaluative->
771 [opinions](https://www.gov.uk/government/publications/development-of-evaluative-))
- 772 Franco-Pedroso, J. and González-Rodríguez, J. (2016). Linguistically-constrained
773 formant-based i-vectors for automatic speaker recognition. *Speech*
774 *Communication* **76**, 61–81. (<http://dx.doi.org/10.1016/j.specom.2015.11.002>)
- 775 French, J.P. (2017). A developmental history of forensic speaker comparison in the
776 UK. *English Phonetics* **21**, 271–286.
- 777 French, J.P. and Harrison, P. (2007). Position statement concerning use of
778 impressionistic likelihood terms in forensic speaker comparison cases.
779 *International Journal of Speech, Language and the Law* **14**, 137–144.
780 (<https://doi.org/10.1558/ijssl.v14i1.137>)
- 781 French, J.P., Nolan, F., Foulkes, P., Harrison P. and McDougall, K. (2010). The UK
782 position statement on forensic speaker comparison: A rejoinder to Rose and
783 Morrison. *International Journal of Speech, Language and the Law* **17**, 143–152.
784 (<https://doi.org/10.1558/ijssl.v17i1.143>)
- 785 Gold, E. and French, J.P. (2011). International practices in forensic speaker
786 comparison. *International Journal of Speech, Language and the Law* **18**, 143–
787 152. (<http://dx.doi.org/10.1558/ijssl.v18i2.293>)
- 788 Gold, E. and French, J.P. (2019). International practices in forensic speaker
789 comparison: Second survey. *International Journal of Speech, Language and the*
790 *Law* **26**, 1–20. (<https://doi.org/10.1558/ijssl.38028>)
- 791 Gruber, J.S. and Poza, F. (1995). Voicegram identification evidence. *American*
792 *Jurisprudence Trials* **54**.

- 793 Gurlekian, J.A., Suligoy, S., Univaso, P., Torres, H., Masessa, E., Molina, N. (2022).
794 Determining the likelihood ratio from perceptual attributes of voice. *Journal of*
795 *Voice*. (<https://doi.org/10.1016/j.jvoice.2022.01.022>)
- 796 Hansen, J.H.L. and Bořil, H. (2018). On the issues of intra-speaker variability and
797 realism in speech, speaker, and language recognition tasks. *Speech*
798 *Communication* **101**, 94–108. (<https://doi.org/10.1016/j.specom.2018.05.004>)
- 799 Hollien, H. (2016). An approach to speaker identification. *Journal of Forensic*
800 *Sciences* **61**, 334–344. (<http://dx.doi.org/10.1111/1556-4029.13034>)
- 801 Hudson, T., McDougall, K. and Hughes, V. (2021). Forensic phonetics. In Knight,
802 R., Setter, J. (eds.), *The Cambridge handbook of phonetics*, pp. 631–656.
803 Cambridge, UK: Cambridge University Press.
804 (<https://doi.org/10.1017/9781108644198.026>)
- 805 Jessen, M. (2018). Forensic voice comparison. In Visconti, J. (ed.). *Handbook of*
806 *communication in the legal sphere*, pp. 219–255. Berlin: De Gruyter.
807 (<https://doi.org/10.1515/9781614514664-012>)
- 808 Jessen, M. (2021). Speaker profiling and forensic voice comparison: The auditory-
809 acoustic approach. In Coulthard, M., May, A. and Sousa-Silva, R. (eds.) *The*
810 *Routledge Handbook of Forensic Linguistics* (2nd edn.), pp. 382–399. London,
811 UK: Routledge. (<https://doi.org/10.4324/9780429030581>)
- 812 Kafadar, K., Stern, H., Cuellar, M., Curran, J., Lancaster, M., Neumann, C.,
813 Saunders, C., Weir, B. and Zabell, S. (2019). *American Statistical Association*
814 *position on statistical statements for forensic evidence*. American Statistical
815 Association. (<https://www.amstat.org/asa/files/pdfs/POL-ForensicScience.pdf>)
- 816 Marquis, R., Biedermann, A., Cadola L., Champod, C., Gueissaz, L., Massonnet, G.,
817 Mazzella W.D., Taroni, F. and Hicks T. (2016). Discussion on how to
818 implement a verbal scale in a forensic laboratory: Benefits, pitfalls and

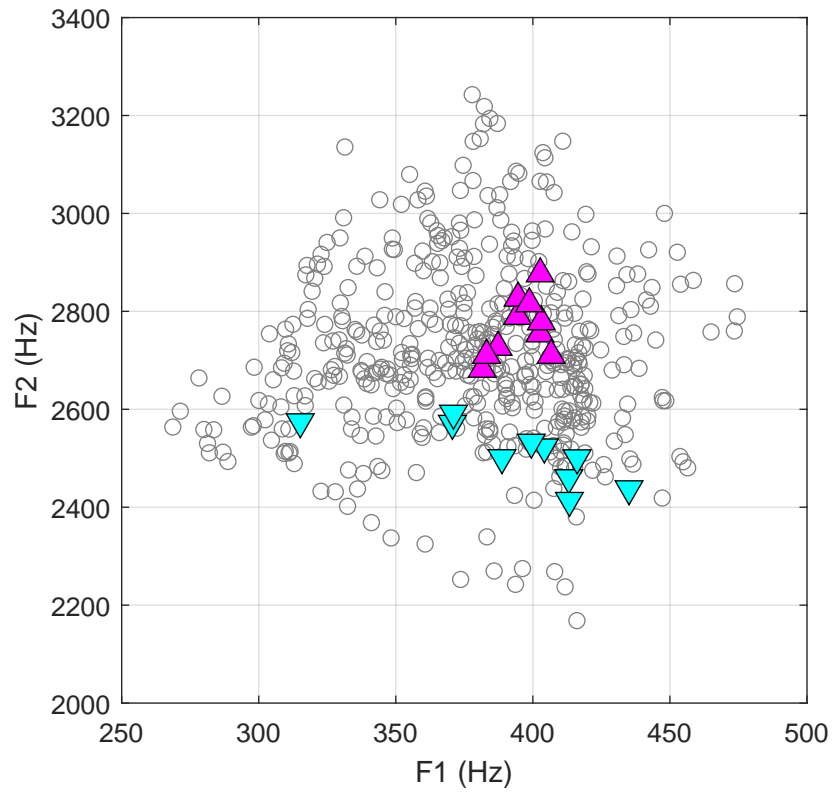
- 819 suggestions to avoid misunderstandings. *Science & Justice* **56**, 364–370.
820 (<http://dx.doi.org/10.1016/j.scijus.2016.05.009>)
- 821 Martire, K.A., Edmond, G., Navarro, D.J., Newell, B.R. (2017). On the likelihood of
822 “encapsulating all uncertainty”, *Science & Justice* **57**, 76–79.
823 (<http://dx.doi.org/10.1016/j.scijus.2016.10.004>)
- 824 Morrison, G.S. (2009). Forensic voice comparison and the paradigm shift. *Science &*
825 *Justice* **49**, 298–308. (<https://doi.org/10.1016/j.scijus.2009.09.002>)
- 826 Morrison, G.S. (2010). Forensic voice comparison. In Freckelton I. and Selby, H.,
827 (eds.) *Expert Evidence*. Sydney, Australia: Thomson Reuters, ch. 99. (preprint:
828 <http://expert-evidence.forensic-voice-comparison.net/>)
- 829 Morrison, G.S. (2013). Tutorial on logistic-regression calibration and fusion:
830 converting a score to a likelihood ratio. *Australian Journal of Forensic Sciences*
831 **45**, 173–197. (<http://dx.doi.org/10.1080/00450618.2012.733025>)
- 832 Morrison, G.S. (2014). Distinguishing between forensic science and forensic
833 pseudoscience: Testing of validity and reliability, and approaches to forensic
834 voice comparison. *Science & Justice* **54**, 245–256.
835 (<http://dx.doi.org/10.1016/j.scijus.2013.07.004>)
- 836 Morrison, G.S. (2017). What should a forensic practitioner’s likelihood ratio be? II.
837 *Science & Justice* **57**, 472–476. (<http://dx.doi.org/10.1016/j.scijus.2017.08.004>)
- 838 Morrison, G.S. (2018a). Admissibility of forensic voice comparison testimony in
839 England and Wales. *Criminal Law Review* **2018** issue 1, 20–33. (preprint:
840 http://geoff-morrison.net/#Admissibility_EW_2018)
- 841 Morrison, G.S. (2018b). The impact in forensic voice comparison of lack of
842 calibration and of mismatched conditions between the known-speaker recording
843 and the relevant-population sample recordings. *Forensic Science International*

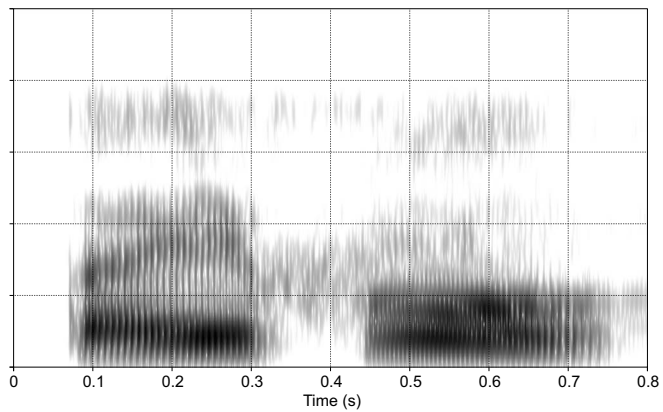
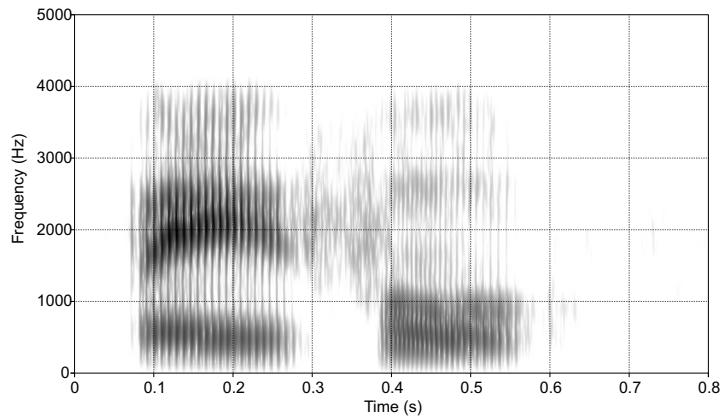
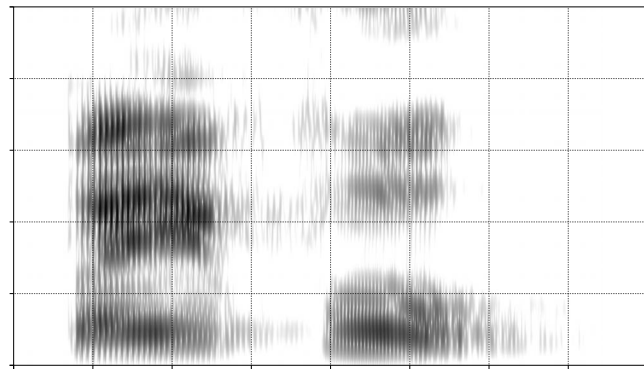
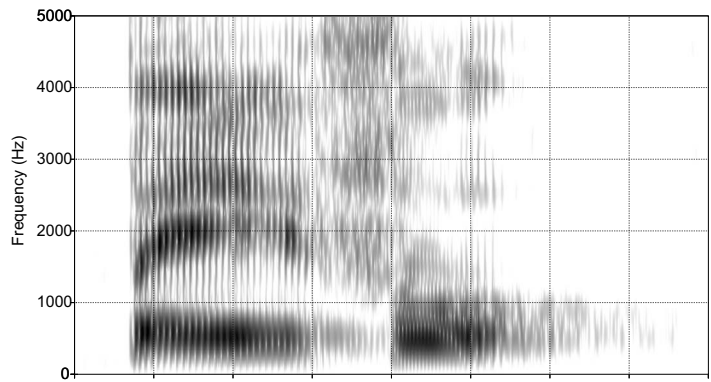
- 844 **283**, e1–e7. (<http://dx.doi.org/10.1016/j.forsciint.2017.12.024>)
- 845 Morrison, G.S., Ballantyne, K. and Geoghegan, P.H. (2018). A response to Marquis
846 et al (2017). What is the error margin of your signature analysis? *Forensic*
847 *Science International* **287**, e11–e12.
848 (<https://doi.org/10.1016/j.forsciint.2018.03.009>)
- 849 Morrison, G.S. and Enzinger, E. (2016). What should a forensic practitioner's
850 likelihood ratio be? *Science & Justice* **56**, 374–379.
851 (<http://dx.doi.org/10.1016/j.scijus.2016.05.007>)
- 852 Morrison, G.S. and Enzinger, E. (2019a). Introduction to forensic voice comparison.
853 In Katz W.F. and Assmann P.F. (eds.) *The Routledge Handbook of Phonetics*,
854 pp. 599–634. Abingdon, UK: Taylor & Francis.
855 (<https://doi.org/10.4324/9780429056253-22>)
- 856 Morrison, G.S. and Enzinger, E. (2019b). Multi-laboratory evaluation of forensic
857 voice comparison systems under conditions reflecting those of a real forensic
858 case (forensic_eval_01) - Conclusion. *Speech Communication* **112**, 37–39.
859 (<https://doi.org/10.1016/j.specom.2019.06.007>)
- 860 Morrison, G.S., Enzinger, E., Hughes, V., Jessen, M., Meuwly, D., Neumann, C.,
861 Planting, S., Thompson, W.C., van der Vloed, D., Ypma, R.J.F., Zhang, C.,
862 Anonymous, A. and Anonymous, B. (2021). Consensus on validation of forensic
863 voice comparison. *Science & Justice* **61**, 229–309.
864 (<https://doi.org/10.1016/j.scijus.2021.02.002>)
- 865 Morrison, G.S., Enzinger, E., Ramos, D., González-Rodríguez, J. and Lozano-Díez,
866 A. (2020). Statistical models in forensic voice comparison. In Banks, D.L.,
867 Kafadar, K., Kaye, D.H. and Tackett, M. (eds.) *Handbook of Forensic Statistics*,
868 pp. 451–497. Boca Raton, FL: CRC. (<https://doi.org/10.1201/9780367527709>)
- 869 Morrison, G.S., Enzinger, E. and Zhang, C. (2016). Refining the relevant population

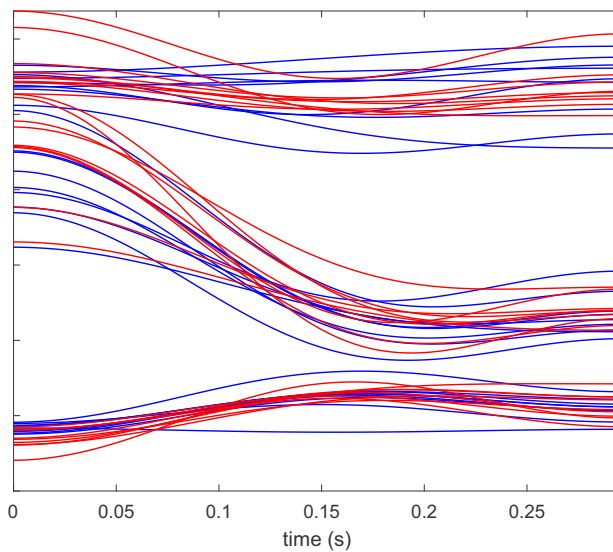
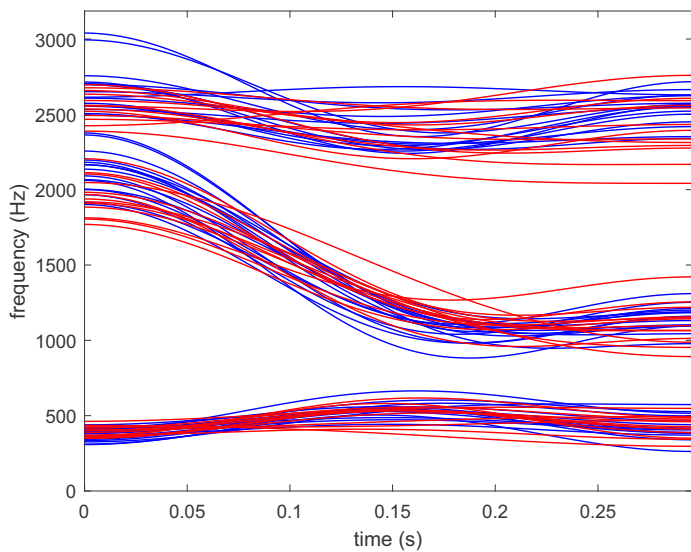
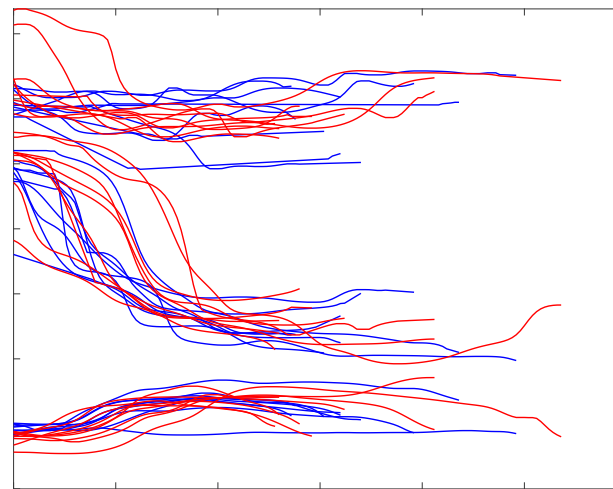
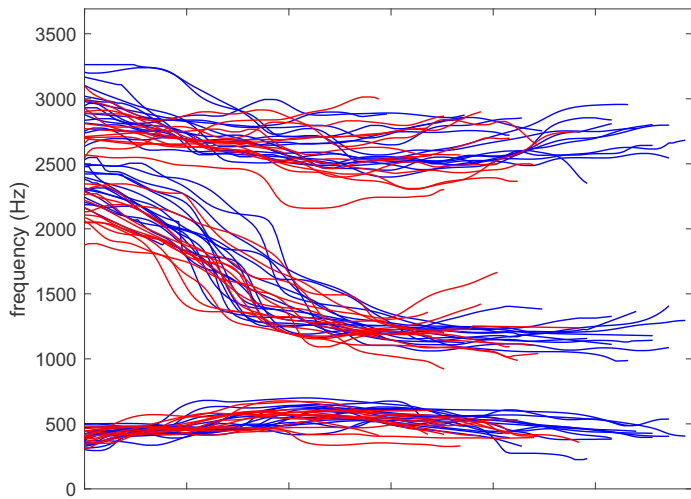
- 870 in forensic voice comparison – A response to Hicks et alii (2015) The
871 importance of distinguishing information from evidence/observations when
872 formulating propositions. *Science & Justice* **56**, 492–497.
873 (<http://dx.doi.org/10.1016/j.scijus.2016.07.002>)
- 874 Morrison, G.S., Enzinger, E. and Zhang, C. (2018). Forensic speech science. In
875 Freckelton, I., Selby, H., (eds.) *Expert Evidence*. Sydney, Australia: Thomson
876 Reuters, ch. 99. (preprint: <http://expert-evidence.forensic-voice-comparison.net/>)
- 877 Morrison, G.S., Kaye, D.H., Balding, D.J., Taylor, D., Dawid, P., Aitken, C.G.G.,
878 Gittelsohn, S., Zadora, G., Robertson, B., Willis, S.M., Pope, S., Neil, M.,
879 Martire, K.A., Hepler, A., Gill, R.D., Jamieson, A., de Zoete, J., Ostrum, R.B.
880 and Caliebe, A. (2017). A comment on the PCAST report: Skip the
881 “match”/“non-match” stage. *Forensic Science International* **272**, e7–e9.
882 (<http://dx.doi.org/10.1016/j.forsciint.2016.10.018>)
- 883 Morrison, G.S., Sahito F.H., Jardine G., Djokic D., Clavet S., Berghs S. and Goemans
884 Dorny C. (2016). INTERPOL survey of the use of speaker identification by law
885 enforcement agencies. *Forensic Science International* **263**, 92–100.
886 (<http://dx.doi.org/10.1016/j.forsciint.2016.03.044>)
- 887 Morrison, G.S. and Thompson, W.C. (2017). Assessing the admissibility of a new
888 generation of forensic voice comparison testimony. *Columbia Science and
889 Technology Law Review* **18**, 326–434. (<https://doi.org/10.7916/stlr.v18i2>)
- 890 Morrison, G.S., Weber, P., Enzinger, E., Labrador-Serrano, B., Lozano-Díez, A.,
891 Ramos, D. and González-Rodríguez, J. (current volume). Forensic voice
892 comparison – Human-supervised-automatic approach. In Houck M., Wilson L.,
893 Lewis S., Eldridge H., Reedy P., Lothridge K. (Eds.), *Encyclopedia of Forensic
894 Sciences* (3rd Ed.). Elsevier.
- 895 National Research Council (1979). *On the theory and practice of voice identification*.

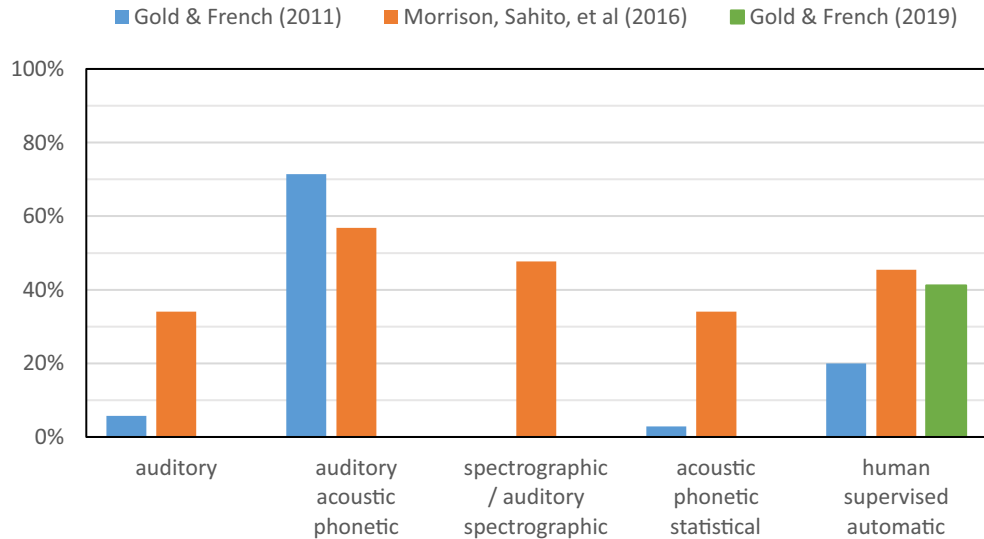
- 896 Washington, DC: National Academies Press.
- 897 Ommen, D.M. and Saunders, C.P. (2021). A problem in forensic science highlighting
898 the differences between the Bayes factor and likelihood ratio. *Statistical Science*
899 **36**, 344–359. (<https://doi.org/10.1214/20-STS805>)
- 900 Poza, F. and Begault, D.R. (2005). Voice identification and elimination using aural-
901 spectrographic protocols. *Proceedings of the Audio Engineering Society 26th*
902 *International Conference: Audio Forensics in the Digital Age*, pp. 1–8.
- 903 President’s Council of Advisors on Science and Technology (2016). *Forensic science*
904 *in criminal courts: ensuring scientific validity of feature-comparison methods*.
905 ([https://obamawhitehouse.archives.gov/administration/eop/ostp/pcast/docsreport](https://obamawhitehouse.archives.gov/administration/eop/ostp/pcast/docsreports/)
906 [s/](https://obamawhitehouse.archives.gov/administration/eop/ostp/pcast/docsreports/))
- 907 Risinger, D.M. (2013). Reservations about likelihood ratios (and some other aspects
908 of forensic ‘Bayesianism’). *Law, Probability and Risk* **12**, 63–73,
909 (<http://dx.doi.org/10.1093/lpr/mgs011>)
- 910 Robertson, B., Vignaux, G.A. and Berger, C.E.H. (2016). *Interpreting evidence:*
911 *evaluating forensic science in the courtroom* (2nd edn.). Chichester, UK: Wiley.
912 (<http://dx.doi.org/10.1002/9781118492475>)
- 913 Rose, P. (2017). Likelihood ratio-based forensic voice comparison with higher level
914 features: research and reality. *Computer Speech & Language* **45**, 475–502
915 (<http://dx.doi.org/10.1016/j.csl.2017.03.003>)
- 916 Rose, P. and Morrison, G.S. (2009). A response to the UK position statement on
917 forensic speaker comparison. *International Journal of Speech, Language and the*
918 *Law* **16**, 139–163. (<http://dx.doi.org/10.1558/ijssl.v16i1.139>)
- 919 San Segundo, E., Foulkes, P., French, P., Harrison, P., Hughes, V. and Kavanagh, C.
920 (2019). The use of the Vocal Profile Analysis for speaker characterization:

- 921 methodological proposals. *Journal of the International Phonetic Association* **49**,
922 353–380. (<https://doi.org/10.1017/S0025100318000130>)
- 923 Speaker Recognition Subcommittee of the Organization of Scientific Area
924 Committees for Forensic Science (2021). *Essential scientific literature for*
925 *human-supervised automatic approaches to forensic speaker recognition*.
926 Organization of Scientific Area Committees for Forensic Science.
927 (<https://www.nist.gov/document/essentialscientificliteratureforhuman>)
- 928 Tosi, O. (1979). *Voice identification: Theory and legal applications*. Baltimore, MD:
929 University Park Press.
- 930 Wagner, I., Boss, D., Svirava, T., Siparov, I., Hughes, V. and Rolfes, M. (2021). *Best*
931 *practice manual for the methodology of forensic speaker comparison*. European
932 Network of Forensic Science Institutes. ([https://enfsi.eu/wp-](https://enfsi.eu/wp-content/uploads/2021/07/2021-07-07-final-draft-BPM-SPEAKER-COMPARISON.pdf)
933 [content/uploads/2021/07/2021-07-07-final-draft-BPM-SPEAKER-](https://enfsi.eu/wp-content/uploads/2021/07/2021-07-07-final-draft-BPM-SPEAKER-COMPARISON.pdf)
934 [COMPARISON.pdf](https://enfsi.eu/wp-content/uploads/2021/07/2021-07-07-final-draft-BPM-SPEAKER-COMPARISON.pdf))
- 935 Willis, S.M., McKenna, L., McDermott, S., O'Donnell, G., Barrett, A., Rasmusson,
936 A., Nordgaard, A., Berger, C.E.H., Sjerps, M.J., Lucena-Molina, J.J., Zadora,
937 G., Aitken, C.G.G., Lunt, L., Champod, C., Biedermann, A., Hicks, T.N. and
938 Taroni, F. (2015). *ENFSI guideline for evaluative reporting in forensic science*.
939 European Network of Forensic Science Institutes. ([http://enfsi.eu/wp-](http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf)
940 [content/uploads/2016/09/m1_guideline.pdf](http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf))
- 941 Zhang, C., Morrison, G.S., Enzinger, E. and Ochoa, F. (2013). Effects of telephone
942 transmission on the performance of formant-trajectory-based forensic voice
943 comparison – female voices. *Speech Communication* **55**, 796–813.
944 (<http://dx.doi.org/10.1016/j.specom.2013.01.011>)
- 945









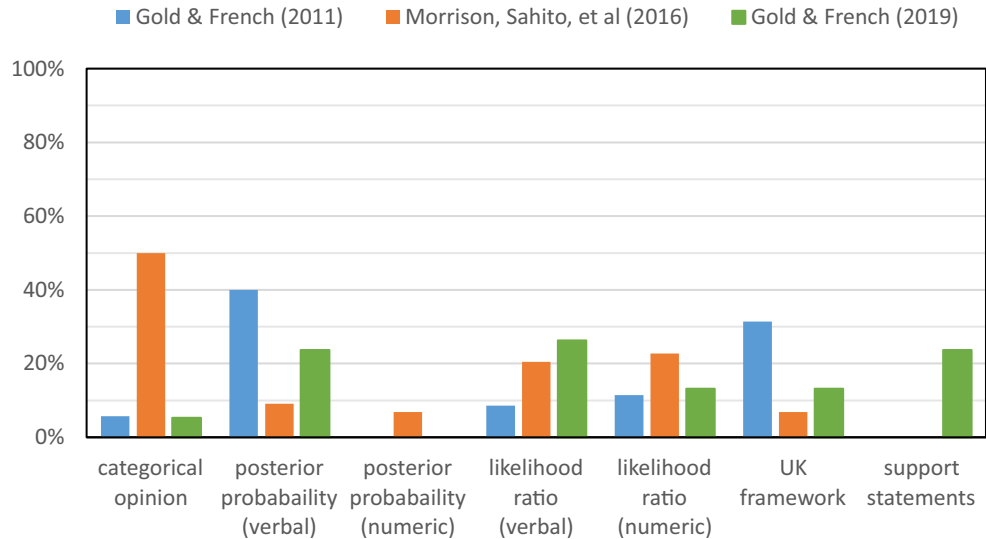


Table 1. Examples of verbal expressions of likelihood ratios intended to correspond to ranges of numeric likelihood-ratio values. If the likelihood-ratio value is less than 1, the same expressions can be used with the ratio inverted and the order of H_1 and H_2 reversed.

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