99
Forensic voice comparison

by

Geoffrey Stewart Morrison BSc, MTS, MA, PhD

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

Author information

Dr. Geoffrey Stewart Morrison was awarded his PhD by the Department of Linguistics, University of Alberta in 2006. His doctoral and postdoctoral work (supported by fellowships from the Social Sciences and Humanities Research Council of Canada) focussed on statistical modelling of second-language speech perception.

He began work on forensic voice comparison in 2007 when he was appointed Research Associate on an Australian Research Council Discovery Project directed by Dr. Philip Rose at the School of Language Studies, Australian National University.

As of September 2010, he is Director of the Forensic Voice Comparison Laboratory at the School of Electrical Engineering & Telecommunications, University of New South Wales. He is lead investigator on an Australian Research Council Linkage Project aimed at improving the validity and reliability of forensic voice comparison via the combination of acoustic-phonetic and automatic approaches (partner organisations include the Australian Federal Police, the National Institute of Forensic Science, and the Australasian Speech Science and Technology Association).

He is also an Invited Lecturer in the Judicial Phonetics Specialisation of the Master in Phonetics and Phonology Programme, Consejo Superior de Investigaciones Científicas [Spanish National Research Council] / Universidad Internacional Menéndez Pelayo.

Dr. Morrison has published papers on forensic voice comparison in the *Australian Journal of Forensic Sciences*, the *International Journal of Speech, Language and the Law*, the *Journal of the Acoustical Society of America*, and *Science & Justice*. He was an invited speaker at the International Conference on Evidence Law and Forensic Science, Beijing, 2009, and has presented tutorials on forensic voice comparison at the International Speech Communication Association's Interspeech conference, Brisbane, 2008, the Audio Engineering Society Conference on Audio Forensics, Hillerød, Denmark, 2010, and the Pan-American/Iberian Meeting on Acoustics, Cancún, Mexico, 2010.

More information about Dr. Morrison's research can be found at http://geoff-morrison.net, http://forensic-voice-comparison.net, and http://forensic.unsw.edu.au.

[A webpage related to this chapter can be found at http://expert-evidence.forensic-voice-comparison.net/. This chapter replaces Rose (2003) *The technical comparison of forensic voice samples* which can now be accessed via the webpage given above. Parts of this chapter are based on parts of Morrison (2009b). Financial support for the writing of this chapter came from Australian Research Council Discovery Project Grant No. DP0774115. The author wishes to express his appreciation to Philip Rose, John Buckleton, Julian Epps, and editor Hugh Selby for helpful comments on earlier drafts of the chapter. To paraphrase Bernard de Chartres, if I have seen far it is because I have been standing on the shoulders of giants. In the class of giants I would like to include my mentors Terrance M Nearey and Philp J Rose.]

CONTENTS

INTRODUCTION

What is forensic voice comparison?

[99.10] Forensic voice comparison is the comparison of one or more audio recordings of the voice of a known speaker with an audio recording of the voice of a speaker of questioned identity for the purpose of presenting expert testimony in court or during pre-trial investigation. Typically the known speaker is a suspect/defendant and the questioned speaker an offender. Here are two representative (fictional) scenarios:

• In a major fraud case involving hundreds of millions of dollars an audio recording of a telephone call made by the offender to the bank is available. An audio recording of a telephone call made by a suspect, a former bank employee, is also available (the defence does not contest the identity of the speaker on this recording). A forensic scientist conducts a forensic comparison of the two voice recordings. In court, the forensic scientist testifies that one would be 2000 times more likely to observe the acoustic differences between the voice recordings under the prosecution's assertion that the recordings are of the same speaker than under the defence's assertion that they are of different speakers. This, along with other evidence, leads to a conviction.

• The police have a telephone-intercept warrant and record a suspected terrorist plotting with a previously unknown associate whom they designate Mr. X. They eventually arrest the suspected terrorist and question a number of his associates, making audio recordings of the interviews. They think that one of the associates, Mr. Y, is Mr. X because to them the voices on the two recordings sound the same. They recommend that Mr. Y be prosecuted, but the prosecutor is of the opinion that the other evidence against him being involved is weak and will not likely lead to a conviction. They provide the audio recordings to a forensic scientist for analysis. The forensic scientist conducts a forensic voice comparison and reports that one would be 1000 times more likely to observe the acoustic differences between the voices in the two recordings under the assumption that they were produced by different speakers than under the assumption that they were produced by the same speaker. The police and prosecutor decide to focus their resources on another suspect who eventually confesses to being Mr. X.

Audience

[99.20] As part of the *Expert Evidence* series this chapter is aimed first at **lawyers**, **judges**, **police officers**, and **potential jury members**; however, it is hoped that this chapter will also be of interest to **forensic scientists**, **phoneticians / speech scientists**, **speech-processing engineers**, and **students of all these disciplines**. It introduces forensic voice comparison in a relatively non-technical way, assuming a reader who has no prior knowledge of the subject For sake of correctness occasional more-technical asides will be necessary, but this chapter will not go into sufficient detail to allow the reader to begin performing forensic voice comparison themselves. The focus will be on the understanding of concepts and the provision of basic knowledge.

Structure

[99.30] The chapter is structured in a logical order suitable for reading from beginning to end, but it may also be possible to read some sections relatively independently of the others. Cross-references point both backwards and forwards in the chapter.

The first four major sections after the introduction describe the new paradigm for forensic-comparison science, including the evaluation of forensic-comparison evidence and the testing of the validity and reliability of forensic-comparison systems. The first three of these major sections (**A Paradigm Shift in Forensic-Comparison Science [99.70] ff**, **The Likelihood-Ratio Framework for the Evaluation of Forensic-Comparison Evidence [99.140] ff**, and **Assessing the Validity and Reliability (Accuracy and Precision) of Forensic-comparison Systems [99.290] ff**) are essential reading for anyone not already familiar with the topics they cover. The fourth (**Problems and Opposition [99.370] ff**) may be skipped.

The next two major sections (**Human Voices [99.440] ff** and **Voice Recording and Voice Transmission [99.600] ff**) may be skipped by readers already familiar with these topics but are highly recommended for other readers.

The next major section (**Approaches to Forensic Voice Comparison [99.650] ff**) describes and evaluates different approaches to forensic voice comparison. This is essential reading. The evaluation sections rely on an understanding of most of the earlier sections in the chapter.

The next major section (**Examples [99.770] ff**) presents two examples of forensic voice comparison, each using a different approach, but each conducted within the new paradigm. This section relies on an understanding of, and draws together, much of the material presented earlier in the chapter.

The final major section (**Comparison of Technical Forensic Voice Comparison and Non-Technical Speaker Identification [99.910] ff**) compares forensic voice comparison performed by forensic scientists working in the new paradigm with speaker identification performed by police officers with no training in forensic voice comparison. It relies in part on an understanding of forensic voice comparison as described in earlier sections. This section may be skipped if the reader's interest is only in forensic voice comparison *per se*.

Appendices provide a copy of the International Phonetic Alphabet **[99.1150]**, and some notes on training and qualifications **[99.1190]** and professional associations **[99.1200]**.

Questions

[99.40] There are a number of questions which investigators, counsel, opposing counsel, judges, and jury members should ask about forensic voice comparison. The exact form of the questions and which questions are most important will differ depending on the questioner's *rôle* in the justice system, but they are fundamentally the same questions. The questions are posed below in a generic form. An investigator, prosecutor, or defence attorney considering soliciting the services of an expert in forensic voice comparison, should be asking whether these things will be done. A judge deciding whether evidence based on forensic voice comparison should be admitted should be asking whether these things are done by the forensic scientist in general and whether they have been done in this case. Counsel should be asking whether these things have been done, expecting to elicit positive

answers so as to instill confidence in the trier of fact with respect to the evidence presented. Opposing counsel should be asking whether these things have been done, expecting to elicit negative answers so as to instill a lack of confidence in the trier of fact with respect to the evidence presented. The trier of fact should be asking whether these things have been done so as to help them to evaluate the evidence presented.

1. Is the voice evidence evaluated using the logically correct framework for the evaluation of forensic-comparison evidence?

2. Is the forensic voice comparison based on objective measurements of the voice recordings? That is, towards the objective end of the subjectivity–objectivity continuum.

3. Is an adequate database of voice recordings of speakers from the relevant population used to assess the typicality of the known- and questioned-voice samples?

4. Have the validity and reliability (accuracy and precision) of the forensic-voice-comparison system been empirically evaluated under conditions similar to those in the present case, and have they been found to be acceptable?

5. What is the strength of evidence for the comparison of the known- and questioned-voice samples in the present case, and what is its estimated precision and error rate?

Much of the remainder of this chapter attempts to provide the reader with an understanding of what these questions mean, why they must be asked, and how to evaluate the answers.

A PARADIGM SHIFT IN FORENSIC-COMPARISON SCIENCE

A paradigm shift

[99.70] Today we are in the midst of what Saks & Koehler (2005) have called a *paradigm shift* in the evaluation and presentation of evidence in the forensic sciences which deal with the comparison of the quantifiable properties of objects of known and questioned origin, e.g., deoxyribonucleic acid (DNA), finger marks, hairs, fibres, glass fragments, tool marks, handwriting, and voice recordings. Saks & Koehler point out that they "use the notion of paradigm shift not as a literal application of Thomas Kuhn's concept (Kuhn, 1962), but as a metaphor highlighting the transformation involved in moving from a pre-science to an empirically grounded science" (p. 892). In Kuhnian terms, Saks & Koehler's paradigm shift might be better described as a shift from a pre-paradigm period towards a period where there is for the first time a single unifying paradigm for conducting normal science, i.e., a shift from a period during which a number of different schools pursue solutions to different sets of problems (with only partial overlap between sets) using different incompatible frameworks, towards a period during which there is agreement throughout the scientific community as to which problems are important (often a superset of the problems addressed by two or more of the pre-paradigm schools), and agreement as to the general procedures for solving these problems and the nature of suitable solutions.

Saks & Koehler (2005) propose that a paradigm shift has already occurred in DNA profile comparison, and that other forensic-comparison sciences are now shifting towards the new paradigm. Forensic voice comparison is one branch of forensic science in which this shift is now well underway but in which it is still far from reaching universal acceptance among researchers and practitioners.

The new paradigm

[99.80] Saks & Koehler (2005) describe the new paradigm as "empirically grounded science" (p. 892) as exemplified by "data-based, probabilistic assessment" (p. 893) as is current practice in forensic DNA-profile comparison. They recommend that other forensic comparison sciences emulate DNA-profile comparison, including that they "construct databases of sample characteristics and use these databases to support a probabilistic approach" (p. 893). They also make it clear that another important aspect of the new paradigm is the quantification and reporting of the limitations of forensic comparison via the measurement of error rates. The new paradigm therefore echos the requirements for admissibility of scientific evidence set out in the US Supreme Court ruling in *Daubert v Merrell Dow Pharmaceuticals* (92-102) 509 US 579 [1993], which Saks & Koehler identify as a driving force for the paradigm shift. The Court ruled that, when considering the admissibility of scientific evidence, the judge must consider the methodology's scientific validity and evidentiary reliability, including whether it has been empirically tested and found to have an acceptable error rate.

The call for other branches of forensic science to be more "scientific", emulate DNA-profile comparison, and conform to the *Daubert* requirements was reiterated in the National Research

Council (NRC) report to Congress on *Strengthening Forensic Science in the United States* (NRC, 2009). Important aspects of a scientific approach identified in the report include "the careful and precise characterization of the scientific procedure, so that others can replicate and validate it; . . . the quantification of measurements . . .; the reporting of a measurement with an interval that has a high probability of containing the true value; . . . [and] the conducting of validation studies of the performance of a forensic procedure" (p. 121); the latter requiring the use of "quantifiable measures of the reliability and accuracy of forensic analyses" (p. 23). The NRC report clearly recommends the use of more objective analytic methodologies over more subjective experience-based methodologies.

Although there does not appear to be any indication that either set of authors were consciously aware of this, there is one other component of the new paradigm which I believe is implicit in Saks & Koehler's (2005) and the NRC report's (2009) recommendation that other forensic comparison sciences emulate forensic DNA-profile comparison: the adoption of the *likelihood-ratio framework* for the evaluation of evidence.

Further reading

[99.90] For a history of the adoption of the new paradigm in forensic-voice-comparison research and practice up to 2009, see Morrison (2009b). Articles emphasising the need for a more "scientific" approach to forensic science include Cole (2006), Edmond *et al*. (2010), Saks & Koehler (2005), and Saks & Faigman (2008).

THE LIKELIHOOD-RATIO FRAMEWORK FOR THE EVALUATION OF FORENSIC-COMPARISON EVIDENCE

Introduction

[99.140] The likelihood-ratio framework has already been described in **Interpreting Scientific Evidence [28]** (Robertson & Vignaux, 2000), and its application to DNA-profile comparison was described in **Statistical Evaluation in Forensic DNA Typing [80A]** (Roberts, 2004). Other descriptions are listed in the **Further reading** section **[99.250]** below. The present section describes the likelihood-ratio framework taking the perspective that the data to be compared come from voice recordings.

The use of the likelihood-ratio framework is required by the *Association of Forensic Science Providers*' (AFSP's) Standards for the Formulation of Evaluative Forensic Science Expert Opinion (AFSP, 2009).

For readers familiar with the Case Assessment and Interpretation (CAI) model (Cook *et al*., 1998a, 1998b), use of which is also required by the AFSP Standards, please note that the description of the likelihood-ratio framework below is provided only at the source level. This is because it is the source level which is the most relevant level for forensic voice comparison. See Cook *et al*. (1998a) on the hierarchy of *source*, *activity*, and *offence* propositions. The activity level is seldom important in forensic voice comparison because issues of transfer and persistence are seldom pertinent: voice recordings are usually deliberately recorded, and those presented for forensic analysis are typically associated with warrants and chain-of-custody documentation. Authentication of audio recordings, and analysis of disputed utterances, should be considered areas of expertise which are distinct from forensic voice comparison (although an individual may be an expert on all three). In forensic voice comparison one must, however, consider the effects of the conversion of the acoustic signal to an electronic signal and often its transmission over a telephone system, which can result in relatively poor quality voice recordings and potentially mismatches between the recording quality of known and questioned samples (**[99.600]**, **[99.610]**). There may also be differences in speaking style, e.g., a lively telephone conversation on the recording of the questioned voice, and subdued answers to questions asked in a police interview on the recording of the known voice. The outcome of a forensic voice comparison may be of direct relevance for the offence level, for example, if the offence is uttering death threats and the questioned-voice recording is a recording of someone uttering death threats.

The likelihood-ratio framework

[99.150] In the likelihood-ratio framework the task of the forensic scientist is to provide the court with a *strength-of-evidence* statement in answer to the question:

> How much more likely are the observed differences between the known and questioned samples to occur under the hypothesis that the questioned sample has the same origin as the known sample than under the hypothesis that it has a different origin?

The answer to this question is quantitatively expressed as a *likelihood ratio*, calculated using Formula 1.

Formula 1

$$LR = \frac{p(E|H_{so})}{p(E|H_{do})}$$

where LR is the likelihood ratio; E is the evidence, i.e., the measured differences between the samples of known and questioned origin; $p(E \mid H)$ is "probability of E given H"; and H_{so} is the same-origin hypothesis, and H_{do} is the different-origin hypothesis.

If the evidence is more likely to occur under the same-origin hypothesis than under the different-origin hypothesis then the value of the likelihood ratio will be greater than one, and if the evidence is more likely to occur under the different-origin hypothesis than under the same-origin hypothesis then the value of the likelihood ratio will be less than one.

The size of the likelihood ratio is a numeric expression of the strength of the evidence with respect to the competing hypotheses. If the forensic scientist testifies that one would be 100 times more likely to observe the differences between the known and questioned samples under the same-origin hypothesis than under the different-origin hypothesis ($LR = 100$), then whatever the trier of fact's belief prior to hearing this, they should now be 100 times more likely than before to believe that the samples have the same origin. Likewise, if the forensic scientist testifies that one would be 1000 times more likely to observe the evidence under the different-origin hypothesis than under the same-origin hypothesis ($LR = 1/1000$), then whatever the trier of fact's prior belief, they should now be 1000 times more likely than before to believe that the samples have different origins.

The numerator of the likelihood ratio can be considered a *similarity* term, and the denominator a *typicality* term. In calculating the strength of evidence, the forensic scientist must consider not only the degree of similarity between the samples, but also their degree of typicality with respect to the relevant population. In fictional television shows, forensic scientists are often portrayed comparing two objects, finding no measurable differences between them, and shouting "It's a match!" Similarity alone, however, does not lead to strong support for the same-origin hypothesis. For example, if two samples are determined to be very similar in terms of some physical properties, this is of little value if these physical properties are also very typical and samples selected at random from any two individuals in the relevant population are likely to be equally or more similar. On the other hand, if two samples are found to be very similar in terms of properties which are very atypical in the population, then samples selected at random from any two individuals in the relevant population are unlikely to be equally or more similar. In general, more similarity and less typicality lead to relatively greater support for the same-origin hypothesis, and less similarity and more typicality lead to relatively greater support for the different-origin hypothesis.

Why the forensic scientist must present the probability of evidence, and must not present the probability of hypotheses

[99.160] A forensic likelihood ratio is an expression of the probability of obtaining the evidence given same- versus different-origin hypotheses. There are logical and legal reasons why the forensic scientist must present a strength-of-evidence statement in this form and must not present the probability of the hypotheses given the evidence.

Determining the probability of guilty versus not-guilty and whether this exceeds a threshold such as "beyond a reasonable doubt" or "on the balance of probabilities" is the task of the trier of fact. If the forensic scientist were to present the probability of same-origin versus different-origin and the evidence were potentially incriminatory, then they would be usurping the *rôle* of the trier of fact.

The trier of fact does not make their decision on the basis of a single piece of evidence, rather their task is to come to a decision after having weighed all the evidence presented in court (a conviction should never be based on a single piece of forensic evidence). What the trier of fact requires from a forensic scientist is a statement of the strength of a specific piece of evidence. One forensic scientist may present the strength of evidence related to specific DNA samples, another may present the strength of evidence related to specific fingermark samples, etc., and the trier of fact will weigh all of these together. Not all the evidence will be forensic comparison evidence evaluated using likelihood ratios, and the trier of fact must also consider the strength of other evidence such as eye-witness testimony. In addition, before any evidence has been presented the trier of fact will have some belief as to the innocence/guilt of the defendant, perhaps influenced by concepts such as "innocent until proven guilty", and this will also contribute to their final decision.

If a forensic scientist wanted to calculate the probability of same-origin versus different-origin hypotheses they would have to apply *Bayes' Theorem*. The odds form of Bayes' Theorem is provided in Formula 2.

Formula 2

$$\frac{p(H_{so}|E)}{p(H_{do}|E)} = \frac{p(E|H_{so})}{p(E|H_{do})} \times \frac{p(H_{so})}{p(H_{do})}$$

posterior odds	likelihood ratio	prior odds

In order to calculate the posterior odds (the relative probability of the same-origin versus the different-origin hypothesis, given the evidence), the forensic scientist would need to know the prior odds. Under one interpretation of Bayes' Theorem, the prior odds would represent the trier of fact's belief in the relative likelihood of the two hypotheses prior to the evidence being presented. Obviously, when conducting their analysis, the forensic scientist cannot know the trier of fact's prior belief. Under another interpretation pragmatic priors can be calculated, e.g., if a crime were committed on an island and there are known to have been 100 people on the island at the time, then pragmatic prior odds could be 1/99; however, this would involve the assumption that each person on the island is equally likely to have committed the crime, and although it may be appropriate for the trier of fact to make such an assumption, it is not appropriate for the forensic scientist to do so.

Also, if other evidence has already been presented in the trial, it is unlikely that the trier of fact's belief as to guilty versus not-guilty would still be 1/99 immediately prior to the presentation of the likelihood ratio from the forensic evidence in question.

It is inappropriate for the forensic scientist to present the posterior odds because the posterior odds include information and assumptions from sources other than a scientific evaluation of the known and questioned samples. If the forensic scientist were to present posterior odds then they would have to supply their own prior odds, and it would be possible that their testimony could be influenced by their own subjective conscious or unconscious opinion as to the guilt or innocence of the defendant. Human bias was a major concern in the NRC report (NRC, 2009, pp. 122–124). It is a strength of the likelihood-ratio framework that it is resistant to influence from such sources of bias.

Terminology

[99.170] Although the likelihood ratio is a component of Bayesian analysis, I have used the term *likelihood-ratio framework* rather than *Bayesian framework* since the latter, unlike the former, could imply that the forensic scientist makes use of priors and calculates posteriors (Buckleton, 2005; Champod & Meuwly, 2000; Rose, 2006). An alternative to the term *likelihood-ratio framework* used by some authors (e.g., Buckleton, 2005) is *logical approach*, I prefer *likelihood-ratio framework* because I believe it is more transparent, and it is also the established term within the forensic-voice-comparison branch of forensic science.

It should also be noted that the fact that forensic scientists present likelihood ratios in court does not imply that the trier of fact must assign numeric weights to evidence which is not forensic comparison evidence, nor that they must arrive at their decision via the rigid application of a Bayes'-Theorem formula such as Formula 2 (*R v Adams* [1996] EWCA Crim 222; *R v Adams* [1997] EWCA Crim 2474; *R v GK* [2001] NSWCCA 413; Balding, 2005, pp. 149–151; Buckleton, 2005; Donnelly, 2005; Morrison, 2009a).

Another terminological point is that in the likelihood-ratio framework the forensic scientist does not perform *identification* or *individualisation*, because these terms imply determining a posterior probability (see Cole, 2009, Saks & Koehler, 2008, and Meuwly, 2006, on terminological and logical problems with *identification* and *individualisation* in forensic science, see Kaye, 2010, for a contrary view). A neutral term such as *comparison* is more appropriate (French & Harrison, 2007). I therefore use the term "forensic voice comparison" rather than either of the traditional terms *forensic speaker identification* or *forensic speaker recognition* (*recognition* also implies a posterior probability, note also that *speaker comparison* would be akin to calling fingermark comparison *toucher comparison*). Following Meuwly's logic I should actually use a term such as *forensic comparison of voice recordings*, i.e., it is the properties of the recordings which are actually compared, not the voices themselves. Since the "of" construction has the potential to interfere with the understanding of sentence structure, I use the somewhat less exact term *forensic voice comparison*.

A database representative of the relevant population

[99.180] The likelihood-ratio framework is a conceptual framework which can be applied to subjective experience-based beliefs as to the likelihoods of the evidence given the competing hypotheses; however, to implement the data-based and quantitative-measurement aspects of the new paradigm, the forensic scientist must have access to a database of samples which are representative of the relevant population. Such a database (a *background database*) is necessary in order to calculate a quantitative estimate of the typicality of the known and questioned samples. Such a database is also needed to implement the validity and reliability testing requirements of the new paradigm **[99.290]**.

The relevant population is the population to which the offender belongs. In forensic voice comparison, this can usually be at least restricted to speakers of the same sex and general age speaking the same language and dialect as can be inferred for the questioned speaker on the basis of the questioned-voice recording. For example, if it were apparent that the speaker on the questioned-voice recording were an adult male (not obviously a child and not obviously very aged) speaking Australian English, and this would not be disputed by either the prosecution or the defence, then an appropriate database would be a database of voice recordings of adult male Australian-English speakers.

The exact nature of the relevant population is, however, dependent on the exact nature of the different-speaker hypothesis advanced by the defence. It would be reasonable for the defence to advance the hypothesis not that the offender is a male Australian-English speaker other than the defendant, but rather a speaker other than the defendant who "sounds like" the defendant, i.e., to a lay person with respect to forensic voice comparison, such as a police officer, a lawyer, a judge, or a jury member, the voice recordings of the offender and the defendant are sufficiently similar that they could generate the hypothesis that the voice recordings were produced by the same speaker – if the voices obviously sounded different then the prosecution would not have advanced the hypothesis that they belonged to the same speaker. In fact voices selected for inclusion in the database should be those which "sound like" the questioned voice recording, rather than the known voice. The defendant's voice (sampled on the known-voice recording) is one voice which sounds like the voice on the questioned-voice recording, which is what lead to the same-speaker hypothesis, and the background database should therefore consist of other voices which "sound like" the offender voice recording. The defence could also advance other hypotheses such as the offender is the defendant's brother who "sounds like" him.

As the relevant population becomes more specific and the voices in the database are restricted to those which to a lay person sound more similar to the voice of the defendant, then one might expect the size of the likelihood ratio to decrease, but note that each likelihood ratio is an answer to a different strength -of-evidence question and should be interpreted by the trier of fact accordingly. For example, for the sake of argument, imagine that prior to hearing the presentation of the forensic evidence the trier of fact believed that any one of a million adult male Australian-English speakers (other than the defendant) living in a geographical area was equally likely to have committed the crime, and the forensic scientist's testimony is that one would be ten thousand times more likely to observe the acoustic differences between the known- and questioned-voice recordings under the assumption that the voices on the known- and questioned-voice recordings were produced by the defendant than under the assumption that the voice on the questioned-voice recording was produced

by some other adult male Australian-English speaker. According to Bayes' Theorem (see Formula 2), the trier of fact's prior odds should be 1/1 000 000, the likelihood ratio is 10 000, and the posterior odds should therefore be 1/100 – one hundred in favour of the different-speaker hypothesis. The trier of fact should require a lot more additional evidence to reach a guilty verdict. Now, imagine that prior to hearing the presentation of the forensic evidence the trier of fact believed that only either the defendant or his brother could possibly have committed the crime and that each was equally likely to have committed the crime, and the forensic scientist's testimony is that one would be ten times more likely to observe the acoustic differences between the known- and questioned-voice recordings under the assumption that the voices on the known- and questioned-voice recordings were produced by the defendant than under the assumption that the voice on the questioned-voice recording was produced by the defendant's brother. The trier of fact's prior odds should be 1 (0.5/0.5), the likelihood ratio is 10, and the posterior odds should therefore be 10 – ten in favour of the same-speaker hypothesis. The trier of fact should still require additional evidence to reach a guilty verdict, but not as much as in the earlier example.

Note that at the time of conducting the forensic voice comparison, the forensic scientist may not be aware of the exact nature of the defence's hypothesis, and may have to anticipate what it will be.

For a broader discussion on the selection of the relevant population see Aitken & Taroni (2004, pp. 274–281), Lucy (2005, pp. 129–133), and Robertson & Vignaux (1995, pp. 33–50).

Differences between DNA data and voice data

[99.190] With respect to the calculation of forensic likelihood ratios, there are some important differences between data extracted from DNA samples and data extracted from voice-recording samples. These differences may lead to differences in the way the results of forensic-DNA-profile comparison and forensic-voice-comparison are presented, which may superficially give the incorrect impression that the two are not evaluated using the same framework. In fact, both DNA-profile-comparison evidence and voice-comparison evidence can and should be evaluated using the likelihood ratio framework.

This section includes a simplified account of forensic-DNA-profile comparison. My purpose is to highlight some basic differences between DNA and voice data, not to discuss issues in the interpretation of DNA evidence. Readers interested in the latter may wish to consult resources such as **Statistical Evaluation in Forensic DNA Typing [80A]** (Roberts, 2004), Balding (2005), and Buckleton, Triggs, & Walsh (2005).

DNA-profile data consist of discrete values (e.g., counts of short tandem repeats) from a finite number of measurements (e.g., pairs of alleles at specific loci). DNA properties are discrete at the molecular level, their values are continuous at the measurement level (which can, for example, be graphically represented as the location and height of peaks on an electropherogram), but they are typically converted back to discrete values to provide data for statistical analysis. It is the latter to which I refer when I use the term "DNA profile". For simplicity I will assume (unrealistically) that it is always the case that DNA profiles have no measurement errors, that samples are not contaminated, that the organisms from which DNA samples originate have not undergone transplants, etc. It is possible to obtain a "match" between two DNA profiles, i.e., for each corresponding locus and allele each of the two profiles has the same discrete value. Under the

assumptions laid out above, the DNA profile of an individual organism does not change from occasion to occasion, hence the probability of obtaining matching DNA profiles given the same-origin hypothesis is one, and the probability of obtaining non-matching DNA profiles given the same-origin hypothesis is zero. The numerator of the likelihood ratio is therefore either 1 or 0 (Aitken & Taroni, 2004, p. 404; Evett, 1998).

If the two samples do not match, the numerator of the likelihood ratio is 0 and the denominator is irrelevant, the value of the likelihood ratio is 0 and via Bayes' Theorem the posterior odds will also be 0, the two samples do not have the same origin.

If the two samples match, the numerator of the likelihood ratio is 1 and the size of the likelihood ratio is then dependant on the denominator, the probability of the DNA profile of the questioned sample matching the DNA profile of the known sample if the questioned sample comes from a source other than the known organism.

Often when the samples match the *match probability* rather than the likelihood ratio is reported in court (*R v Doheny & Adams* [1996] EWCA Crim 728 directed DNA experts to provide match probabilities; see also Evett, 1998; and Balding, 2005, pp. 151–153). The match probability is simply the denominator of the likelihood ratio, or equivalently the inverse of the likelihood ratio given in Formula 1, i.e., it is the probability of obtaining the matching DNA profile in question under the different-origin versus the same-origin hypothesis (Balding, 2005, p. 24; Foreman *et al*., 2003, p. 484).

An acoustic-phonetic or automatic forensic-voice-comparison system would be based on measurements of acoustic properties of voices (see **[99.700]** and **[99.720]**). These acoustic properties are continuous, not discrete. There is also substantial within-speaker variation, even if a speaker says exactly the same words twice in a row it would be extremely unlikely for there not to be measurable differences in the acoustic properties of the two utterances. Note that this is not just the precision of the measurement techniques, it is also intrinsic variability in the source. In practice a speaker is unlikely to repeat long stretches of exactly the same words, and there will likely also be variability due to factors such as phonetic context and speaking style (and also often channel effects, **[99.600]**, **[99.610]**).

For continuously valued properties with this sort of variation a "match", in terms of two samples being indistinguishable within the precision of measurement techniques, or in terms of not having (at some frequentist alpha level) a statistically significant difference for the combination of intrinsic and measurement variability, or in terms of some pre-determined difference threshold (whether experience or empirically based), suffers from a cliff-edge effect (Robertson & Vignaux, pp. 118–120; Evett, 1998; Rose & Morrison, 2009). For example, if the threshold were set at 10 Hz then a value of 9.99 Hz would be declared a match, but an almost identical value of 10.01 Hz would be declared a non-match (think about an airline charging you for excess baggage if your suitcase is 1 g over the specified weight limit). "Match" is therefore not a useful concept for the acoustic properties of voices (the same can probably also be said for the properties of objects of comparison in many other branches of forensic science).

The numerator of a likelihood ratio calculated from a forensic voice comparison cannot therefore be either 0 or 1, a match probability cannot be calculated, and the results must be reported in the form of a full likelihood ratio.

Some would argue that, since the simplifications made above with respect to DNA-profile comparison are not valid, DNA results should also be reported as full likelihood ratios (personal communication from Didier Meuwly, April 2009). See Kaye & Sensabaugh (2008, §30:41) on problems with the process of converting continuous electropherogram data to discrete values.

Calculating a forensic likelihood ratio

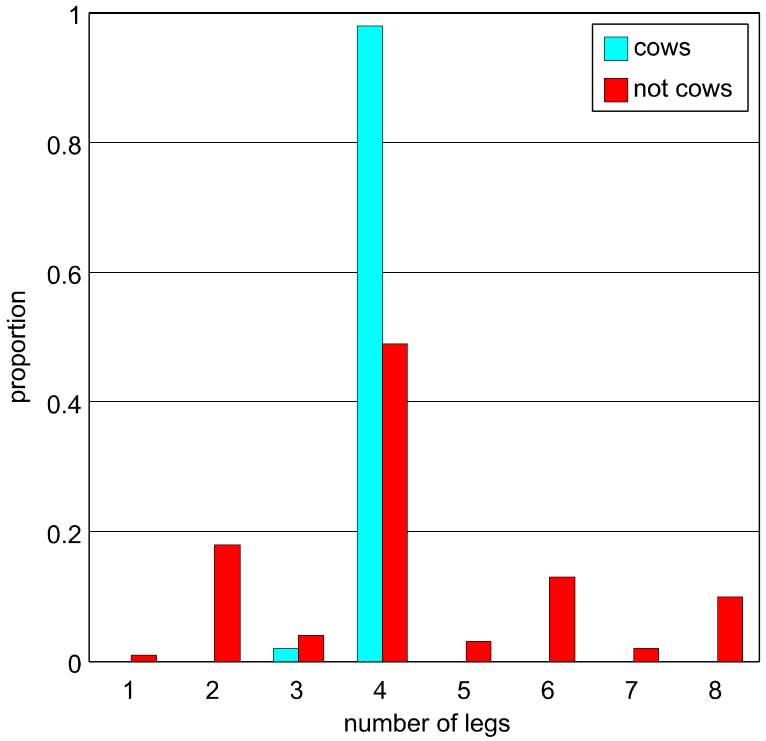
[99.200] This section describes how to calculate a forensic likelihood ratio at a general conceptual level, not at a detailed mathematical level. At a detailed mathematical level there are multiple different procedures for calculating forensic likelihood ratios, many of which are much more complicated than those presented here. The aim of this section is to provide the reader with a basic understanding of how a forensic likelihood ratio is calculated and also of some factors affecting the size of the likelihood ratio. All the data presented in this section are simple artificial data designed for illustrative purposes and they are not intended to be realistic.

Calculating a forensic likelihood ratio from discrete data

[99.210] Let us begin with a fanciful discrete-data example. Imagine the competing hypotheses are H_1: "the animal is a cow", and H_2: "the animal is not a cow", and our evidence consists only of a count of the number of legs on the animal. First we need some data, I go out to the countryside and look for animals and whenever I see an animal I record whether it is or is not a cow and the number of legs that it has (assume that animals only have whole numbers of legs, no half legs etc., also assume that at this stage there are no disputes about what is and what is not a cow). At the end of the day I calculate the proportion of the total number of cows which had one leg, two legs, three legs, four legs, etc. I do the same for non-cows. I display this information graphically as the *histograms* in Figure 1. It turns out that 2% of the cows I saw had three legs and the rest had four legs. I also saw some sheep and horses, most with four legs but some with three, some ducks and chickens including a one-legged duck, and also some insects and spiders (I didn't see any snakes or earthworms, or centipedes or millipedes).

Now I am told that the evidence is that the animal in question has four legs. How do I calculate the likelihood ratio p (4 legs | cow) / p (4 legs | not cow)? In Figure 1 I go to number of legs = 4, and take the relative proportions of cows with four legs and non-cows with four legs: 0.98 / 0.49 = 2. Having four legs is twice as likely if the animal is a cow than if it is not a cow. Whatever one believed before hearing this evidence, one should now be two-times more likely than before to believe that the animal is a cow.

FIGURE 1. Histograms of discrete data.



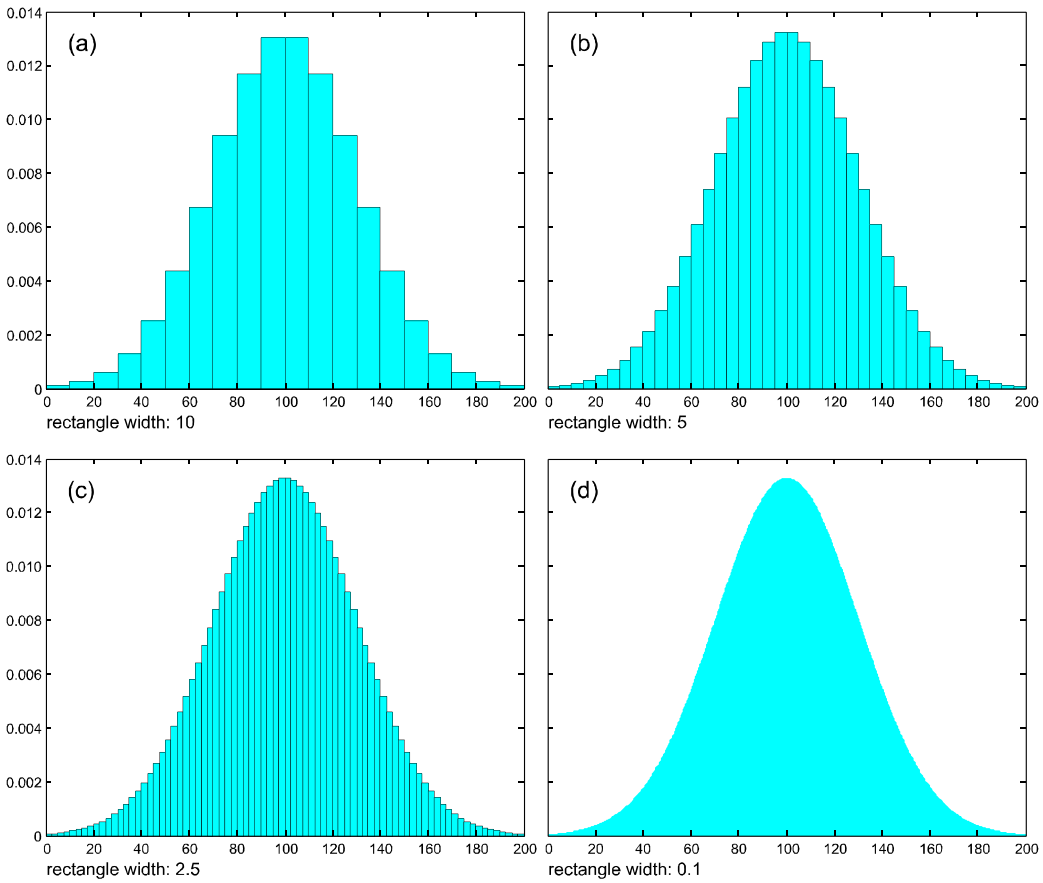
From discrete data to continuous data

[99.220] As noted in **[99.190]** voice data are normally continuous, not discrete. For calculations based on continuous data, *histograms* are replaced by *probability density functions*. In a histogram of continuous data there are no gaps between the rectangles and each covers a range of values. For example, if each rectangle is 10 units wide then one rectangle could cover the range $40 \leq x < 50$ (x is greater than or equal to 40 and less than 50), and the next rectangle would cover the range $50 \leq x < 60$ (see Figure 2a). The area of a rectangle represents the proportion of data points that fall within the range it covers, e.g., if 2.5% of the data fall in the range $40 \leq x < 50$ then the rectangle will be 0.0025 units tall to give it an area of $0.0025 \times 10 = 0.025$. The sum of the areas of all the rectangles must equal 1.

Now, imagine that we have a very large amount of data so that we can reduce the widths of the rectangles and still have enough data to be able to calculate a meaningful value for the proportion of data points within each rectangle's range. Say we start by reducing the width of each rectangle to 5 units, one rectangle could cover the range $40 \leq x < 45$ and the next $45 \leq x < 50$, etc. (see Figure 2b). We now see more detail in how the proportions change as the x value changes. As before, the area of the rectangle represents the proportion of data points which falls within the range it covers,

e.g., if 1% of the data points fall in the range $40 \leq x < 45$ then the rectangle will be 0.002 units tall to give it an area of $0.002 \times 5 = 0.01$. The sum of the areas of all the rectangles must still equal 1.

FIGURE 2. Transition from histogram to probability density function for continuous data.



As the widths of the rectangles are reduced (Figures 2a through 2d), the size of the steps between rectangles decrease, not just the widths of the steps but also their height differences (assuming the proportions change gradually and do not have discontinuities). Eventually the tops of the rectangles will look like a smooth curve rather than a series of steps (see Figure 2d). If we make some assumptions about the shape of this curve, such as that it is a *Gaussian* curve (also called a normal curve), then even with relatively small amounts of data we can skip straight to an estimate of the shape of the curve. For a Gaussian curve we only need to estimate the mean and standard deviation. Note that the total area under the curve is still equal to 1.

The curve is the calculated probability density function of the data. When combined with any prior belief about the shape of the probability density function of the relevant population, the calculated probability density function of the background data can be used to estimate the shape of the probability density function of the relevant population (if uniform priors are assumed, the calculated probability density function of the background data is the estimate of the probability density function of the relevant population).

Calculating a forensic likelihood ratio for continuous data

[99.230] As mentioned above **[99.220]**, for calculations based on continuous data, histograms are replaced by probability density functions, but otherwise the same procedures as in the discrete-data example **[99.210]** can be followed.

Let us imagine this time that each of our data points is a measurement of the mean fundamental frequency (f0) of a voice in a voice recording. This voice property is described in **[99.540]**, what matters here is that f0 can differ between speakers (some speakers have a higher mean f0 value and others a lower mean f0 value), and also within speakers (on one occasion a speaker may produce a higher mean f0 value and on another occasion a lower mean f0 value).

We collect a database of voice recordings of speakers from the relevant population and measure the mean f0 of each recording and calculate the probability density function for these values. This is plotted in Figure 3. Likewise we collect multiple non-contemporaneous recordings of the voice of the known speaker and calculate the probability density function for the mean f0 from each of these recordings. This is also plotted in Figure 3. The former probability density function is known as the *background model*, and the latter is known as the *suspect model* (the speaker whose identity is known is usually the suspect).

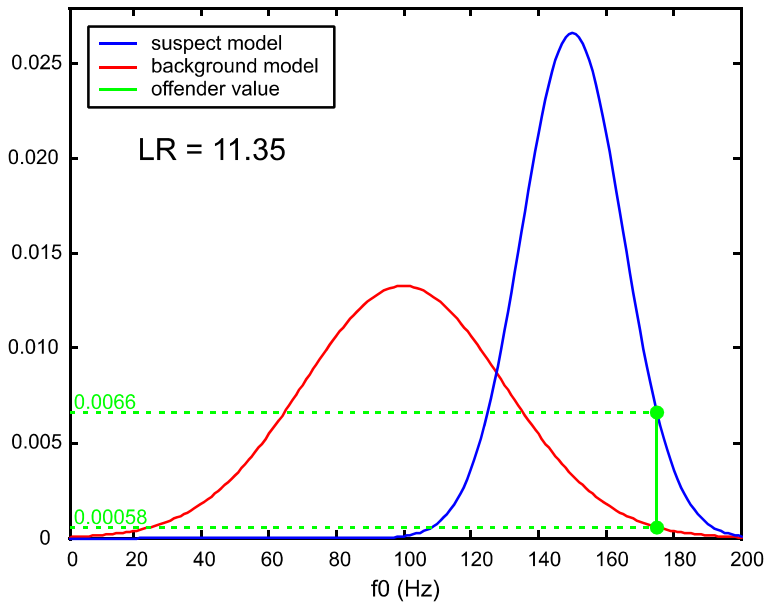
In this example the background model is a Gaussian distribution with a mean of 100 Hz and a standard deviation of 30 Hz, and the suspect model is a Gaussian distribution with a mean of 150 Hz and a standard deviation of 15 Hz.

To calculate a likelihood ratio, we find the mean f0 value of the questioned voice in the questioned-voice recording, then find the relative heights of the suspect-model and background-model curves in Figure 3. If the offender value is 175 Hz, the probability-density-function value of the suspect model at 175 Hz is 0.0066, the probability-density-function value of the background model at 175 Hz is 0.00058, and via Formula 1 the likelihood ratio is therefore 0.0066 / 0.00058 = 11.35. One would be approximately 11 times more likely to obtain the f0 value of 175 Hz in the questioned-voice sample if the questioned-voice sample had been produced by the suspect than if it had been produced by a different speaker.

Theoretically we should build an offender model from the questioned-voice data (the questioned voice is usually that of the offender) and test at the value extracted from the suspect sample, but in practice we do not normally have enough questioned-voice data to be able to do this. In the present example we would need multiple recordings of the questioned voice. In casework there may only be one recording of the questioned voice and it is not possible to obtain more (we don't know who the questioned speaker is). In contrast, it is usually possible to obtain additional recordings of the known voice. Even if there are multiple recordings which are ostensibly of the questioned speaker, a forensic scientist cannot assume that all these are recordings of the same questioned speaker,

whereas this is not a problem for multiple known-voice recordings. In other cases we might extract data from multiple tokens of a particular speech sound within a single recording, but again the limiting factor is that questioned-voice recordings are usually quite short whereas this is of less concern for known-voice data.

FIGURE 3. Calculation of a likelihood ratio from a suspect model and a background model.



What if, instead of 175 Hz, the mean f0 of the questioned-voice recording was 150 Hz, right at the mean value for the known-voice recordings? In this case, as shown in Figure 4, instead of 11.35, the likelihood ratio would be 8.02.

At first it may seem counter-intuitive that an offender value closer to the suspect mean has a lower likelihood ratio than one which is further away, but note that it is also closer to the background-model mean, i.e., closer to our estimate of the mean for the relevant population – the questioned-voice sample is more typical.

Note that even if there is no measurable difference between the mean f0 in the suspect recordings and the mean f0 in the offender recording (a situation in which some would say the suspect and offender data are "identical"), this does not lead to a positive identification of the offender as being the suspect. In fact, in this example, one would be only 8 times more likely to obtain no measurable difference between the known-voice and questioned-voice samples if the questioned-voice sample had been produced by the known-speaker than if it had been produced by someone other than the known-speaker.

FIGURE 4. Calculation of a likelihood ratio from a suspect model and a background model.

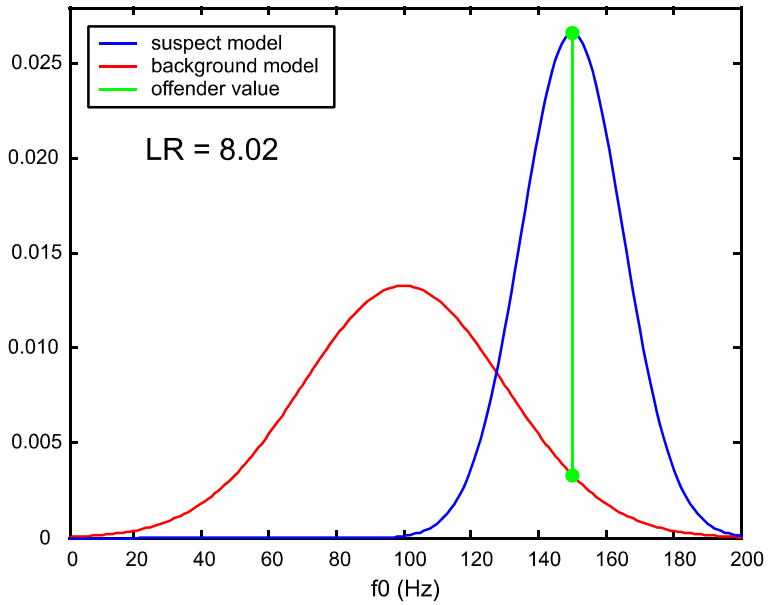
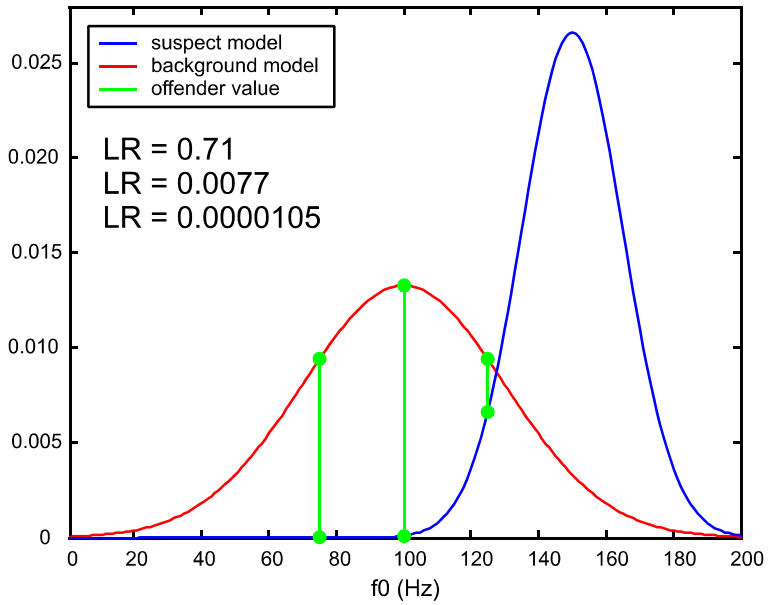


FIGURE 5. Calculation of likelihood ratios from a suspect model and a background model.



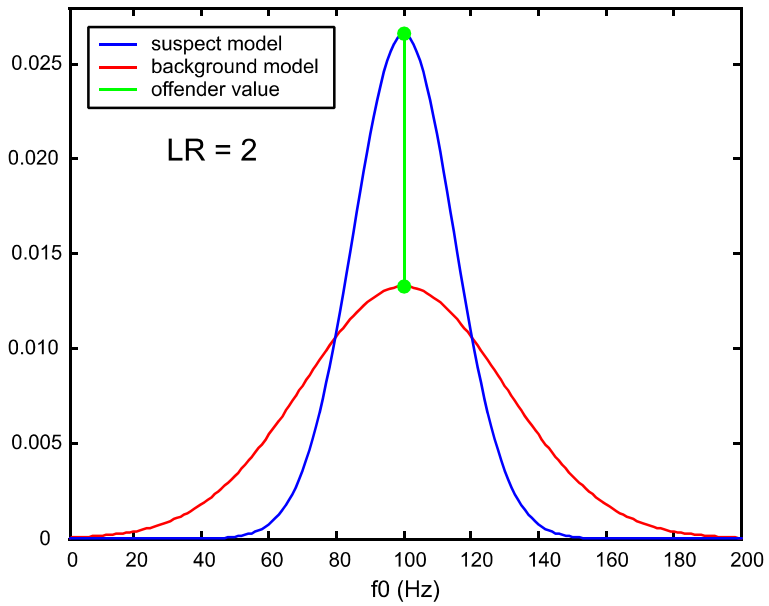
What if the questioned-voice sample were even more typical and had a mean f0 of 125 Hz or 100 Hz? In these cases, as shown in Figure 5, the likelihood ratios would be 0.71 (0.71 in favour of the same-speaker hypothesis, or 1 / 0.71 = 1.42 in favour of the different-speaker hypothesis) and 0.0077 (129 in favour of the different-speaker hypothesis) respectively.

If the questioned-voice sample is atypical in the opposite direction to the atypicality of the known-voice samples, then the support for the different-speaker hypothesis is even higher, e.g., if the questioned-voice sample has a mean f0 of 75 Hz then the likelihood ratio is 94 810 in favour of the different-speaker hypothesis.

Note that at a value of approximately 128 Hz, the likelihood ratio would be 1 – one would be equally likely to obtain this value irrespective of whether the questioned-voice recording had been produced by the known-speaker or by a different speaker.

In the previous examples the suspect-model was relatively atypical. What if the known-voice recordings were more typical? In Figure 6 the suspect model has the same mean as the background model (100 Hz), and is thus maximally typical. The standard deviations are unchanged from the previous examples. As was the case in Figure 4, the mean f0 value for the questioned-voice sample is at the mean value for the suspect model, but instead of being 8.02, because the suspect model is now more typical the likelihood ratio is only 2.

FIGURE 6. Calculation of a likelihood ratio from a suspect model and a background model.



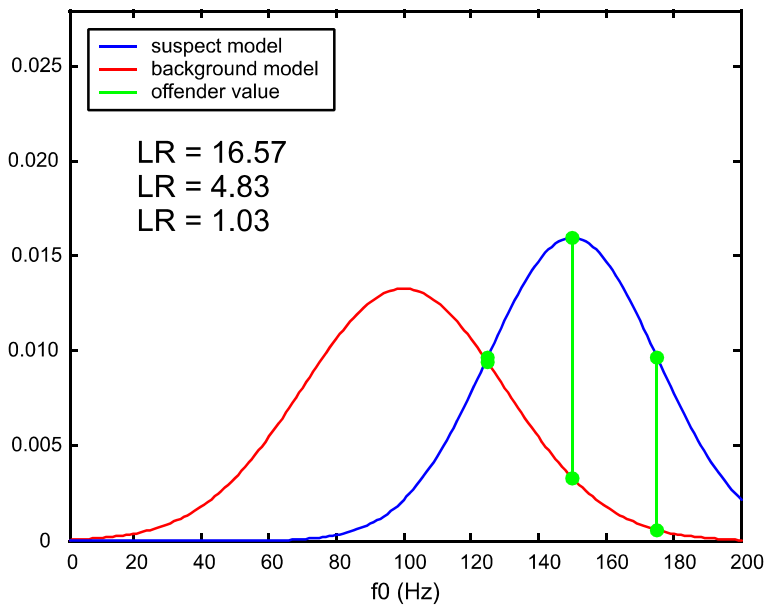
Note that even though in this example the suspect model is maximally typical and has the same mean as the background model, the likelihood ratio is not 1, and one is still more likely to obtain a

mean f0 at the maximally typical value if the questioned-voice sample had been produced by the known speaker than if it had been produced by some other speaker. This is because not all speakers in the population are maximally typical and because some are atypical they are less likely to produce the maximally typical mean f0 value, which contributes to this also being less likely for the population as a whole.

What if the within-speaker variability were greater? The suspect model in Figure 7 has a standard deviation of 25 Hz as opposed to 15 Hz as was the case in the earlier examples. The mean for the suspect model is 150 Hz as was the case in Figures 3 through 5. The first-three questioned-speaker values of 175 Hz, 150 Hz, and 125 Hz, which previously resulted in likelihood ratios of 11.35, 8.02, and 0.71, now result in likelihood ratios of 16.57, 4.83, and 1.03. Questioned-voice values close to the suspect-model mean now result in smaller likelihood ratios and questioned-voice values relatively far from the suspect-model mean now result in larger likelihood ratios.

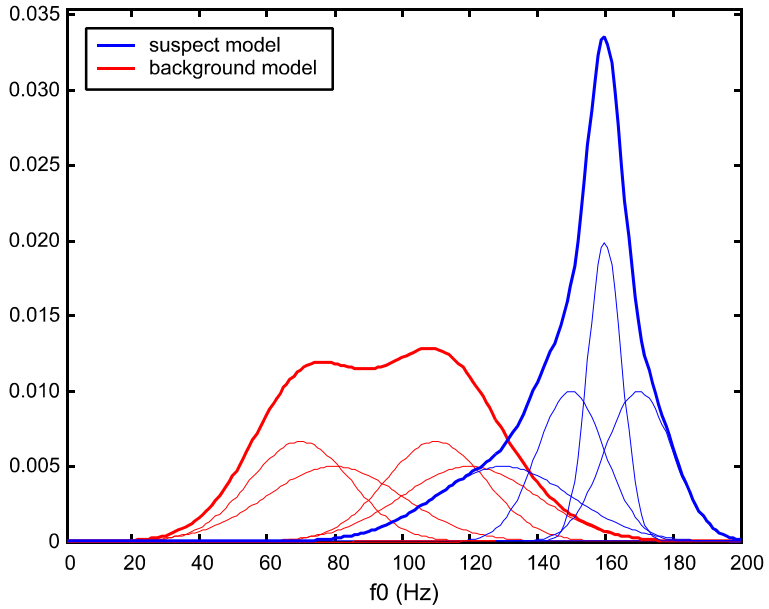
In general, the smaller the within-speaker variability relative to between-speaker variability, the better the performance of the forensic-comparison system (system validity and reliability are discussed in **[99.290]**). Most speakers will be relatively typical (by definition) so most suspect models will have means close to the background-model mean, and, as the within-speaker variance approaches the between-speaker variance, most speaker models will have variances close to the background-model variance. The suspect-model and background-model curves will get closer together, and therefore most likelihood ratios will approach 1 and not provide strong support for either hypothesis.

FIGURE 7. Calculation of likelihood ratios from a suspect model and a background model.



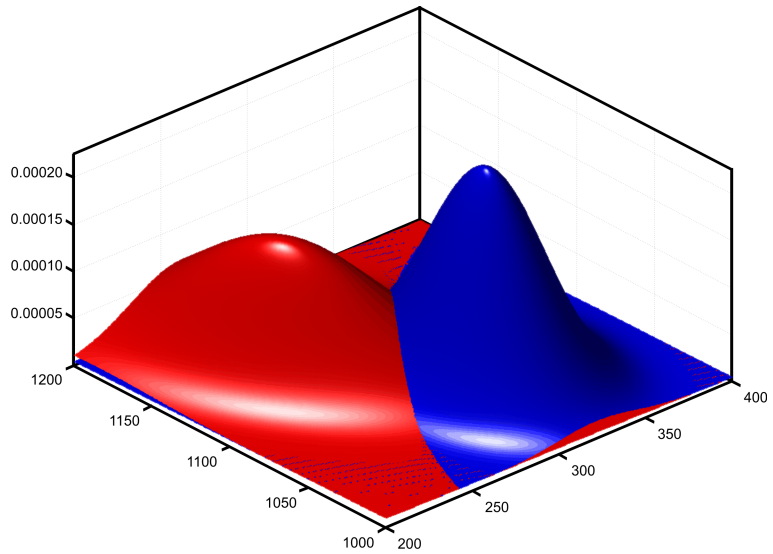
All the previous examples have used models consisting of a single Gaussian as a background model and a single Gaussian as a suspect model; however, more complex models are usually used in forensic voice comparison. For example, procedures based on *Gaussian mixture models* (GMMs) are common. Multiple Gaussians are used to fit a more complex distribution than can be achieved using a single Gaussian. Figure 8 provides an example of a background and a suspect model each based on four Gaussians. The GMMs, shown as the thick lines, are the result of summing the individual Gaussians shown as thin lines (in this case each individual Gaussian is given equal weight).

FIGURE 8. Gaussian mixture models.



All the previous examples have been unidimensional, using measurements of a single acoustic property of the voice samples. In practice forensic voice comparison is usually based on measurements of multiple acoustic properties of voice samples and this has the potential to lead to much larger likelihood ratios in favour of one hypothesis or the other. A voice sample may be only moderately atypical on measurements of each of a number of acoustic properties, but the particular combination of these measured values may be highly atypical. Figure 9 provides an example of a background model and a suspect model (each a single Gaussian) in a two-dimensional space. Use of multidimensional Gaussian mixture models is common in practice.

FIGURE 9. Two-dimensional suspect and background models.



Calibration and fusion

[99.240] The models used in the description of the calculation of likelihood ratios above **[99.230]** are theoretically correct, but in real life there may be a number of practical difficulties. These could, for example, be related to whether the model is appropriate for the true structure of the distribution of the data, whether there are sufficient data to build models which are sufficiently accurate and precise estimates of the true distributions, or whether aspects of the likelihood-ratio calculation procedure violate statistical assumptions. There is also the problem of how to combine multiple estimates of likelihood ratios on the same data, e.g., multiple spectral measurements taken at different points in time in the speech in the offender data (see **[99.720]**), or from different but parallel data, e.g., likelihood ratios based on different phonemes in the same recordings (see **[99.700]**).

The practical solutions to these problems are called *calibration* and *fusion*, and a single procedure, *logistic regression*, can be used to do both (Brümmer *et al.*, 2007; Brümmer & du Preez, 2006; González-Rodríguez *et al*., 2007; Morrison, 2009c; Morrison & Kinoshita, 2008; Morrison, Thiruvanran, & Epps, 2010a; Pigeon, Druyts, & Verlinde, 2000; Ramos Castro, 2007; van Leeuwen & Brümmer, 2007). One way to view calibration is to consider the raw likelihood ratios calculated using the sorts of procedures described above **[99.230]**, not as likelihood ratios *per se*, but rather simply as *scores* which quantify the degree of similarity of pairs of samples while also taking account of their typicality (but which are not directly interpretable as likelihood ratios). Calibration then converts a single set of scores into likelihood ratios, or fusion converts multiple parallel sets of scores into likelihood ratios (the latter akin to multivariate likelihood-ratio calculation, see Figure 9).

Further reading

[99.250] General descriptions of the likelihood-ratio framework can be found in numerous textbooks and articles including Aitken & Taroni (2004), Balding (2005), Buckleton (2005), Evett (1998), Lucy (2005), Robertson & Vignaux (1995). Of these, Robertson & Vignaux (1995) and parts of Balding (2005) may be most accessible for a reader with a limited technical background. Descriptions of the application of the likelihood-ratio framework to forensic voice comparison can be found in several textbooks and articles including Champod & Meuwly (2000), González-Rodríguez *et al*. (2006), González-Rodríguez *et al*. (2007), and Rose (2002, 2003, 2006).

ASSESSING THE VALIDITY AND RELIABILITY (ACCURACY AND PRECISION) OF FORENSIC-COMPARISON SYSTEMS

[99.290] The issues of validity and reliability are of great concern in forensic science (*Daubert v Merrell Dow Pharmaceuticals* (92-102) 509 US 579 [1993]; Cole, 2006; Law Commission of England & Wales, 2009; NRC, 2009; Saks & Faigman, 2008; Saks & Koehler, 2005); however, in judicial and forensic-science literature the word *reliability* has often not been explicitly defined or has been defined in terms of a measure of validity. In statistics and scientific literature *validity* and *reliability* mean different things – the first is synonymous with *accuracy* and the second with *precision*.

To illustrate the difference between accuracy and precision, imagine a device for measuring a person's height. It consists of a base which sits on the ground, a vertical pole with marks on it indicating distance from the top of the base, and a horizontal arm which slides up and down the pole. A person stands on the base, the arm is placed on top of their head, and the distance from the top of the base to the bottom of the arm is read off as the value marked on the pole. This is taken as a measure of the person's height.

Now imagine that this device is broken and rather than being vertical (fixed at 90° to the base), the pole is somewhat loose and sometimes the person is measured with the pole at 85°, other times at 95°, and various other angles in between. For the sake of argument, let us also assume that a person's height is fixed and that we have an oracle who can tell us a person's true height. We measure the same person's height multiple times using the broken device. Sometimes we measure their height as 177 cm, sometimes as 173 cm, and other values in between. We take the mean of all the measurements and we find it to be 175.1 cm. The oracle tells us that in fact the true height of this person is 175.0 cm. Our measuring device is very accurate, averaging over multiple measurements it has come up with an answer which is only 1 mm (0.057%) away from the true value. In contrast, the measuring device is not very precise, our measurements range from approximately 2 cm below to 2 cm above the mean value.

Now imagine that the measuring device has been repaired and the pole is now at 90° to the base. We measure the same person again multiple times and we get values which range from 176.9 cm to 177.1 cm with a mean of 177.0 cm. The device is now much more precise, our measurements only range from 1 mm below to 1 mm above the mean value, but its accuracy is now poor, the mean of our measurements is 2 cm too high! Upon inspection, we discover that as part of the "repair" the pole was made shorter, removing 2 cm from the bottom.

Ideally, for any system, we would like to have both a high degree of accuracy and a high degree of precision.

Measuring the accuracy of a forensic-comparison system

[99.300] The accuracy of the output of a forensic comparison system can be assessed by testing it on a large number of pairs of samples (a test set) where it is known for each pair whether its members have the same origin or different origins, then comparing the system's output with this knowledge about the input.

A measure of accuracy which has often been presented in the forensic-science literature (e.g., Found & Rogers, 2008; NRC, 2009, pp. 116–122; Saks & Koehler, 2005) is correct-classification rate, i.e., the proportion of true positives (the proportion of same-origin pairs correctly classified as same origin) and the proportion of true negatives (the proportion of different-origin pairs correctly classified as different origin); or alternatively, classification-error rates i.e., the proportion of false positives (the proportion of different-origin pairs incorrectly classified as same origin) and the proportion of false negatives (the proportion of same-origin pairs incorrectly classified as different origin). Classification-error rates are simply the inverse of correct-classification rates.

Classification-error rates (and correct-classification rates) are the result of binary (same or different) decisions made on the basis of posterior probabilities. Because it is based on posterior probabilities, this approach is inconsistent with the likelihood-ratio framework. The binary nature of the decisions is also inconsistent with the likelihood-ratio framework.

Likelihood ratios greater than one favour the same-origin hypothesis and likelihood ratios less than one favour the different-origin hypothesis; however, forensic comparison of known and questioned samples is not a binary decision task but rather the task of determining the strength of evidence with respect to the same-origin versus different-origin hypotheses, i.e., the extent to which likelihood ratios are greater than or less than one, equivalently the extent to which log likelihood ratios are greater than or less than zero.

It is often convenient to convert likelihood ratios to *log likelihood ratios* since the latter are symmetrical about zero, e.g., likelihood ratios of 1000 (one thousand in favour of the same-origin hypothesis) and 1/1000 (one thousand in favour of the different-origin hypothesis) become log-base-ten likelihood ratios of +3 and −3 respectively, and likelihood ratios of 10 000 and 1/10 000 are log-base-ten likelihood ratios of +4 and −4 respectively. A likelihood ratio of 1 is a log likelihood ratio of 0.

Ideally, for a same-origin pair the forensic-comparison system should produce a large positive log likelihood ratio, and for a different-origin pair it should produce a large negative log likelihood ratio. For a same-origin comparison, a small positive log likelihood ratio is not as good as a large positive log likelihood ratio, a small negative log likelihood ratio is worse than a small positive log likelihood ratio, and a large negative log likelihood ratio is worse than a small negative log likelihood ratio (*mutatis mutandis* for a different-origin comparison). It is worse to report a likelihood ratio of 1000 in favour of a contrary-to-fact hypothesis than it is to report a likelihood ratio of 10 in favour of a contrary-to-fact hypothesis, because the former provides greater support for the contrary-to-fact hypothesis and therefore has greater potential to contribute towards a miscarriage of justice.

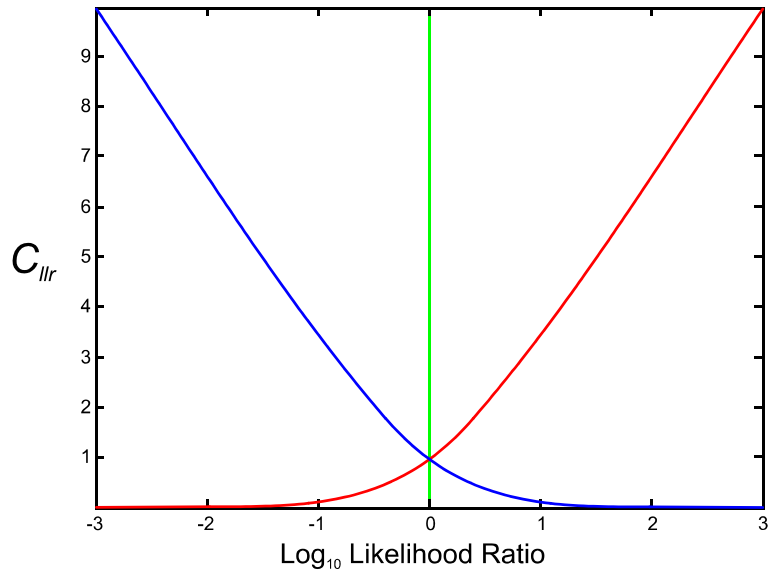
A measure of accuracy which is consistent with the likelihood-ratio framework is the *log-likelihood-ratio cost* (C_{llr}; Brümmer & du Preez, 2006; van Leeuwen & Brümmer, 2007). C_{llr} was developed for use in automatic speaker recognition and has subsequently been applied to forensic voice comparison, e.g., González-Rodríguez *et al*. (2007), Morrison (2009c), Morrison & Kinoshita

(2009), Ramos Castro (2007). In contrast to classification-error rates, C_{llr} has the desired properties of being based on likelihood-ratios, and of being continuous and more heavily penalising worse results.

To calculate C_{llr}, one must first calculate a C_{llr} component value for the likelihood ratio from each test pair. Figure 10 provides a plot of the function for calculating a C_{llr} component value when the input to the system is a same-origin pair (blue line). Large positive log likelihood values which correctly support the same-origin hypothesis are assigned very low C_{llr} component values, log likelihood values close to zero provide little support for either the same-origin or different-origin hypothesis and are assigned moderate C_{llr} component values, and negative log likelihood values which contrary-to-fact support the different-origin hypothesis are assigned high C_{llr} component values which increase rapidly as the log likelihood values become more negative and provide stronger contrary-to-fact support for the contrary-to-fact different-origin hypothesis. The function for calculating a C_{llr} component value when the input to the system is a different-origin pair (red line in Figure 10) is a mirrored version of the same-speaker function.

To calculate C_{llr}, one finds the mean of all the C_{llr} component values from same-origin pairs, the mean of all the C_{llr} component values from different-origin pairs, and then takes the mean of the latter two means. The formula for calculating C_{llr} is given in Formula 3.

FIGURE 10. Plot of the function for calculating a C_{llr} component value for a same-origin comparison (blue line) and a different-origin comparison (red line).



Formula 3

$$C_{llr} = \frac{1}{2}\left(\frac{1}{N_{ss}}\sum_{i=1}^{N_{ss}}\log_2\left(1+\frac{1}{LR_{ss_i}}\right)+\frac{1}{N_{ds}}\sum_{j=1}^{N_{ds}}\log_2\left(1+LR_{ds_j}\right)\right)$$

where N_{ss} and N_{ds} are the number of same-speaker and different-speaker comparisons, and LR_{ss} and LR_{ds} are the likelihood ratios derived from same-speaker and different-speaker comparisons. A same-origin component value is $\log_2(1 + 1/LR_{ss})$, and a different-origin component value is $\log_2(1 + LR_{ds})$.

The lower the C_{llr}, the better the performance of the system. If several systems are tested using the same set of test data, then the most accurate system is the system which results in the lowest C_{llr} value.

It is important to note that (as with other measures of accuracy such as classification-error rates) C_{llr} depends on the test data as well as the forensic comparison system. To be meaningful in casework, the test set should therefore be drawn from a database which samples the relevant population **[99.180]**, and the quantity and quality of each test pair should be matched as closely as possible to the quantity and quality of the known and questioned samples, e.g., for voice recordings this would include attempting to match duration, recording quality, and speaking style. In this way, the accuracy of the system can be assessed under conditions appropriate to the case.

Within the likelihood ratio framework it is also possible to report an *error rate* for the specific likelihood ratio which is calculated for the comparison of the known and questioned samples. For example, if a likelihood ratio of 100 in favour of the same-origin hypothesis is obtained, an error rate can be reported as the proportion of different-origin pairs in the test data which resulted in likelihood ratios of equal to or greater than 100 in favour of the same-origin hypothesis.

Examples of measurement of the accuracy of forensic-voice-comparison systems are provided in **[99.770]**.

Measuring the precision of a forensic-comparison system

[99.310] It is important to consider the precision of a forensic-comparison system as well as its accuracy. All else being equal, a system which outputs a likelihood ratio of 1000 in response to a pair of samples known to have the same origin is preferable to a system which outputs 100 for the same comparison. However, if additional pairs of samples taken from the same source were used to test the systems multiple times and the values produced by the former system ranged from 10 in favour of the different-speaker hypothesis to 10 000 000 in favour of the same-speaker hypothesis, whereas for the latter system they ranged from 79 to 126 in favour of the same-speaker hypothesis, then the latter would have better precision and would be preferred over the former. In any particular instance, such as when comparing the known- and questioned-voice recordings in casework, the former system may vastly overestimate or vastly underestimate the strength of evidence, but we would not know where in the range the particular result fell. The latter system is on average more conservative in expressing support for the same-origin hypothesis (let us assume that this hypothetical system is symmetrically conservative and in other instances it would also be

conservative in expressing support for the different-origin hypothesis, i.e., it produces smaller magnitude log likelihood ratios in general), and on average it may be less accurate, but we would be less concerned as to whether the results in a particular instance came from the top or the bottom of the range. We might be willing to say that the likelihood-ratio values for the latter system, including anywhere from the bottom to the top of the range, provide roughly the same strength of evidence; whereas it would seem entirely unreasonable to say the same for the former system.

In contrast to research on accuracy, there has been much less research aimed at developing a measure of the precision of likelihood-ratio results. Morrison (2010), Morrison, Thiruvaran, & Epps (2010b), and Morrison, Zhang, & Rose (2010) describe empirical procedures for estimating the 95% credible interval for the likelihood-ratio output of a forensic-comparison system. Assuming sufficient test data are available, for any particular likelihood ratio value (such as the value obtained from the comparison of the known- and questioned-voice samples in casework) this allows the forensic scientist to state the range of values which they are 95% certain contains the true value of the particular likelihood-ratio value. For each different-speaker and same-speaker pair in the test set, if multiple recordings are available, multiple comparisons are made using non-overlapping pairs of recordings (e.g., if speaker A recording 1 is compared with speaker B recording 1 and A2 with B2, then A1 with B2 and A2 with B1 are not compared since they overlap with the earlier pairs). For each different-speaker and same-speaker pair, the mean of the log-likelihood-ratio results is calculated and for each log-likelihood-ratio result its distance from this mean is calculated (deviation-from-mean value). The estimate of the 95% credible interval is a calculation of the boundary between the 95% of deviation-from-mean values which are closest to the means and the 5% which are furthest from the means. This approach is applicable to forensic-voice comparison (and potentially many other branches of forensic science); for a different approach applicable to certain precision problems in DNA-profile comparison, see Curran (2005).

Examples of measurement of the precision of forensic-voice-comparison systems are provided in **[99.770]**.

Using measures of accuracy and precision

[99.320] Measures of accuracy and precision are useful in several ways:

If a forensic scientist does not consider the system they are using to be sufficiently accurate and precise, then they can conduct additional research modifying some or all components of the system so as to achieve higher degrees of accuracy and precision. While developing systems, accuracy and precision measures can be used to determine which system is most effective and which is therefore the best candidate for use in casework.

If a forensic scientist does consider their system to be sufficiently accurate and precise (and they are working in a jurisdiction with *Daubert*, or *Daubert*-like, admissibility standards), then they can submit the accuracy and precision test results to the court so that the judge can consider whether the system is sufficiently valid and reliable that evidence evaluated using the system should be admitted (the accuracy and precision results could be presented at a formal admissibility hearing if the judge deems one necessary). At this point, the C_{llr} measure of accuracy and the credible intervals over the whole test set are probably particularly pertinent, because at an admissibility hearing the issue is the

general validity and reliability of the system. More detailed information about the system could be presented using Tippett plots including credible intervals (see **[99.330]** and **[99.770]**).

When giving evidence at trial, the forensic scientist can also present the trier of fact with accuracy and precision results so that the trier of fact can assess the likelihood-ratio strength-of-evidence statement accordingly. At this point the C_{llr} measure of general system accuracy is probably less pertinent, and may be too technical to be presented to a jury. Likewise, Tippett plots may be too technical. Reporting the error rate and the credible interval for the particular likelihood-ratio value obtained would be more appropriate. For example, the forensic scientist could give a statement of the following form:

> Based on my evaluation of the evidence, I have calculated that one would be X times more likely to obtain the acoustic differences between the voice samples if the questioned-voice sample had been produced by someone other than the accused than if it had been produced by the accused. Based on my calculations, I am 95% certain that it is at least X_{lower} times more likely and not more than X_{upper} times more likely. In tests of my forensic-voice-comparison system, Y% of same-speaker comparison results provided support for the different-speaker proposition which was greater than or equal to that found for the comparison of the known- and questioned-voice samples.

There would probably need to be substantial explanation leading up to such a concise summary statement of the strength-of-evidence and its validity and reliability, especially if dealing with a lay jury.

Tippett plots

[99.330] A graphical method for presenting the results of running a likelihood-ratio forensic-comparison system on a set of test data is a *Tippett plot*. Tippett plots were introduced in Meuwly (2001) (inspired by the work of C. F. Tippett and Evett & Buckleton, 1996), and are now a standard method for presenting results in likelihood-ratio forensic-voice-comparison research. Tippett plots provide more detailed information about the results than is available from a summary measure such as C_{llr}. This section provides a guide to the interpretation of Tippett plots.

Figures 11 through 13 provide a series of Tippett plots drawn on the basis of hypothetical sets of output from forensic-comparison systems. The lines rising to the right represent the results from same-speaker comparisons in the test set, the cumulative proportion of log likelihood ratios less than or equal to the value indicated on the x axis. The lines rising to the left represent the results from different-speaker comparisons in the test set, the cumulative proportion of log likelihood ratios greater than or equal to the value indicated on the x axis. (Some authors draw both same-speaker and different-speaker lines as the cumulative proportion of log likelihood ratios greater than or equal to the value indicated on the x axis.) In these hypothetical results the same-speaker and different-speaker lines are symmetrical and cross at a log likelihood ratio of zero; this need not be the case for real test results.

FIGURE 11. Tippett plot of hypothetical test results.

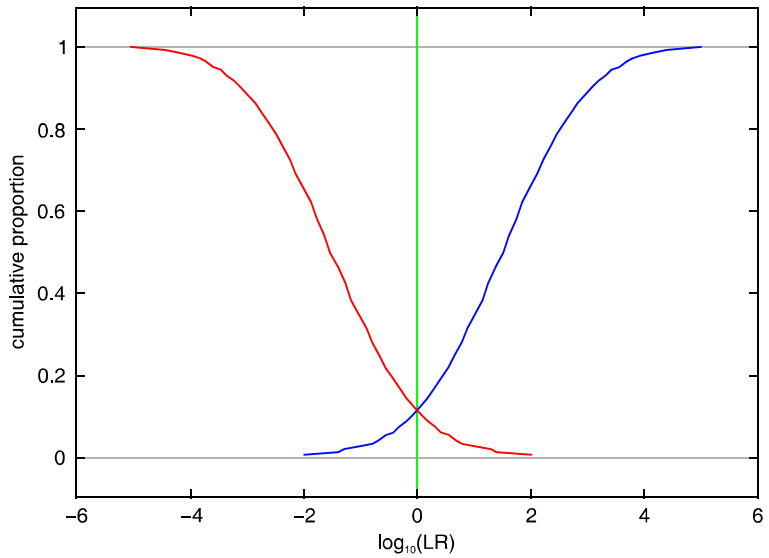


FIGURE 12. Tippett plot of hypothetical test results.

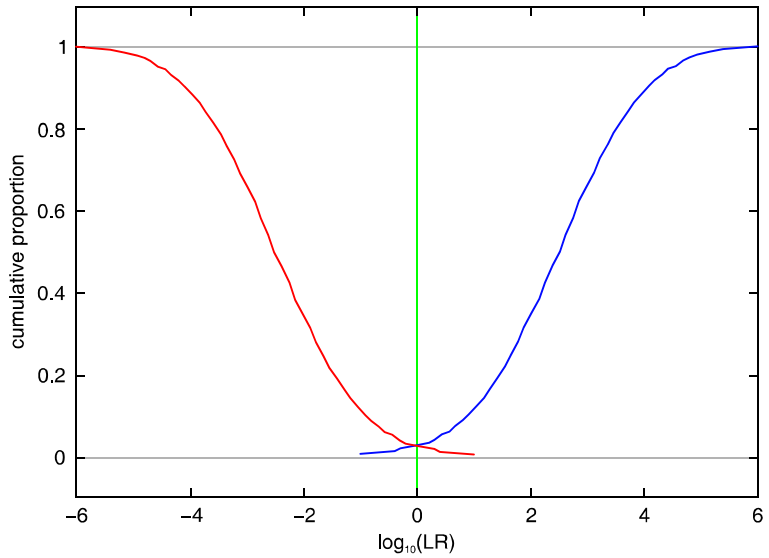
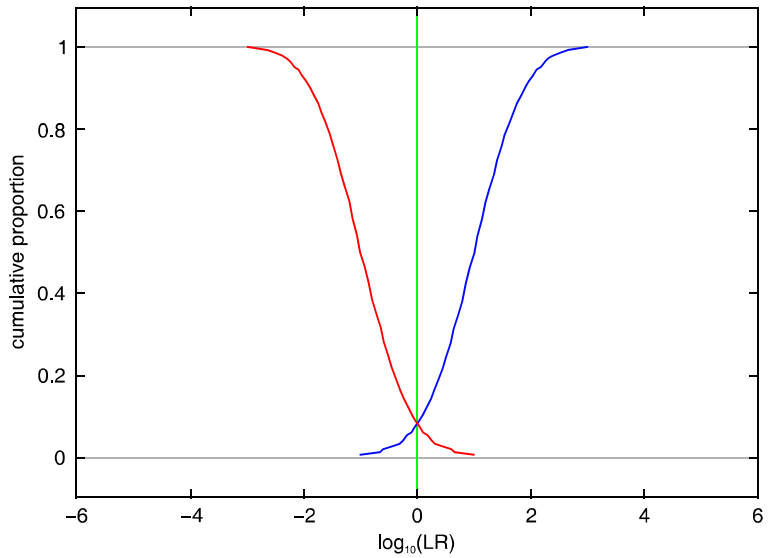


FIGURE 13. Tippett plot of hypothetical test results.



As discussed in section **[99.300]**, an ideal forensic-comparison system should produce a large positive log likelihood ratio for a same-origin comparison, and a large negative log likelihood ratio for a different-origin comparison. Large-magnitude log likelihood ratios which support the consistent-with-fact hypothesis are better than small-magnitude log likelihood ratios which support the consistent-with-fact hypothesis. Log likelihood ratios which support the contrary-to-fact hypothesis are bad, and the larger their magnitude the worse they are. Therefore, in Tippet plots the further apart the same-speaker and different-speaker lines (the further to the right the same-speaker line and the further to the left the different-speaker line) the better the results. The results presented in the Tippett plot in Figure 12 are therefore better than those presented in the Tippett plot in Figure 11.

Note, however, that (consistent with the C_{llr} metric) log-likelihood-ratio results which support contrary-to-fact hypotheses are of greater concern than whether the consistent-with-fact log-likelihood-ratio results are relatively small or large – a system which minimises support for contrary-to-fact hypotheses is preferable even if this leads to a reduction in its strength of support for consistent-with-fact hypotheses. The results presented in the Tippett plot in Figure 13 are therefore also better than those presented in the Tippett plot in Figure 11.

Tippet plots based on real test results are shown in examples of forensic voice comparison provided in **[99.770] ff**. These include 95% credible intervals.

PROBLEMS AND OPPOSITION

Misinterpretations of forensic likelihood ratios

[99.370] The main problems with the presentation of likelihood-ratio strength-of-evidence statements in court relate to misinterpretations of those statements. Misinterpretations can occur in the mind of a lawyer, judge, or jury member, or can be inadvertently caused by a forensic expert misphrasing their strength-of-evidence statement. Forensic scientists should be careful not to inadvertently cause a misinterpretation via misphrasing, and as far as is possible try to prevent others from misinterpreting correctly phrased statements. Lawyers and judges should also be careful not to induce misinterpretations in the minds of jury members. The two principle misinterpretations are commonly referred to as *the prosecutor's fallacy* and *the defence attorney's fallacy*. These are summarised here and are also discussed in many publications including Aitken & Taroni (2004, §3.3), Balding (2005, ch. 9), Buckleton (2005), and Robertson & Vignaux (1995, ch. 6).

The prosecutor's fallacy

[99.380]

Forensic Scientist: One would be one thousand times more likely to obtain the acoustic differences between the voice samples if the questioned-voice sample had been produced by the accused than if it had been produced by some other speaker. What this means is that whatever one believed prior to being presented with this evidence, one should now be one thousand times more likely than before to believe that the questioned-voice sample was produced by the accused than that is was produced by some other speaker.

Prosecutor: So, to simplify for the benefit of the jury if I may, what you are saying Doctor is that it is a thousand times more likely that the voice on the telephone intercept is that of the accused than that it is of any other speaker.

The prosecutor's fallacy is also know as the *transposition of conditionals*. It consists of ignoring the prior odds and interpreting the likelihood ratio (probability of evidence given hypotheses) as the posterior odds (probabilities of hypotheses given evidence), see **[99.160]**.

To understand why this is such a serious error let us return to the cow example from **[99.210]**. Imagine that I tell you I have a cow somewhere out of sight and I ask you "what is the probability that it has four legs given that it is a cow?" $p(E = 4 \text{ legs} \mid H_{cow})$. In the imaginary data which I reported in **[99.210]** I said that 98% of cows had four legs, which corresponds to a probability of 0.98. In reality the number of cows that do not have four legs is probably a lot smaller than my imaginary data suggest and I would expect the probability of an animal having four legs given that it is a cow to be much closer to 1.

Now let me ask the transposed-conditional question: "What is the probability that an animal is a cow given that it has four legs?" $p(H_{cow} \mid E = 4 \text{ legs})$. It should be immediately obvious that the answer to this question is certainly not a probability of very close to 1 – lots of animals including sheep, pigs, horses, dogs, cats, giraffes, and elephants typically have four legs, and the proportion of four-

legged animals in the world which are cows is probably quite small, maybe less than 0.01 (1%), i.e., close to 0, not close to 1.

The prosecutor's fallacy is to take the statement that "the probability of the animal having four legs given that it is a cow is very high" (or "the probability of obtaining the acoustic differences between the speech samples is many times more likely if the questioned-voice sample had been produced by the known-speaker than if it had been produced by some other speaker"), and misphrase or misinterpret it as "given that the animal has four legs, the probability of it being a cow is very high" (or "given the acoustic differences between the speech samples, it is many times more likely that the questioned-voice sample was produced by the known-speaker than that it was produced by some other speaker").

Of course what is missing was the prior probability of an animal being a cow. If, for sake of argument, we accepted the probabilities of evidence calculated in **[99.210]**, and, again for sake of argument, used the prior probability of being a cow of 0.01, then the posterior probability for being a cow would be as in Formula 4 (which is an alternate formulation of Bayes' Theorem used to calculate a posterior probability rather than posterior odds – Formula 4 can be mathematically derived from Formula 2 and vice versa).

Formula 4

$$p(H_{cow} \mid E = 4 \text{ legs}) = \frac{p(E = 4 \text{ legs} \mid H_{cow}) \times p(H_{cow})}{p(E = 4 \text{ legs} \mid H_{cow}) \times p(H_{cow}) + p(E = 4 \text{ legs} \mid H_{not\ cow}) \times p(H_{not\ cow})}$$

$$0.0198 = \frac{0.98 \times 0.01}{0.98 \times 0.01 + 0.49 \times 0.99}$$

Remembering that probability values range between 0 and 1, the probability of an animal being a cow given that it has four legs, 0.0198, is very different from the probability of an animal having four legs given that it is a cow, 0.98.

Avoiding the prosecutor's fallacy: Although the term "prosecutor's fallacy" may suggest that only a prosecutor would transpose the conditionals, it is in fact a mistake which can very easily be unintentionally made by prosecutors, defence council, judges, jury members, and even forensic scientists (see, for example, Thompson & Schumann, 1987). A way to help avoid making this mistake is to always ask whether the probability/likelihood statement refers to the hypothesis/hypotheses (same speaker or different speaker). If so, and it is an interpretation of the evidence presented by a forensic scientist, then it is probably example of the prosecutor's fallacy. To hopefully lessen the probability of the trier of fact making the transposition-of-conditionals mistake, during examination in chief it may be advisable for the forensic scientist to be asked questions which allow them to explain the potential misinterpretation problem so that the trier of fact can be warned to avoid it. The forensic scientist should explicitly state that they are not providing the probability that the known- and questioned-voice samples were produced by the same speaker. See Buckleton (2005) for additional advice on how to avoid the prosecutors fallacy.

The defence attorney's fallacy

[99.390]

Forensic Scientist: One would be one thousand times more likely to obtain the acoustic differences between the voice samples if the questioned-voice sample had been produced by the accused than if it had been produced by some other adult male Australian-English speaker. What this means is that whatever one believed prior to being presented with this evidence, one should now be one thousand times more likely than before to believe the that questioned-voice sample was produced by the accused than that it was produced by some other adult male Australian-English speaker.

Defence attorney: So, given that there are approximately a million adult male Australian-English speakers in the region and assuming initially that any one of them could have made the intercepted telephone call, we begin with prior odds of one over one million, and the evidence which has been presented has resulted in posterior odds of one over one hundred thousand. One over one hundred thousand is a very small number. Since it is one hundred thousand times more likely that the voice on the telephone intercept was that of an adult male Australian-English speaker other than my client than that it is the voice of my client, I submit that this evidence fails to prove that my client was the speaker on the intercepted telephone call and as such it should not be taken into consideration by the jury.

The logic of the defence attorney's fallacy is correct until the final conclusion. What the defence attorney's fallacy does is ignore all other evidence presented at trial so as to imply that a particular piece of evidence is of no value. Different types of evidence can reasonably be assumed to be statistically independent, hence the prior odds and the likelihood ratios from different evidence can simply be multiplied together to arrive at the posterior odds (*naïve Bayes* combination). Let us assume prior odds of 1/1 000 000, and likelihood ratios of 1000 from forensic voice comparison evidence, 100 from fingermark evidence, and 10 000 from DNA evidence. The posterior odds would then be: 1/1 000 000 × 1000 × 100 × 10 000 = 1000 (the multiplication could be done in any order).When the forensic-voice-comparison evidence is included the conclusion is that it is one thousand times more likely that the crime was committed by the accused than by some other person. Had the forensic-voice-comparison evidence not been included, then the posterior odds would have been 1/1 000 000 × 100 × 10 000 = 1, equal probability that the crime was committed by the accused as by some other person. In this example the forensic-voice-comparison evidence is certainly of value.

If only one piece of evidence were presented and the resulting posterior odds were 1/10 000 then this should not lead to a conviction. In fact, even if the posterior odds were 10 000 or higher, a single piece of evidence should still not lead to a conviction – likelihood ratios are probabilistic not definitive and additional evidence pointing in the opposite direction could override even a very large likelihood ratio.

Opposition to the adoption of the new paradigm

[99.400] In forensic voice comparison, as in other branches of forensic science, there has been opposition to the adoption of the new paradigm, including the adoption of the likelihood-ratio framework. See Buckleton (2005), Evett (1991), and Meuwly (2006) on other branches of forensic science and on forensic science in general; and Morrison (2009b), Rose (2002, pp. 66–78), and Rose & Morrison (2009) on forensic voice comparison in particular.

One reason for opposition to the adoption of the likelihood-ratio framework component of the new paradigm seems to be a lack of understanding of the likelihood-ratio framework, possibly combined with a belief that the likelihood-ratio framework is difficult to understand. I do not believe that the likelihood-ratio framework is inherently difficult to understand – the reader may judge for themself depending on the difficulty they had in understanding sections **[99.150]** and **[99.160]** above. As in many other fields, difficulties in understanding may stem from preconceived ideas which interfere with a person's ability to understand new information when it is presented to them. Even if the likelihood-ratio framework were relatively difficult to understand, this would not constitute a rational argument against its adoption. Is it better to use the logically correct framework for the evaluation of evidence, or to use a logically incorrect framework which may be easier to understand?

The other principal reason for opposition to the adoption of the new paradigm appears to be related to the difficulty and expense involved in the collection and analysis of databases of recordings of voices of speakers from the relevant population. Database collection and analysis can indeed be time consuming and financially costly; however, it is an absolute necessity in order to be able to calculate the typicality of voice samples and also to be able to empirically assess the accuracy and precision of forensic-voice-comparison systems.

In what appears to me to have been at least in part an attempt to resist pressure to adopt the new paradigm, in 2007 a group of forensic speech scientists in the United Kingdom published a position statement in which they described a framework for the forensic comparison of voice samples (French & Harrison, 2007). Superficially the UK framework may appear to be consistent with the likelihood-ratio framework, and it claims to be conceptually identical to the framework used for DNA-profile comparison, but in fact it is not. Amongst its flaws, in two instances it advocates giving posterior probabilities, and it suffers from the cliff-edge effect (see **[99.190]** above). Rather than background databases and objective measurements, the authors of the UK position statement appear to have had an experienced-based auditory-acoustic-phonetic approach (see **[99.660]** and **[99.670]** below) in mind:

> it involves 'separating out' the samples into their constituent phonetic and acoustic 'strands' (e.g., voice quality, intonation, rhythm, tempo, articulation rate, consonant and vowel realisations) and analysing each one separately. . . . by drawing upon research literature and *general experience*, the analyst may provide an assessment of the degree to which the features common to the questioned voice and that of the suspect are unusual or distinctive. (French & Harrison, 2007, p. 138, emphasis added)

This is confirmed by French *et al*. (2010):

> The framework set out in that document serves to remind forensic phoneticians of the need to judge the distinctiveness of the features found in the criminal and suspect samples, and this implies comparison with a broader population, albeit informally via the analyst's experience

> and general linguistic knowledge rather than formally and quantitatively. (French *et al*., 2010, p. 144)
>
> we feel that the 2007 framework provides a sound practical framework for the expression of conclusions arrived at through auditory-acoustic phonetic comparison. (French *et al*., 2010, p. 150)

French *et al*. (2010) defend the inclusion of a posterior-probability-based definitive exclusion option as part of the UK framework. The also interpret Rose & Morrison (2009) as allowing such an option – this interpretation is incorrect, neither definitive nor gradient posterior probabilities should never be presented by a forensic scientist.

The UK position statement also fails to make any mention of validity and reliability. See Rose & Morrison (2009) and Morrison (2009b) for more detailed critiques of the UK position statement.

At the time of writing (May 2010) it is probably still the case that the majority of persons who would present themselves as experts in forensic voice comparison have not adopted the new paradigm. In contrast to the AFSP, the *International Association for Forensic Phonetics and Acoustics* (IAFPA) http://www.iafpa.net/, the only international association whose focus is forensic voice comparison, does not require its members to use the likelihood-ratio framework, nor does it require them to demonstrate the validity and reliability of the methods by which they arrive at the evidentiary statements which they present in court.

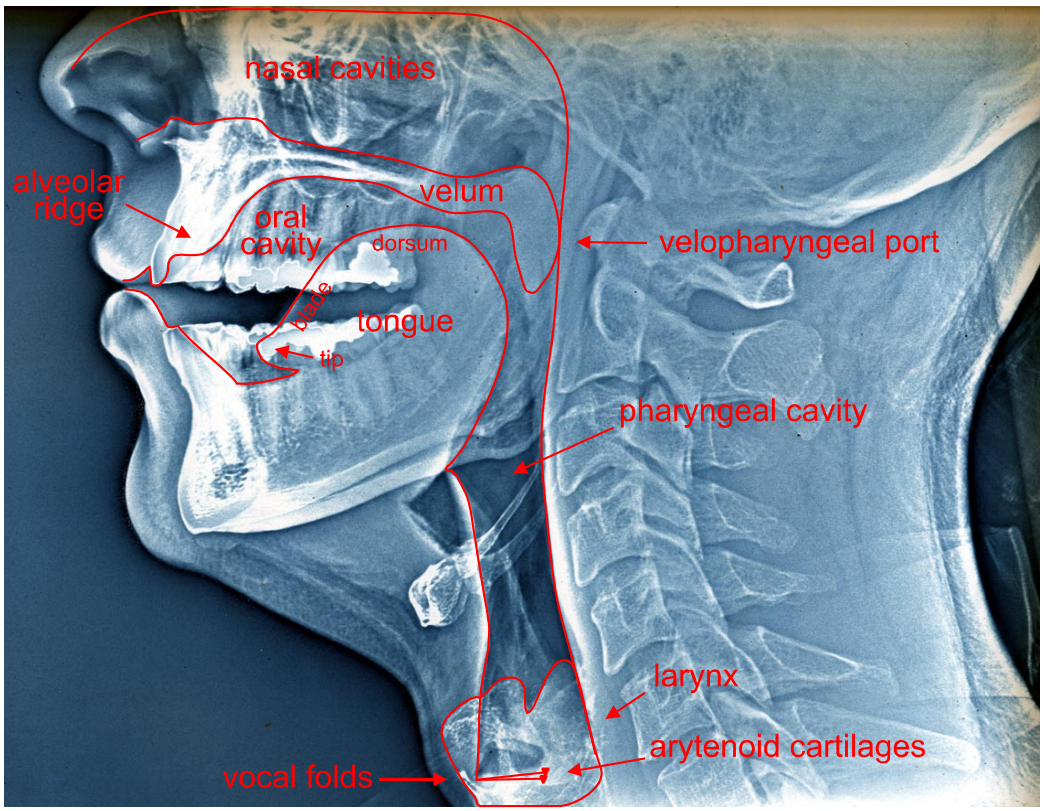
HUMAN VOICES

Introduction

[99.440] *Phonetics* is the study of the physical aspects of the production, transmission, and perception of human speech. This section provides a brief introduction to articulatory and acoustic phonetics, which cover the production and transmission of speech. The intent is to provide the reader with a basic understanding of some of the phonetic terms and concepts which may be used in reports on forensic voice comparison. It is the acoustic speech signal which is measured and analysed in forensic voice comparison.

This introduction is not meant to be exhaustive. A reader wishing to gain a wider and deeper understanding of phonetics may wish to start by consulting some of the works listed under **Further reading [99.560]**.

FIGURE 14. X-ray and tracing of a vocal tract.



Vocal tract

[99.450] Humans make speech sounds using their *vocal tracts*. The vocal tract is essentially a tube consisting of the mouth (*oral cavity*) and throat (*pharyngeal cavity*), with the *lips* at one end and the *larynx* at the other (the *vocal folds* are in the larynx), see Figure 14 (this is an X-ray of Philip Rose with the vocal tract highlighted). The length of the tube can be slightly increased by rounding and protruding the lips and by lowering the larynx (raising the larynx will slightly shorten the tube). The nose forms another tube (*nasal cavities* from the *nostrils* to the *velopharyngeal port*) which can be connected to the *oropharyngeal* tube (pharyngeal cavity plus oral cavity) by lowering the soft palate (*velum*) to open the velopharyngeal port, see Figure 14. The jaw can be lowered or raised and the tongue can be moved to change the shape of the oropharyngeal tube.

Vowels

Description

[99.460] The vocal tract is similar to a musical instrument, a wind instrument such as a clarinet or a trombone. To play these instruments, one must blow air into them. Air is blown into the vocal tract by compressing the *lungs* so as to push air between the vocal folds. However, simply blowing into a trombone will not make a musical sound, one has to "blow a raspberry" forcing air between ones lips so that they vibrate, opening and closing many times per second. Similarly, a reed needs to be fitted to the mouthpiece of a clarinet so that when one blows into the mouthpiece the reed vibrates. In the same way, to make a *voiced* sound (including a vowel), one has to hold one's vocal folds together and under tension so that when air is forced between them they vibrate, opening and closing many times per second (see **[99.540]** for details). Note that, as one can open one's lips and not "blow a raspberry", one can open one's vocal folds and not make a voiced sound; in fact, the latter is the normal state when one is breathing.

To verify that the difference between a voiced and a *voiceless* sound is the vibration of the vocal folds, put your fingers on your throat, in front of your larynx, and say "sssssss" (this is a voiceless sound), then say "zzzzzzzzz" (this is a voiced sound) – you should be able to feel the vibration with your fingers. Now try saying "bus" and "buzz" – in both cases the vowel is voiced but the following consonant in "bus" isn't and the vibrations should stop sooner in "bus" than in "buzz".

To get different notes out of a trombone you have to move the slider, which changes the length of the tube. To get different notes from a clarinet you have to open and close holes – the length of the tube is the distance from the mouthpiece to the nearest open hole. When the tube is longer the note sounds lower, and when the tube is shorter the note sounds higher – also think about long fat tubes and short thin tubes in a pipe organ. The difference in the notes of a wind instrument are usually not caused by differences in the rate of vibration at the mouthpiece, rather they are caused by differences in the length of the tube which cause the tube to have different *resonance frequencies* (frequencies at which the sound is amplified) – longer tubes have lower resonance frequencies and shorter tubes have higher resonance frequencies.

The resonance frequencies of a simple tube can be easily calculated mathematically, one only needs to know the length of the tube and the speed of sound (the cross-sectional area of the tube also has a small effect). Tubes have multiple resonances, not just one, and a simple tube 16 cm in length

(about the average length of adult-human vocal tracts) will have resonances at about 500 Hz, 1500 Hz, 2500 Hz, etc. The amplitude (loudness) of the resonances gets less as the frequency gets higher.

The resonance frequencies of vocal tracts are usually called *formants*. Although a human can increase the length of their vocal tract by rounding and protruding the lips and by lowering their larynx, this increase in length is limited and the primary way in which a human changes the resonance frequencies of their vocal tract is by lowering or raising their jaw and moving their tongue. Part of the tongue is moved towards part of the roof of the mouth or the back of the throat causing a *constriction* in the oropharyngeal tube. The vocal tract is then a complex-shaped tube rather than a simple tube. In a simple description of the complex tube, the location, length, and cross-sectional area of the constriction, and the concomitant lengths and cross-sectional areas of the parts of the tube behind and in-front of the constriction, determine the resonance frequencies of the vocal tract.

Try saying the vowel sound "ee" from the word "heed", keep saying it, don't stop. Your jaw is probably quite high and the front-to-middle part of your tongue is probably quite close to the roof of your mouth. Now slowly open your mouth and lower your tongue – you should hear the "ee" sound change to sound like the vowel sounds in "hid" then "head", then "had" (how well the sounds correspond with the words may depend on your accent, for example, in some Scottish, New Zealand, and Southern US accents "head" may sound like "heed" or "hid" said by speakers of other accents of English). The different mouth shapes result in different resonance frequencies which make the sound of different vowels. The primary acoustic differences between the vowels in "heed", "hid", "head", and "had" are that the first formant (F1) increases as the constriction widens and second formant (F2) decreases.

Now say the "ee" sound from "heed" again, but this time move your tongue back until you are saying the vowel sound from "who" – you have probably also gone from *spread lips* (like smiling) to *rounded lips* (in Figure 14 the speaker is saying the vowel sound of "who"). It turns out that moving your tongue back in your mouth lowers F2 and that rounding your lips also lowers F2, so doing both together has a larger effect. The most important acoustic difference between the vowel sounds in "heed" and "who" is the change in F2 (F1 stays about the same).

The *phonetic symbols* of the International Phonetic Association (IPA) http://www.langsci.ucl.ac.uk/ipa/ can be used to represent many speech sounds, and *diacritics* (extra smaller symbols put above, below, or after a main symbol) can be used to represent small differences between speech sounds; the IPA alphabet is reproduced in **Appendix A [99.1150]**. The symbols for the vowel sounds in "heed", "hid", "head", "had", and "who" are /i/, /ɪ/, /ɛ/, /æ/, and /u/ respectively. Slashes /_ / are put around phonetic symbols in *broad transcription*, indicating the sounds which contrast in a given language or dialect (*phonemes*), and square brackets [_] are used to indicate finer phonetic detail in a *narrow transcription*, e.g., "buzz" /bʌz/ may be realised as [bʌːs], the vowel is long ([ː] is the diacritic for long duration) and the final consonant is voiceless.

FIGURE 15. Spectra of vowels /i/, /ɛ/, and /u/.

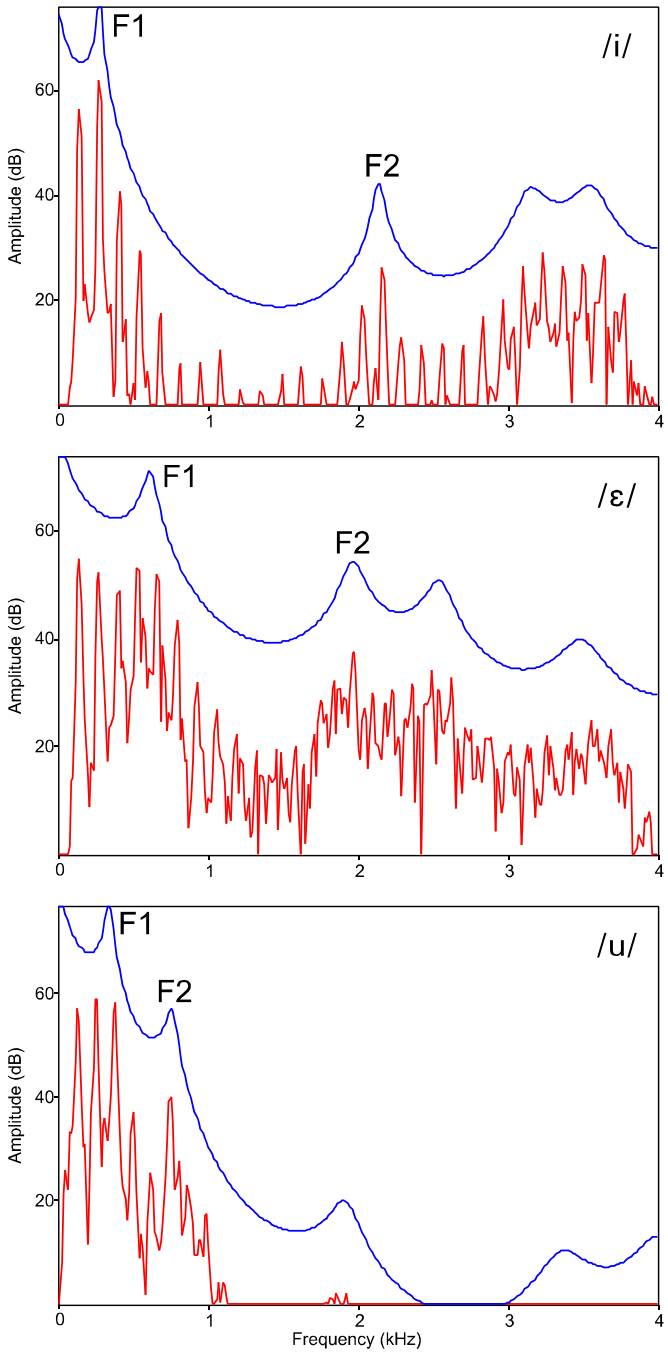
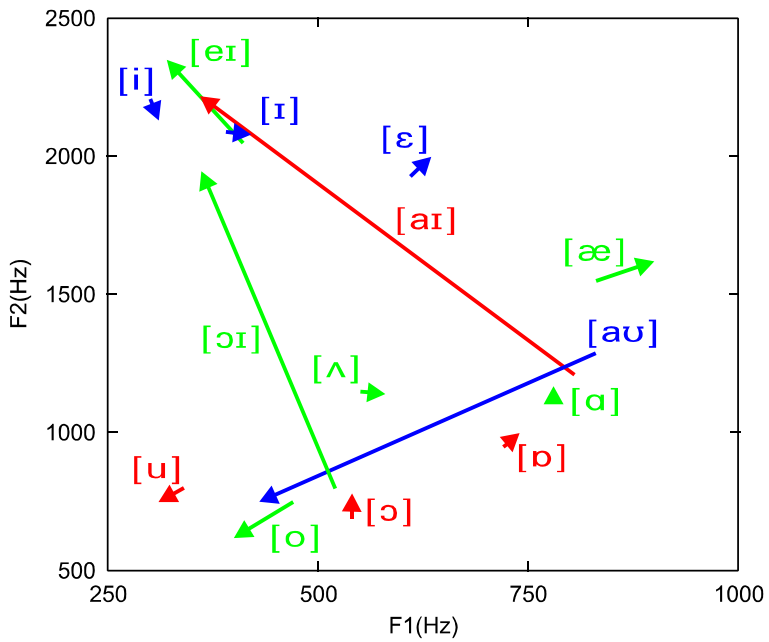


Figure 15 shows the *spectra* (singular: *spectrum*) of the vowels /i/, /ɛ/, and /u/ (spoken by the author). Frequency is on the *x* axis and amplitude on the *y* axis. These spectra were measured at a point in time 25% of the way between the beginning and the end of the vowel. The jagged red lines are raw measurements and the blue lines are smoothed measurements. The peaks in the smooth lines are the formant measurements, the first two peaks from the left are F1 and F2. Note that for /ɛ/ F1 is higher and F2 lower than for /i/, and for /u/ F1 is about the same but F2 is much lower than for /i/. Note that there are also other differences in the shape of the spectra.

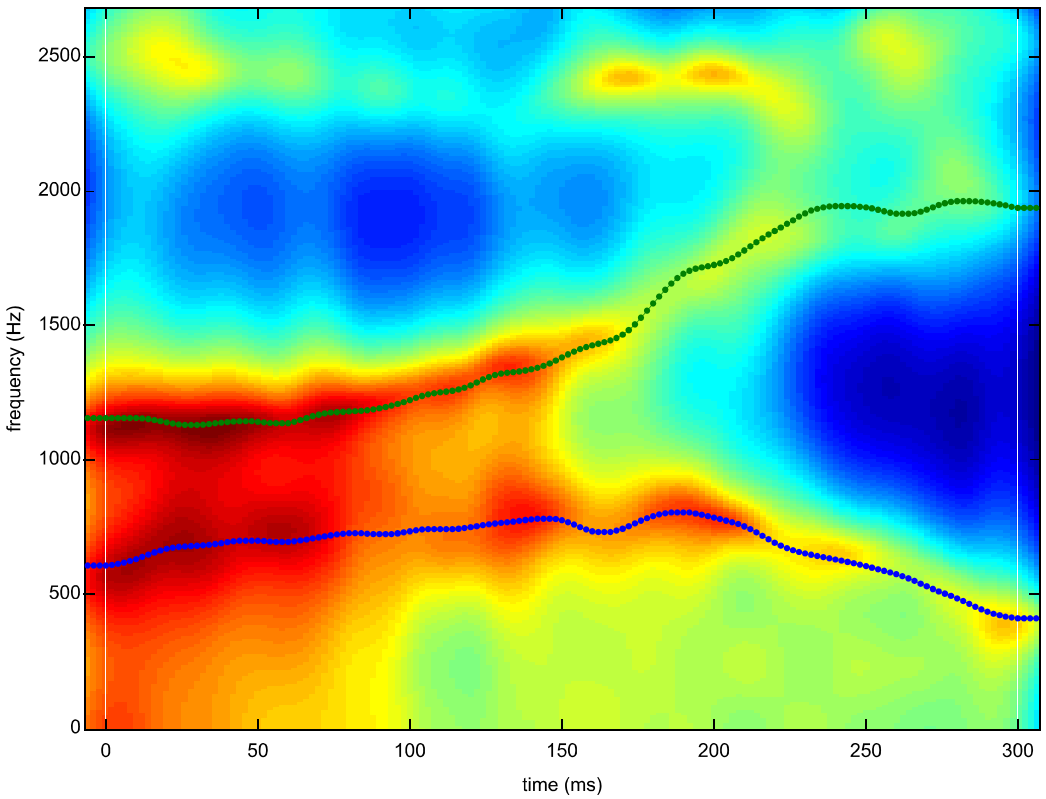
In many languages F1 and F2 peaks are the primary acoustic indicators of vowel category (vowel phoneme) identity (the peak formant values rather than the exact shape of the spectra are perceptually relevant), and vowels are often graphically represented via a two-dimensional plot of F1 and F2 as in Figure 16 (vowels spoken by the author). This plot has arrows pointing from measurements taken at 25% of the duration of the vowel to measurements taken at 75% of the duration of the vowel. Some vowels have very little formant movement, and others have substantial formant movement, the former are known as *monophthongs*, and the latter as *diphthongs*, for example, the vowel /aɪ/ as in the word "hide"starts off with high F1 and intermediate F2, somewhere between /æ/ and /ɑ/ (the first vowel in "father"), and ends up with a low F1 and a high F2, somewhere between /ɪ/ and /i/ (in a broad Australian-English accent this might be realised as [ɔɪ] rather than [aɪ], and in a Canadian-English accent "hide" may be realised as [haːd̥], [̥] is the voiceless diacritic, but "height" as [hʌɪt]).

FIGURE 16. Plot of F1 and F2 measurements of a set of English vowels.



Another graphical method for representing the acoustic speech signal is a *spectrogram*. A spectrogram is made by measuring the spectrum of the speech signal every few milliseconds, then lining those spectra up in order so that time is on the *x* axis and frequency is on the *y* axis. On a three-dimensional plot amplitude can be represented on the *z* axis (this is called a *waterfall plot*), but it is more common to produce a two-dimensional plot with darkness of a monochrome scale or colours on a multi-coloured scale used to represent amplitude. Spectrograms can represent fine details of the acoustic signal across time, frequency, and amplitude. Figure 17 provides an example of a colour spectrogram of a token of the diphthong /aɪ/ spoken by an adult male speaker of Australian English (the highest amplitudes are in dark red, and the lowest in dark blue, measurements of the first two formant peak frequencies have been overlayed).

FIGURE 17. Spectrogram of /aɪ/.



In addition to F1, F2, and diphthongisation, vowel duration can be an important cue to vowel phoneme identity in English (for example, in addition to spectral differences, all else being equal, /i/ is longer than /ɪ/ in most dialects of English). In some languages, such as French, other acoustic properties such as third formant (F3) and nasalisation (see **[99.480]**) can be important for vowel phoneme identity.

In English, vowels can be *stressed*, in which case they are relatively long and have well defined formant values, or they can be *non-stressed* in which case they are relatively short and the vocal tract approximates a rest position or the position needed to make the preceding or following speech sounds, which results in some degree of neutralisation of the vowel's formant values. The ultimate non-stressed vowel is schwa [ə] for which the original identity of the vowel phoneme is lost, for example, the second vowel in "photograph" is realised as a schwa as are the first and third vowels in "photographer".

Forensic value

[99.470] The acoustic properties of speech will be useful forensically to the extent that they have relatively large between-speaker variation and relatively small within-speaker variation.

From the discussion above, it should be clear that vocal-tract length has a major effect on formant frequencies; men generally have longer vocal tracts than women but there is also variation within each sex. Additional anatomical differences in the shape of the vocal tract and idiosyncrasies in control of the muscles of the tongue, lips, etc. may also be reflected in vowel spectra. Speakers may also exhibit *idiolectal* differences which are more subtle versions of the sort of dialectal differences mentioned above. Acoustic properties which are not important for vowel phoneme identity, such as the higher formants (F3 and above) and the shape of the whole spectrum, may also contain information which can help differentiate speakers.

Although there may be a great deal of anatomical and idiosyncratic variation between speakers, the ability of a forensic-voice-comparison system to exploit this may be limited. Much of the information may not be available or may not be extractable from the acoustic signal. Transmission of the acoustic signal through a telephone system will alter the shape of the spectrum and, depending on the vowel phoneme, may make both F1 and higher formants unusable, see **[99.610]**. Also, unlike DNA profiles or fingermarks, within-speaker variability of many of the acoustic properties of speech may be very high and may approach the level of between-speaker variability.

Research papers on forensic voice comparison in the likelihood-ratio framework using measurements of vowel formants include Becker, Jessen, & Grigoras (2008, 2009), González-Rodríguez *et al*. (2007), and Morrison (2009c). Research papers on forensic voice comparison in the likelihood-ratio framework using whole spectra (of the whole acoustic speech signal, not just vowels) include Alexander *et al*. (2005), González-Rodríguez *et al*. (2006), González-Rodríguez *et al*. (2007), and Thiruvaran, Ambikairajah, & Epps (2008).

Nasals

Description

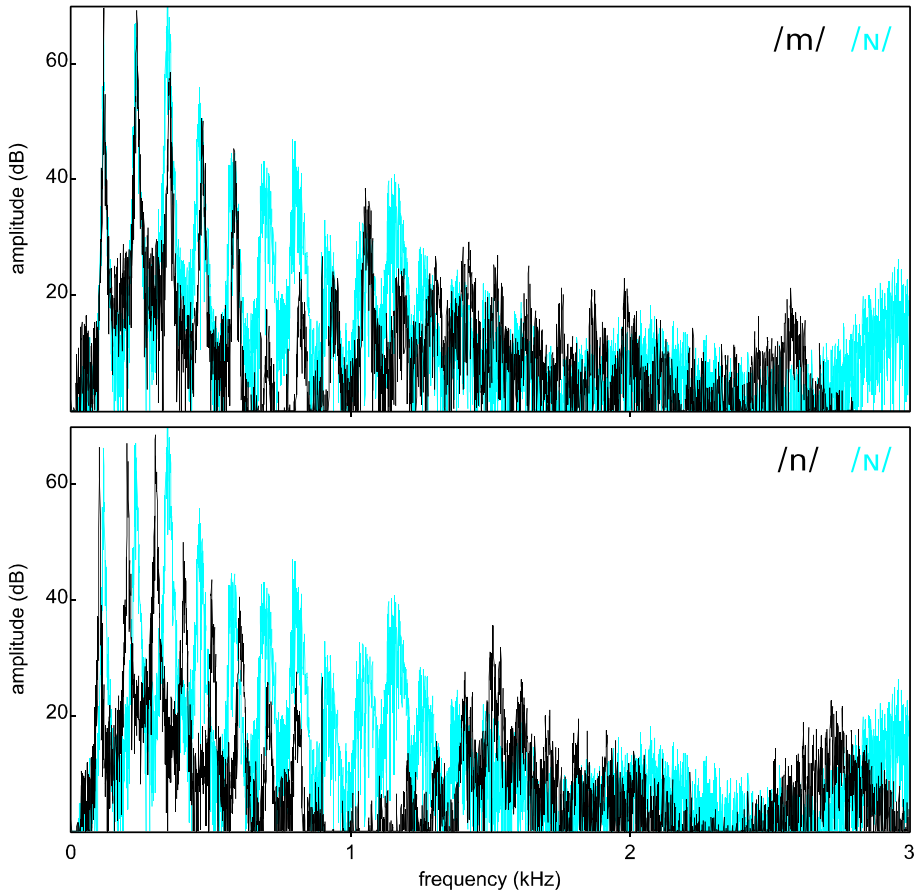
[99.480] *Nasals*, such as /m/, /n/, and /ŋ/ (the last sound in "sum", "sun", and "sung" respectively) are produced by producing voicing, opening the velopharyngeal port so that air can flow through the nasal cavities (see Figure 14), and making a closure in the oral cavity (the velopharyngeal port is also held open when one is breathing through one's nose). The tracing of the nasal cavities in Figure 14 is greatly simplified, and in reality the shape of nasal cavities is very complex, including several side-branches (sinuses).

For /m/ the lips are held together and the oral cavity is a relatively long side-tube on the *nasopharyngeal* tube (nasal cavities plus pharyngeal cavity). For /n/ the *tip* and *blade* of the tongue (see Figure 14) are held against the *alveolar ridge* to make a closure and the oral cavity tube is shorter than for /m/ (if you put the tip of your tongue on your upper lip, then gradually move it backwards past your upper incisors and gums and keep going, you get to a ridge near the front of the roof of your mouth, this is the alveolar ridge, see Figure 14). For /ŋ/ the closure is made between the *dorsum* of the tongue and the velum (see Figure 14), and the oral cavity tube is very short.

The acoustic differences in the spectra of nasals, which makes them sound different, is due to the different *anti-resonances* of the different lengths of the oral-cavity tube. Rather than adding a resonance, a closed side-tube to a main-tube subtracts an anti-resonance. Figure 18 shows the raw spectra of /m/ and /n/ compared to /ɴ/; /ɴ/ is a nasal where the closure is a little further back than for English /ŋ/ such that the length of the oral cavity side-tube is zero – Figure 18 therefore compares the spectrum of the nasopharyngeal tube with the spectra of the nasopharyngeal tube plus the different-length oral-cavity side-tubes. The first anti-resonance for /m/ can be seen as the lower amplitude of the /m/ spectrum compared to the /ɴ/ spectrum at around 750 Hz, for /n/ the anti-resonance is more pronounced and occurs at and just above 1 kHz. For both /m/ and /n/, the spectra above the first anti-resonance are also shifted down in frequency relative to the /ɴ/ spectrum.

Both the nasal and oral cavities can be open at the same time, i.e., the velopharyngeal port is open and there is no closure in the oral cavity. When there is also voicing from the vocal folds, this results in a *nasalised vowel*. The spectra of nasalised vowels can be quite complex because both the oral and nasal cavities contribute resonances and anti-resonances. In some languages, such as French, nasalised and non-nasalised (oral) vowels serve as different phonemes, a word containing the nasalised version of the vowel has one meaning and an otherwise phonemically identical word with the oral version of the vowel has another meaning (e.g., "matais" /matɛ/ form of verb TO SUBDUE and "matin" /matɛ̃/ MORNING, [˜] is the nasalisation diacritic). In English, nasalisation does not distinguish phonemes but vowels preceding nasals are often nasalised – the velum is lowered during the vowel in preparation for saying the nasal consonant (when articulations for making one speech sound overlap with the articulation of earlier or later speech sounds, this is known as *coarticulation*).

FIGURE 18. Spectra of nasals /m/ and /n/ compared to /ɴ/.



Forensic value

[99.490] Nasal cavities are very complex with potentially large between-speaker variability leading to the potential for large between-speaker variability in their acoustic spectra. Nasal cavities are static structures and hence have essentially no within-speaker variability, but the degree of opening of the velopharyngeal port can be varied and this will affect the acoustic spectra. Nasal congestion due to colds or allergies will also affect the acoustic spectra. Mobile-telephone systems do not explicitly encode anti-resonances and parts of the acoustic spectra of nasals may be lost, see **[99.610]**.

A research paper on forensic voice comparison in the likelihood-ratio framework which uses measurements of nasal spectra is Rose, Osanai, & Kinoshita (2003).

Fricatives

Description

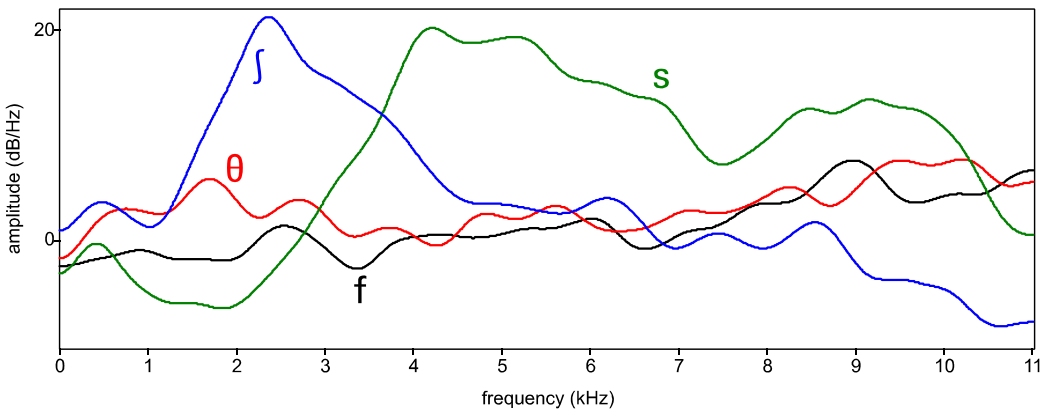
[99.500] Most dialects of English have five voiceless *fricatives*, the first sound in each of "fish" /f/, "thick" /θ/, "sip" /s/, "ship" /ʃ/, and "hip" /h/, and four voiced fricatives, the first sound in each of "villa" /v/, "the" /ð/, "zip" /z/, and "genre" /ʒ/ (Scottish English has an additional voiceless fricative /x/, the last sound in the Scottish word "loch").

Fricatives are produced by making a constriction in the vocal tract which is narrower than the constriction for making a vowel and forcing air through the constriction quickly so that it makes a noise. The noise is the result of *turbulent airflow* – imagine a wide deep slow moving river, this has *laminar flow* (the airflow for making vowels and nasals is also laminar), now imagine a shallow fast-flowing river with lots of white water, this is turbulent flow.

The difference between the voiceless and voiced fricatives is that the vocal folds are vibrating for the voiced fricatives and held open for the voiceless fricatives.

/f/ and /v/ are produced by forcing air between a constriction made by holding the lower lip on the upper incisors – the air escapes between the teeth and out the sides of the mouth. For /θ/ and /ð/ the tip of the tongue is held against the upper incisors. For /s/ and /z/ the turbulence is caused by shaping the tongue so as to aim a jet of air at the upper incisors. For /ʃ/ and /ʒ/ the turbulence is caused by shaping the tongue so as to aim a jet of air at the lower incisors. /h/ is produced by holding the vocal folds close together and forcing air through the narrow opening between them.

FIGURE 19. Smoothed spectra of English voiceless fricatives.



Fricatives cause the resonances of the vocal tract to be excited but by a noise source, rather than by periodic voicing as is the case for vowels. /h/ can actually be analysed as a voiceless vowel rather than a fricative, e.g., /hi/ is [i̥i] and /hɛ/ is [ɛ̥ɛ] – if you whisper you are also producing voiceless vowels. For /s/ and /ʃ/ the shape of the vocal tract is, however, quite different from any vowel and

the source of the noise is near the front of the mouth rather than at the vocal folds. Since the source of the noise for /f/ and /θ/ is produced right at the opening of the mouth, the resonances of the vocal tract have little effect on their acoustic spectra. Smoothed spectra of English voiceless fricatives (spoken by the author) are shown in Figure 19.

Forensic value

[99.510] With the exception of /s/ and /ʃ/, fricatives are generally quiet which can make them difficult to measure if there is any background or channel noise. /f/ and /θ/ have similar spectra (see Figure 19) which can make them difficult to distinguish perceptually, and much of the spectral difference between /s/ and /ʃ/ occurs in the higher frequencies (see Figure 19) which are usually lost in telephone transmission, see **[99.610]**. The spectra of fricatives may be too much affected by telephone transmission to make them particularly useful for forensic voice comparison.

A research paper on forensic voice comparison in the likelihood-ratio framework which uses measurements of fricative spectra is Rose, Osanai, & Kinoshita (2003).

Plosives

Description

[99.520] English has three voiced and three voiceless *plosives* made using complete closures of the vocal tract at the lips, e.g., the first sounds in "bat" /b/ and "pat" /p/, at the alveolar ridge, e.g., the first sounds in "dip" /d/ and "tip" /t/, and at the velum, e.g., the first sounds in "gap" /g/ and "cap" /k/. First a closure of the oral cavity and a closure of the velopharyngeal port are made. Then the lungs are compressed to pump air into the oropharyngeal tube, increasing the air pressure behind the closure in the oral cavity. Finally the oral closure is released and a burst of air escapes.

For English utterance-initial voiced plosives the vocal folds usually begin vibrating just after the oral closure has been released. For English utterance-initial voiceless plosives the vocal folds don't usually begin vibrating until at least 30 ms after the oral closure has been released, and there is turbulent airflow at the vocal folds before voicing begins, this is called *aspiration*. The phonetic details of utterance-initial English plosives are therefore, for example, /b/→[p] and /p/→[p^h] ([^{h}] is the aspiration diacritic). The time between the release of the oral closure and the beginning of voicing is called *voice-onset time* (VOT).

The acoustics of plosives include the release burst, aspiration, and *formant transitions*. The formant transitions reflect the change in the shape of the mouth from the closed position of the plosive to the relatively open position of the adjacent vowel. Formant transitions are different for different places of articulation of the plosive, *bilabial* (lips), alveolar, and velar, but also depend on the shape of the mouth needed for the adjacent vowel, and hence the characteristic formants of the adjacent vowel. Formant transitions may be measurable during the aspiration stage as well as the voiced stage (see discussion of /h/, **[99.500]**). An utterance-final plosive may not have an audible release burst and the formant transitions from the preceding vowel may be the only acoustic indication that a plosive was made.

Forensic value

[99.530] Acoustic properties of plosives such as VOT and formant transitions could potentially be used for forensic voice comparison, but since they are dependent on both the plosive phoneme and the vowel phoneme they would probably be most effective for particular plosive-vowel or vowel-plosive sequences. These sequences would form phonetic units longer than a single phoneme and may have less variability. For example, if there are enough tokens of "day" in the known- and questioned-voice recordings then using the whole of the formant trajectory from the /d/ transition through the /eɪ/ may be more effective than using a collection of /eɪ/ tokens from multiple consonantal environments.

Laryngeal activity

Description

[99.540] The rate at which the vocal folds vibrate during voicing is known as the *fundamental frequency* (f0). Some speakers have longer and more massive vocal folds and others have shorter and less massive vocal folds, on average adult males have larger vocal folds than adult females but there is also variation within each sex. All else being equal, larger vocal folds vibrate at a lower f0 and smaller vocal folds vibrate at a higher f0 (think of long fat piano wires and shorter thinner piano wires). Fundamental frequency averages around 125 Hz for adult males and 200 Hz for adult females.

A speaker can stretch and tighten and lengthen their vocal folds or relax and slacken and shorten them. The tightening and slackening is like turning the tuning peg on a stringed instrument or stretching a rubber band. When the vocal folds are longer and tighter they vibrate at a higher frequency. Although all else being equal longer strings vibrate at a lower frequency than shorter strings, a string which is longer because it has been stretched vibrates at a higher frequency than the same string when it is shorter because it is slacker.

Humans can have quite fine control over their f0, and it is f0 values, not formant values, which correspond to notes in singing. In some languages, such as Standard Chinese, f0 is used to distinguish phonemes, e.g., "mā" /ma˥/ MOTHER with a steady high f0, and "mà" /ma˥˩/ SCOLD with a falling f0 ([˥] and [˥˩] indicate the f0 patterns). The different f0 patterns are called *tones*. English does not use f0 to distinguish phonemes but does use it to signal other differences such as increasing f0 towards the end of an utterance to indicate a question "it's three o'clock?" versus lowering it to indicate a statement "it's three o'clock". This use of f0 is called *intonation*. In addition, f0 can signal a speaker's emotional state – someone who is depressed may speak with a low frequency monotone, whereas someone who is excited may use a wide range of frequencies including high frequencies.

Voicing is caused by the speaker holding their vocal folds together and under tension, and compressing the lungs so that the air pressure below the vocal folds is higher than above the vocal folds. The increased air pressure pushes on the bottom of the vocal folds and eventually forces them to open. Air then escapes through the gap between the vocal folds and the air pressure below the vocal folds decreases. The air escaping between the vocal folds pulls them back together (this is due to an aerodynamic effect – the same effect causes an open door to slam shut on a windy day), and

the elasticity of the vocal folds also pulls them back together. Once the vocal folds are closed the pressure below them begins to rise again and the opening and closing cycle repeats.

Different speakers may differ not only in the frequency at which their vocal folds normally vibrate but also in other aspects of voicing. The relative amount of time in each cycle during which the vocal folds are open versus closed (*open quotient* versus *closed quotient*) may differ. The vibrations are unlikely to be perfectly regular and speakers may differ in the degree to which the amplitude of voicing varies across cycles (*shimmer*) and the degree to which the duration of individual cycles varies (*jitter*). It is possible to produce voicing without complete closure of the vocal folds, in particular a gap may be left between the *arytenoid cartilages* at the back of the vocal folds (see Figure 14). If there is turbulent airflow through this gap, *breathy voicing* is produced. Some speakers may have a habitually breathy voice (Marilyn Monro was famous for having a breathy voice). Holding the vocal folds tight together but with little tension results in irregular voicing which can be half the frequency of normal voicing. This is called *creaky voice*. Some speakers may have a habitually creaky voice (Louis Armstrong was famous for having a creaky voice (although his singing style may have used another type of voicing known as ventricular)). Damage to the vocal folds can cause habitual creaky voice. Some speakers have temporary creaky voice if they have not spoken for a long time, such as when they first speak in the morning. In some languages and dialects, such as Northern Vietnamese, creaky voice signals a phonemic contrast, e.g., "mì" /mi^{32}/ WHEAT and "mị " /mi^{21}/ [m̰i^{21}] COAX ([̰] is the diacritic for creaky voice and the numbers indicate the tones, historically the lowest tones have become creaky). Young speakers of North American and Australian English may use creaky voice sociolinguistically to indicate friendliness.

Forensic value

[99.550] Given the large within-speaker variation in f0, acoustic properties such as mean f0 are not expected to be particularly useful for forensic voice comparison. Properties such as the shape of f0 trajectories (changes over time) on tones or intonation patterns may have some value. Properties such as jitter may be more closely related to the physiology of the vocal folds and may have lower within-speaker variation.

A research paper on forensic voice comparison in the likelihood-ratio framework which uses measurements of f0 is Kinoshita, Ishihara, & Rose (2009).

Further reading

[99.560] There are number of introductory phonetics textbooks currently available. Textbooks at a basic introductory level include Ladefoged (2001, 2006) and Rogers (2000). Since its first edition in 1975, Ladefoged (2006) has probably been the most widely read phonetics textbook. Rogers (2000) includes a chapter describing a number of different dialects of English including Australian English and New Zealand English. Slightly more advanced textbooks include Clark, Yallop, & Fletcher (2007) and Johnson (2003).

VOICE RECORDING AND VOICE TRANSMISSION

Voice recording

[99.600] Almost all audio recording is now digital, and if an analogue recording were presented for forensic voice comparison it would be digitised before analysis. Sound is a pattern of vibrations in the air. When the vibrations hit a microphone, part of the microphone vibrates and produces an analogue electrical signal. This analogue electrical signal is then converted to a digital signal.

There are two factors affecting the quality of the digitisation itself: *sampling frequency*, and *word size* (multiplied by the sampling frequency to get *bit rate*). Sampling frequency refers to the number of times per second a measurement (a sample) is taken. A typical high-quality sampling rate is 44.1 kHz, which can be used to record audio frequencies up to 22.05 kHz. This is more than adequate for most audio recordings because humans cannot hear frequencies above about 20 kHz and this upper threshold decreases with age. Word size refers to the number of binary digits (bits) used to encode the amplitude of the signal at each sample. The larger the word size, the more detailed the representation of amplitude can be and the better the recording quality. A common word size is 16 bits which allows 65 536 different levels of amplitude to be encoded. Each sample has an amplitude value and moving from one sample to the next the amplitude value usually changes. At low sampling frequencies and low word sizes, there can be big steps from sample to sample, but at high sampling frequencies and high word rates the steps are very small and approximate smooth transitions (similar to the histogram example in **[99.220]**).

There are a number of additional factors which can affect recording quality. The speaker may be far from the microphone, or talking quietly, or there may be an object between the speaker and the microphone, in which case the electrical signal produced by the microphone in response to the acoustic signal of the speaker's voice may be of low amplitude, and when digitised the signal may only use a small part of the possible range of digital amplitude values. If there are other noises in the place the recording is being made, these may be loud relative to the acoustic signal from the speaker's voice hitting the microphone and both will be combined on the audio recording, with the noises partially obscuring the speaker's voice. Analogue components of recording systems, including microphones and amplifiers, produce their own electrical noise which will also form part of the recording. Turning up the gain on a microphone to compensate for a quiet audio signal may not work well because any acoustic noise and the electrical noise of the system will also be amplified.

On the other extreme, if the speaker's voice is too loud, or the gain on the microphone is too high, the electrical signal may exceed the maximum and minimum which can be digitally encoded and the highest amplitude parts of the signal are truncated. This is a phenomenon known as *clipping*, and results in the recording sounding fuzzy (a fuzz box attached to an electric guitar deliberately causes the signal to be clipped).

Different microphones and recording systems can have different frequency responses to the same acoustic signal, and the distance and angle of the speaker's mouth relative to the microphone and reflections of the sound off walls etc., can also result in differences in frequency response (such differences are collectively known as *channel effects*). Since known- and questioned-voice recordings may be made on different recording systems under different recording conditions, this can introduce a source of variability which is conflated with within-speaker variability.

Voice transmission

[99.610] In forensic-voice-comparison casework, questioned- and/or known-voice recordings are often recordings of telephone conversations. Telephone systems introduce additional channel effects, which can be a source of variability which is conflated with within-speaker variability.

Landline-telephone systems are now usually digital (at the level of the local exchange if not at the level of every handset), and the signal is usually digitised at a sampling frequency of 8 kHz with a bit rate of 64 kbits/s (word length of 8 bits, i.e., 256 levels of amplitude). Landline telephone systems only transmit frequencies between about 300 Hz and 3.4 kHz (this is known as a *bandpass*) and distort frequencies close to the edges of the bandpass. The bandpass is superimposed on the spectrum of the incoming audio signal. Some vowels such as /i/ and /u/ have intrinsically low F1 which for male speakers may be affected by the low end of the bandpass. F3 and above for females and F4 and above for males are likely to be affected by the high end of the bandpass. Although f0 falls below the low end of the bandpass it is recoverable from other parts of the spectrum (*harmonics*, which occur at multiples of the f0 value). For more on the effects of landline-telephone transmission on formant measurements, see Künzel (2001).

Mobile-telephone systems also apply a bandpass to the signal; the low end of the bandpass is maintained at 100 Hz (lower than for a landline system), and the high end varies between 2.8 kHz and 3.6 kHz. But in addition, mobile systems use compression and decompression algorithms (*codecs*) to reduce the amount of data sent, and this results in further deterioration of the signal. Some information in the signal is lost as the codecs are designed to reduce the amount of information sent by only keeping information which is most important for speech intelligibility. Some of the information lost may be information which would otherwise have been useful for forensic voice comparison. Since mobile telephone handsets have to communicate with base stations via radio transmissions, the quality of the transmission is affected by the presence of obstacles such as buildings between the handset and the nearest base station. Quality may also be affected by the number of users requesting service from that base station. Quality can change radically within a few seconds (in theory it could change every 20 ms), and, if the user is on the move, quality can change radically within a few metres. The signal is not sent continuously, but rather it is cut up into *packets* and each packet is sent in turn. Sometimes entire packets may be lost, but the system tries to ensure that as many packets as possible are received by lowering the amount of information about the speech signal in each packet, which also reduces the quality of the speech signal (bit rates can vary between 4.75 and 12.2 kbits/s, much lower than for a landline). Occasional missing packets may be replaced by the immediately preceding packets. For a more thorough relatively non-technical description of mobile-telephone systems and effects of mobile-telephone transmission on formant measurements, see Guillemin & Watson (2008).

Direct-to-satellite mobile-telephone systems compress the speech data more than terrestrial mobile-telephone systems leading to an even greater loss of information which may otherwise have been useful for forensic voice comparison.

Voice-Over-Internet-Protocol (VoIP) systems are broadly similar to mobile- telephone systems, but generally have a wider bandpass and use higher bit rates. They therefore transmit a higher quality signal than mobile-telephone systems and the problems in quality are associated with routing the information through the internet rather than with radio communication between handsets and base stations.

A challenge in forensic voice comparison has been and is to find acoustic properties of speech which are relatively robust to channel effects, and/or statistical procedures which compensate for the predictable aspects of channel effects. The challenges of working with mobile-telephone systems are generally greater than for landline systems.

APPROACHES TO FORENSIC VOICE COMPARISON

Introduction

[99.650] Historically there have been four basic approaches to forensic voice comparison: *auditory*, *spectrographic*, *acoustic-phonetic*, and *automatic*. What I mean by *approach* here is a methodology for extracting information from voice samples for the purpose of conducting forensic voice comparison. To a greater or lesser extent the approach can be independent of the *framework* for evaluating the evidence. For example, the acoustic-phonetic and automatic approaches can be characterised as general ways of turning acoustic information into numbers, which could then be evaluated using the likelihood-ratio framework or using some other framework. In this section, each of the approaches is described and then evaluated with respect to its compatibility with the new paradigm.

Auditory approach

Description

[99.660] Descriptions of the auditory approach (and auditory-acoustic-phonetic approach) can be found in Jessen (2008), Nolan (1997), and Rose (2002, 2006).The auditory approach is practised by phoneticians who may be drawing on years of training and experience in auditory phonetics, a tradition which includes using phonetic symbols and diacritics to transcribe the speech sounds which are heard. The phoneticians listen to the known and questioned voice samples and comment on any properties of the voices which may be shared and which in their experience they consider unusual, distinctive, or otherwise noteworthy, or any features which are noteworthy because they are present in one sample and unexpectedly absent in another.

Audible features which are exploited could be the sorts of differences which distinguish dialects, e.g., consider the way the word "height" (phonemically transcribed /haɪt/) would be pronounced by English speakers from the US Mid-West, Southern US, Canada, and Australia (in broad phonetic transcription these could be [haɪt], [haːt], [hʌɪt], and [hɔɪt] respectively). Such large dialectal differences are often salient even to the untrained listener, but an expert trained in auditory phonetics will be able to notice and systematically label smaller idiolectal differences.

Audible features could also be related to laryngeal activity, e.g., whether the voice is breathy or creaky **[99.540]**, or could be what might be considered speech impediments of varying severity, e.g., pronouncing "r"s as "w"s (/r/ as [w]). Again, although some of these features may be salient to untrained listeners, an expert trained in auditory phonetics will also be able to notice and systematically label smaller idiolectal differences.

In the auditory approach, the phonetician carefully documents all the features which they deem relevant and considers the ensemble of all of these features in arriving at an evidentiary statement for presentation in court. They will typically also consider some basic acoustic measurements thus implementing a hybrid auditory-acoustic-phonetic approach.

Evaluation

[99.670] Theoretically it would be possible to implement the new paradigm using some auditory features. For example, if the speaker in a voice sample had a stutter it would be possible to calculate frequencies of the occurrence of stuttering according to phoneme uttered and context (e.g., utterance initial or utterance medial). If such numeric data were derived from the known- and questioned-voice samples, and also from voices in a background database, then it would be possible to calculate a likelihood ratio. As far as I am aware, calculation of likelihood ratios from auditory features has only been reported in two unpublished theses (Elliot, 2002; Kirkland, 2003), and neither of the authors is now active in research or casework. Both Elliot (2002) and Kirkland (2003) calculated likelihood ratios on the basis of frequencies of the occurrence of allophonic variants of a number of phonemes.

The majority of features used in the auditory approach are intrinsically qualitative – sounds are fitted into boxes according to the phonetician's perception of them – making it impossible to calculate objective gradient measures of their similarity and typicality. Even if frequencies for different perceived allophonic variants of given phonemes can be calculated, more objective and detailed information could be extracted via acoustic measurements.

It would be possible to arrive at a subjective likelihood ratio via the phonetician's experience-based estimates of the similarity and typicality of the ensemble of all the features considered. In practice, however, (apart from the two unpublished theses mentioned above) I am not aware of anyone using the auditory approach (or auditory-acoustic-phonetic approach) who expresses their conclusions in the form of a likelihood ratio.

Theoretically it would be possible to measure the accuracy and precision of a practitioner of the auditory approach by having them compare a large number of pairs of samples, each known by the tester (but not the testee) to be of same or different origin. The subjective conclusions of an experienced practitioner of the auditory approach (or auditory-acoustic-phonetic approach) could turn out to be more accurate and precise than the output of an objective-acoustic-measurement data-based system. A small-scale evaluation including practitioners of the auditory-acoustic-phonetic approach was reported in Cambier-Langevald (2007), but, as far as I am aware, no large-scale evaluations have been conducted.

The Cambier-Langevald (2007) evaluation included five auditory-acoustic-phonetic submissions, four acoustic-phonetic submissions, two automatic submissions, and one spectrographic or auditory-spectrographic submission, tested on ten comparison pairs. It did not use accuracy and precision measures consistent with the new paradigm. On average, practitioners of the auditory-acoustic-phonetic approach were willing to give same-or-different responses to a larger percentage of the test pairs than were practitioners of other approaches (87.5% v 65%), but made a larger percentage of errors on the test pairs for which same-or-different answers were supplied (11.4% v 2.6%). Given the small size and other constraints on the evaluation, one should be cautious about generalising these results.

In conclusion, (with the exception of the two unpublished theses mentioned above) there do not appear to have been any attempts by practitioners of the auditory approach (or auditory-acoustic-phonetic approach) to work within the likelihood-ratio framework, and in general its reliance on experience-based subjective decisions rather than objective measurements and databases mean that the auditory approach is not well suited for use within the new paradigm.

Spectrographic approach

Description

[99.680] The spectrographic approach is based on visual inspection of spectrograms (see Figure 17 in **[99.460]**). The spectrographic approach is also known as *voicegram identification*, and as *voiceprinting* ("voiceprint" is a trademark, although not currently held by an entity which does forensic work). It is typically combined with listening, resulting in an auditory-spectrographic hybrid approach. The spectrographic / auditory-spectrographic approach is described in Kersta (1962), NRC (1979), and Rose (2002, pp. 107–111). Protocols for performing auditory-spectrographic forensic voice comparison have been developed by the *Federal Bureau of Investigation* (FBI) http://www.fbi.gov/, the *International Association of Voice Identification* (IAVI), the *International Association for Identification* (IAI) http://www.theiai.org/ (the IAVI became part of the IAI in 1980), and the *American Board of Recorded Evidence* (ABRE) http://www.abreboard.us/. The IAI no longer promulgates forensic-voice-comparison protocols (personal communication from J. Polski, Chief Operations Officer, IAI, February 2010). No response was received to an enquiry sent to ABRE.

The basic approach involves making a spectrogram of a stretch of speech, e.g., a word or a phrase, in the questioned-voice recording, and also making spectrograms of the same word or phrase spoken by the known speaker and by a number of foil speakers.

> The known, or another unknown voice sample, must be either wholly verbatim (preferred), or partially verbatim to allow meaningful comparisons with unknown voice samples. (ABRE, 1999, §5.2).
>
> Only speech sounds of similarly spoken words should be compared between voice samples. Comparison of the same speech sound but in different words, should be avoided. (ABRE, 1999, §7.1.2)

This usually requires recording the known speaker, and any foil speakers, saying the particular word or phrase which was said in the questioned-voice recording, either by reading aloud a written text or repeating the words spoken by a model speaker. The model speaker is not the questioned-speaker (or the questioned-voice recording) but should:

> recite the phrases in the same manner as the unknown speaker and have the suspect repeat them in a similar fashion. Ideally, the exemplar should be spoken in a manner that replicates the unknown speaker, to include speech rate, accent (whether real or feigned), hoarseness, or any abnormal vocal effect. (ABRE, 1999, §3.3.1)

Multiple phrases are collected and multiple repetitions of each phrase are collected so as to obtain an indication of intra-speaker variability.

The examiner is presented with the questioned-voice spectrogram and with a set of spectrograms for comparison. Where foil speakers are employed, the set of spectrograms will include spectrograms from foil speakers and may or may not contain the spectrograms from the known speaker. The task of the examiner is to choose the spectrograms in the comparison set which came from the same speaker who produced the questioned-voice spectrogram, or to say that none of the spectrograms in the comparison set came from the questioned-speaker. This would be repeated for multiple phrases, and in the aural-spectrographic approach the examiner would also listen to the

original recordings. Poza & Begault (2005) recommend the use of foils, but the ABRE protocols do not require this and the comparison may be made only between the known- and questioned-voice recordings.

The ABRE protocols §7.1.5 require the examiner to visually compare "General formant shaping and positioning", "Pitch striations", "Energy distribution", "Word length", "Coupling", i.e., nasality, and "Other"; and to auditorily compare "Pitch", "Stress/Emphasis", "Rate", "Disguise", "Mode", i.e., abruptness of voicing onset, "Psychological state", "Speech defects", "Vocal quality" (related to laryngeal activity), and "Other". In contrast, Poza & Begault (2005) recommend a gestalt approach:

> rather than trying to somehow rate individual characteristics, such as formant positions, phoneme durations, etc., the examiner allows experience in spectrographic pattern matching and knowledge of acoustic phonetics to guide him [*sic*] in evaluating the aggregate of appropriate patterns at his disposal. (Poza & Begault, 2005, p. 1)

The ABRE protocols require the examiner to state their conclusions on a seven-step subjective-posterior-probability scale: "Identification, Probable Identification, Possible Identification, Inconclusive, Possible Elimination, Probable Elimination, or Elimination" (ABRE, 1999, §7.3).

The FBI allows its auditory-spectrographic examiners to provide advice for investigative purposes, but does not allow them to testify in court.

Evaluation

[99.690] Through the 1960s to the 1980s there was a great deal of controversy over the spectrographic approach. A relatively brief review of the controversy can be found in Rose (2002, pp. 107–122), and a very thorough review can be found in Gruber & Poza (1995); the latter is highly recommended to anyone wishing to obtain a deeper understanding of the issues.

At least in the past at least some proponents of the spectrographic approach have made extravagant unproven claims about the accuracy of the procedure, claiming near infallibility, and claiming that real-world performance would be better than performance in controlled laboratory experiments (Tosi *et al*., 1972), the latter known as the *Tosi Extrapolation*. This lead to criticism and an increasingly vociferous debate; see, for example, Hollien's (1990, ch. 10; 2002, pp. 24–25, ch. 6) criticisms of the spectrographic approach and its practitioners, and Koenig's (2002) response. The 1991 version of the IAI protocols stated that the IAI "does not support or approve the use of any other voice identification technique not listed within these standards" (quoted from Gruber & Poza, 1995, §57). Note that the IAVI precursor to the IAI was established by proponents of the aural-spectrographic approach. In contrast, IAFPA, which is dominated by practitioners of auditory-acoustic-phonetic and acoustic-phonetic approaches, passed a resolution in 2007 stating that "The Association considers this approach to be without scientific foundation, and it should not be used in forensic casework" http://www.iafpa.net/voiceprintsres.htm. Practitioners of other approaches are likely to take offense if someone describes them as doing voiceprinting.

> Opponents claim the following: (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 [*sic*] 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.
>
> The strong opposition by so many scientists (as well as legal academics) to the easy and widespread acceptance of this type of evidence is partly a reaction to the image of near infallibility of voicegram examiners maintained by some of those advocating voicegram evidence, and to their apparent self-interest in defining the community of persons whose opinions should be considered in this matter. (Gruber & Poza, 1995, §6)

> While the community of "certified" forensic examiners has seemed intent on maintaining an image of near infallibility with regard to the potential for errors of false identification, in fact we have practically no information about error rates or accuracy under real-world conditions. (Gruber & Poza, 1995, §8)

It is rather interesting to compare the claims made by proponents of the spectrographic approach with similar claims made by proponents of current practice in fingermark comparison (and some other branches of forensic science). It is only relatively recently that criticism of the latter has emerged (Cole, 2006, 2009, 2010; Koehler, 2010; Saks & Faigman, 2008).

The procedures for collecting the known-voice recording have also come in for criticism, with the fear that a known-voice recording of an innocent speaker who is a good mimic could end up being identified as the source of the questioned-voice (Gruber & Poza, 1995, §63, §70; Rose, 2002, pp. 112–113); although it should be noted that in the ABRE protocol, the known-speaker is never asked to directly mimic the questioned-voice recording, and if there are foil speakers they are recorded under the same conditions. Practitioners of other approaches typically avoid explicitly making recordings of the known-speaker for the purposes of forensic voice comparison, and instead rely on existing recordings, such as police interviews or telephone calls made from remand centres, without requiring that they contain the same phrases spoken in the same way as on the questioned-voice recording.

Courts in some jurisdictions have excluded testimony based on the spectrographic approach, but courts in some other jurisdictions have allowed it. In *California v Siervonti* [(1985) No. CR-21695, Municipal Court of the Chico Judicial District, County of Butte, California] under the *Frye* standard [*Frye v United States*, 293 F. 1013 (D.C. Cir. 1923)] the court found that "the aural spectrographic analysis of the human voice for the purposes of forensic identification has failed to find acceptability and reliability in the relevant scientific community, and that therefore, there exists no foundation for its admissibility into evidence in this hearing pursuant to the law of California." In *State v Coon* [974 P.2d 386, 95 A.L.R. 5th 729 (Alaska 1999)] under *Daubert* the court admitted testimony based on the spectrographic approach but allowed both sides of the controversy to be presented (see discussion in Faigman *et al.*, 2008, §37.2; and Solan & Tiersma, 2003, pp. 423–425). In contrast, In *United States v Angleton* [269 F. Supp. 2d 892 (S.D. Tex. 2003)], also under *Daubert*, testimony based on the spectrographic approach was excluded (see discussion in Faigman *et al.*, 2008, §37.3). It is probably also the case that, although there has been much debate in the United States, in some other parts of the world testimony based on the spectrographic approach has been, and continues to be, admitted in court unopposed.

One of the dangers of the spectrographic approach is that the conversion of voice-samples from an acoustic signal to a picture may give a layperson (e.g., a police officer, a lawyer, a judge, or a jury member) the impression that the procedure is scientific, when in fact the comparisons of the spectrograms are made subjectively on the basis of the practitioner's experience.

It is also worth pointing out that the spectrographic approach is now anachronistic. The spectrographic approach was initially developed in the late 1950s / early 1960s using an analogue technology which was itself developed in the 1940s (Potter, Kopp, & Green, 1947; Joos, 1948). The analogue devices were superceded by specialised digital hardware in the 1980s, which were in turn superceded by software running on standard computers in the 1990s. Since the preliminary steps for creating a digital spectrogram on a modern computer include making objective numeric measurements of the acoustic signal and manipulating them using signal processing algorithms, objective numeric information extracted at the algorithmic stage can be directly used as the basis for forensic voice comparison, rather than relying on a human's subjective conclusion based on a graphical representation of this information.

From the perspective of the new paradigm, the spectrographic approach suffers from the same drawbacks as the auditory approach: Conclusions are subjective and experience-based rather than based on objective measurements and databases. Conclusions are, by prescription, expressed using a subjective posterior-probability scale. Although at least one large scale evaluation of the spectrographic approach (using hundreds of speakers) has been performed (Tosi et al., 1972), the applicability of its results to real-world conditions has been disputed (Gruber & Poza, 1995, §82–§87). Also, it does not appear to be current practice to assess validity and reliability by running a large-scale evaluation of each individual practitioner (or an associated group of practitioners who routinely supply second opinions on each other's work) before allowing them to testify in court (a recommendation to do this was made by Gruber & Poza, 1995, §61).

Acoustic-phonetic approach

Description

[99.700] Descriptions of the acoustic-phonetic approach can be found in Jessen (2008), Nolan (1997), and Rose (2002, 2006). The acoustic-phonetic approach is practised by phoneticians trained in acoustic phonetics and involves making quantitative measurements of acoustic properties of voice samples and statistically analysing the resulting numeric data.

The acoustic properties measured by practitioners of the acoustic-phonetic approach are typically those which have been found to be relevant in empirical studies of speech production and speech perception, for example formant frequencies, fundamental frequency, and VOT. The acoustic properties of many of the features used in the auditory approach can also be quantitatively measured to provide acoustic-phonetic features.

Typically, comparable phonetic units are identified in both known- and questioned-voice samples and then acoustic properties of these units are measured. A phonetic unit could be a phoneme, or a major allophone, but could also cover a shorter or longer stretch of speech. For example, a phonetic unit could be the most general allophone of the phoneme /aɪ/ (the vowel sound in the words "hi", "buy", "side" etc.), excluding allophones following /r/, /l/, /w/ (e.g., in "right", "light", and "wipe")

and the nasalised allophone preceding nasals – coarticulation with these consonants can result in tokens of these allophones having quite different acoustic properties from tokens of the general allophone. Another example of a phonetic unit is the /raɪt/ sequence in the word "right" up until the /t/ closure. Longer phonetic units will tend to have less contextual variation and also could potentially contain more acoustic information pertinent to speaker identity; for example, the formant trajectories in tokens of /raɪt/ may be more complex than those in tokens of /aɪ/ in general, and consonant transitions of /aɪ/ tokens taken from /raɪt/ will be more consistent than /aɪ/ tokens taken from numerous different consonantal contexts.

Usually the questioned-voice recording is shorter than the known-voice recording, so tokens of potentially usable phonetic units are first identified and marked in the questioned-voice recording. If the questioned-voice recording contains multiple tokens of a particular phonetic unit, and the number of tokens is considered sufficient for statistical analysis, then tokens of the same phonetic unit are sought in the longer known-voice recording, and if sufficient tokens are also found in that recording then acoustic properties of the tokens of this phonetic unit in both recordings are measured and subjected to statistical analysis ("sufficient" is ultimately related to the degree of accuracy and precision desired for the system). The procedure is typically applied to multiple phonetic units from the same sets of voice samples.

For examples of the use of this approach within the new paradigm, see González-Rodríguez et al. (2007), Morrison (2009c, 2010), and Morrison, Zhang, & Rose (2010).

It is also possible to use acoustic-phonetic-type measurements without explicit use of phonetic units. For example, formant frequencies or the fundamental frequency can be measured at regular intervals over the entire voiced portion of the voice sample, without regard to speech-sound identity beyond voiced versus voiceless, and these values can then be subjected to statistical analysis.

For examples of this approach, see Becker, Jessen, & Grigoras (2008, 2009), and Kinoshita, Ishihara, & Rose (2009).

Acoustic measurements are made using software implementations of signal-processing algorithms with human supervision of which parts of the voice samples to measure and of the parameter settings used by the algorithms.

Evaluation

[99.710] The acoustic-phonetic approach is well suited for use within the new paradigm: The acoustic measurements made are relatively objective, the numeric data generated can be, and have been, used to calculate likelihood ratios, and tests of validity and reliability can be, and have been, conducted.

It should be noted that use of the acoustic-phonetic approach does not guarantee compatibility with the new paradigm. The analysis of the numeric data could be performed in ways incompatible with the likelihood-ratio framework; for example, they could be subjected to discriminant analysis, which provides posterior probabilities for a closed set of speakers (see Morrison, 2008, Appendix).

Some acoustic-phonetic features, such as fundamental frequency and the second formant, have the advantage that they are expected to be relatively robust to channel effects, **[99.600]**, **[99.610]**. Another advantage is that human supervision of the signal-processing algorithms is expected to yield

higher accuracy and precision in the measurements. A disadvantage is that a great deal of human labour is involved in identifying and marking the phonetic units and in supervising the signal-processing algorithms. This will usually be the limiting factor in the number of phonetic units examined and the number of voice samples included in the background database.

The time and expense involved in implementing the acoustic-phonetic approach is its major drawback. To date (May 2010), evaluations of the accuracy and precision of acoustic-phonetic forensic-voice-comparison systems have only been conducted using test databases including voice recordings of tens of speakers and these have generally been high-quality recordings rather than recordings reflecting speaking style and recording quality conditions typical in forensic casework.

A criticism of the acoustic-phonetic approach, sometimes raised by signal-processing engineers, is that there is a degree of subjectivity involved. For example, when the acoustic-phonetician says they measured the stressed tokens of /aɪ/, the engineer wants to know, among other things, the explicit criteria for deciding what a stressed /aɪ/ token is and the explicit criteria for where it begins and where it ends. Although the acoustic-phonetician may be able to describe some general rules of thumb (heuristics), in any particular instance their decision may to a large extent depend on experience. However, if it is the case that human supervision generally leads to more accurate and precise measurements than using fully automatic procedures, and this results in greater accuracy and precision in the likelihood-ratio output of the forensic-voice-comparison system, then having a human in the loop will clearly be advantageous. The latter argument may have theoretical merit, but, as yet, there is no empirical evidence that this is the case. If tasks performed by a human as part of the forensic-voice-comparison system could be replaced by computer software without overly deleterious effects on the system's accuracy and precision, then the system would be more objective and probably operate more quickly and economically.

Automatic approach

Description

[99.720] Descriptions of the automatic approach to forensic voice comparison can be found in Jessen (2008) and Ramos Castro (2007). For a review of approaches to automatic speaker recognition in general, see Kinnunen & Li (2010).

The automatic approach to forensic voice comparison was developed by signal-processing engineers, and draws heavily on research on automatic speaker recognition developed for non-forensic applications, e.g., a voice-recognition password system for telephone banking. Much of the acoustic analysis, signal processing, and statistical modelling applied in automatic forensic voice comparison is the same as that applied in automatic speaker recognition for other applications, but for forensic applications the final stage is to produce a forensic likelihood ratio.

As with the acoustic-phonetic approach, the automatic approach is based on quantitative measurements of acoustic properties of speech, but typically no attempt is made to exploit information relating to phonetic units. Also, as implied by the name, once the system has been designed and built it operates fully automatically, with the human *rôle* restricted to inputting the audio recordings and reading the output.

Typical features in an automatic system are *mel-frequency cepstral coefficients* (MFCCs) representing the shape of the spectrum. The spectrum is measured in a frame of length 20–30 ms, and this frame is moved through the entire speech-active portion of the recording in steps of 10–20 ms. This results in a long series of sequential MFCC measurements of the spectra of the speech signal.

Some automatic-speaker recognition systems make use of "higher level features", for example, they use automatic-speech-recognition systems to divide the acoustic signal into phonetic units (which may not correspond to the phonetic units which a phonetician would extract), or they automatically extract fundamental-frequency trajectories (Shriberg & Stolke, 2008). Such features could also be exploited by automatic forensic-voice-comparison systems. Some human supervision could also be incorporated into an acoustic-phonetic-automatic hybrid approach.

For examples of the use of the automatic approach within the likelihood-ratio framework, see González-Rodríguez *et al*. (2006), González-Rodríguez *et al*. (2007), Meuwly & Drygajlo (2001), Morrison, Thiruvaran, & Epps (2010b), Thiruvaran, Ambikairajah, & Epps (2008).

Evaluation

[99.730] The automatic approach is well suited for use within the new paradigm: The acoustic measurements made are objective, the numeric data generated can be, and have been, used to calculate likelihood ratios, and tests of validity and reliability can be, and have been, conducted.

It should be noted that, as with the acoustic-phonetic approach, use of the automatic approach does not guarantee compatibility with the new paradigm. It is especially important to distinguish an automatic forensic-voice-comparison system compatible with the new paradigm from an automatic speaker-recognition system developed for some other purpose. The latter sort of system usually makes a yes/no decision on the basis of its calculated posterior probability for the voice belonging to a particular speaker on which it has been trained, or belonging to one of a limited number of speakers on which it has been trained. Most researchers working on automatic speaker recognition are not knowledgeable about forensic science, and some may mistakenly believe that their existing systems can be used as-is for forensic work.

The great advantage of the automatic approach to forensic voice comparison is that it is automatic, and can therefore analyse massive amounts of data without concerns about human labour costs. Assessments of automatic-speaker-recognition systems routinely use data from hundreds of speakers. In evaluations of automatic forensic-voice-comparison systems the primary constraint may be the number of speakers from the relevant population from whom one can collect voice samples, rather than the time and expense involved in analysing those samples.

Typical automatic features such as MFCCs are not particularly robust to channel effects, but statistical procedures have been developed to compensate for channel differences.

Some phoneticians have criticised the automatic approach for not being based on theoretical and empirical research on human speech production and perception, a criticism which they have also made about the spectrographic approach; however, although a typical automatic forensic-voice-comparison system does not explicitly exploit information about phonetic units, its ability to process much more data may allow it to outperform an acoustic-phonetic system in accuracy and precision

(to date, May 2010, no large-scale direct comparisons of these two approaches have been conducted).

A danger with an automatic system is that, although the system may be properly designed, it is a piece of software which could be inappropriately used by someone who is not sufficiently knowledgeable about phonetics and forensic science. As with any software system, if you put garbage in you get garbage out (GIGO). The operator needs to be aware of potential problems related to issues such as speaking-style mismatches, recording quality, and selection of the relevant population from which to collect voice recordings for the background and test databases.

EXAMPLES

Introduction

[99.770] This section provides two examples of forensic voice comparison conducted within the new paradigm. The first example uses an acoustic-phonetic approach and the second uses an automatic approach.

The examples are taken from research studies aimed at testing the validity and reliability of different techniques, and in terms of database size and selection and recording conditions they do not meet all of the conditions which would be imposed for use in casework. This is a standard research strategy. Testing a number of new techniques on smaller "easier" databases is more economical, and those which show promising results can later be tested on larger, more forensically realistic, more challenging databases.

Acoustic-phonetic example

[99.780] This example is based on research reported in Morrison (2010) and Morrison, Thiruvanran, & Epps (2010a) which itself builds on earlier research reported in Morrison (2009c). More details can be found in those papers.

The system is based on measurements of the shape of the formant trajectories of diphthongs. Research on speech perception indicates that the initial and final formant values in a diphthong are important for the perception of phoneme identity, but the exact shape of the changes over time in formant values between the initial and final values are not important for the perception of phoneme identity (Morrison, 2011). This means that, to a large extent, as long as an individual speaker produces appropriate formant values at the beginning and end of a diphthong, they are free to choose a trajectory between these values which suits the physiology of their vocal tract and any personal preferences or idiosyncrasies. Theoretically, therefore, the details of the shapes of formant trajectories in diphthongs potentially have large between-speaker variation relative to within-speaker variation, making them potentially good features for use in forensic voice comparison.

Data

[99.790] This research made use of a database of voice recordings collected by Yuko Kinoshita and first reported in Kinoshita & Osanai (2006). The recordings were laboratory quality. Twenty-seven adult male speakers of Australian English were recorded on two separate occasions reading sentences which contained tokens of five vowel phonemes /aɪ/, /eɪ/, /oʊ/, /aʊ/, and /ɔɪ/.

The recording conditions and speaking style were not realistic compared to typical conditions for forensic-voice-comparison casework, and the database was small, but it was used to test some new techniques and allow comparison with old techniques previously applied to the same database.

Acoustic analysis

[99.800] Tokens of the five vowel phonemes were isolated and the values of the first three formants measured at intervals of 2 ms. The trajectories of F2 were used in the current example. A curve known as a *discrete cosine transform* (DCT) was fitted to each F2 trajectory (see details below). A third-order DCT was used, this results in four numbers which together describe the shape of the curve fitted to a trajectory.

In a DCT, each number (a *coefficient estimate*) is a multiplier applied to the amplitude of a simple curve (a *basis function*). Figure 20 shows the four basis functions of a third-order DCT. The basis functions and their corresponding coefficients are numbered zeroth, first, second, and third. Figure 21 shows the DCT curve (solid line) fitted to the F2 trajectory of an /aɪ/ token. The coefficient estimates for this DCT curve were +1491, −447, +79, and +34 (in hertz scale). The zeroth coefficient value, +1491, is the mean of the F2 values over the entire trajectory, the first coefficient value, −447, is the size of an S-shaped deviation from the mean, the second coefficient value, +79, is the size of a U-shaped deviation from the curve resulting from adding the mean and the S-shaped deviation, etc.

FIGURE 20. Zeroth through third DCT basis functions.

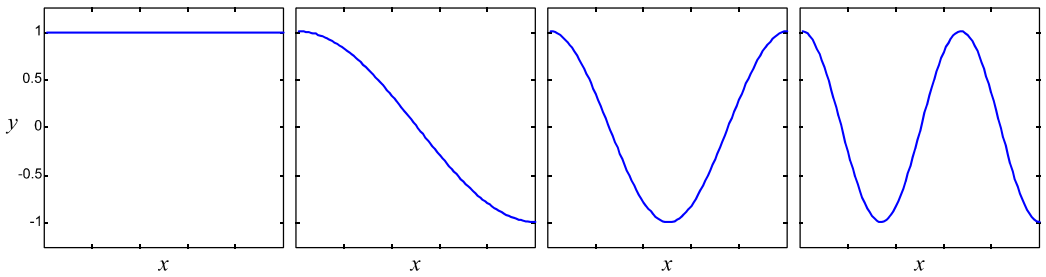
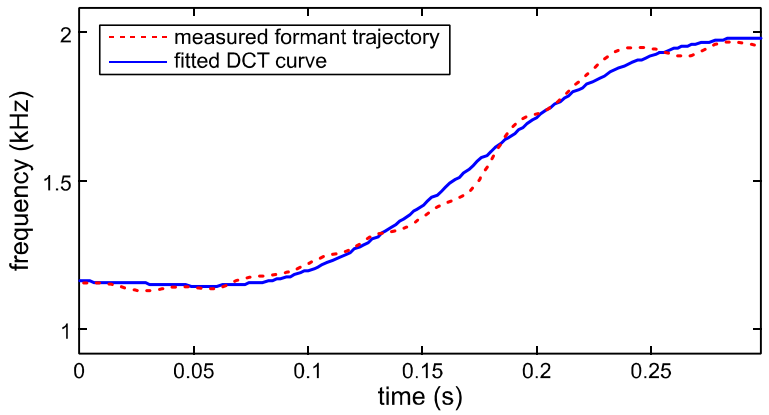


FIGURE 21. DCT curve fitted to F2 trajectory.



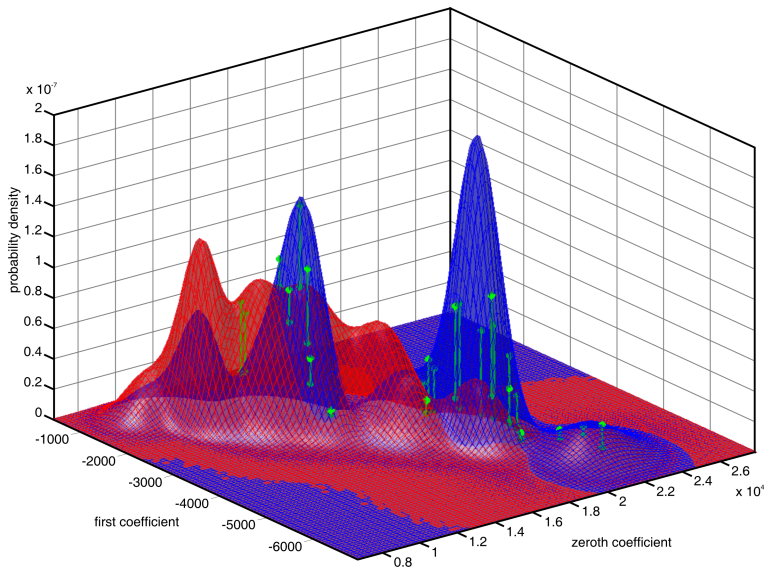
The DCT coefficient values were estimated for each vowel token in the data, and the calculation of likelihood ratios was based on these numbers.

Likelihood-ratio calculation

[99.810] Likelihood ratios were calculated using Gaussian mixture models (GMMs), see **[99.230]**. A procedure known as *cross validation* was adopted such that for the calculation of each likelihood ratio the background database consisted of data from all speakers except for the speaker or speakers whose data were being compared. Likelihood ratios were calculated for each possible same-speaker and each different-speaker combination in the data. Separate sets of likelihood ratios were calculated for each of the five vowel phonemes.

Figure 22 shows an example of a background model, a suspect model, and offender data points being tested for /aɪ/. There were 14 Gaussians in each model in the actual analysis. Only the distribution of the first two DCT coefficients and a mixture of seven Gaussians are shown in Figure 22 (the coefficient values were not scaled in hertz).

FIGURE 22. GMM background and suspect models.



The suspect model was built on the basis of multiple tokens of the vowel phoneme from the test suspect recording. The procedure actually uses the background model as a start point and then adapts the model towards the suspect data (Reynolds, Quatieri, & Dunn, 2000). Multiple tokens of the vowel phoneme from the test offender recording were then used as probe data: the likelihood of obtaining the values of each token given the suspect model and given the background model were

calculated (for an example, see the green dots and connecting lines in Figure 22), and a likelihood ratio calculated as the ratio of these for each token. The likelihood ratios from all these tokens were then multiplied together to derive a single score for the comparison of this simulated suspect sample and this simulated offender sample. Although this is a standard procedure (Reynolds, Quatieri, & Dunn, 2000) these scores cannot be treated as likelihood ratios because to do so would be a violation of the assumption of statistical independence. The assumption of statistical independence would have to be valid if simple multiplication of likelihood ratios were to be used to produce a fused likelihood ratio; however, measurements based on multiple tokens of the same vowel phoneme from the same recording cannot be assumed to be statistically independent (nor could statistical independence be assumed for parallel sets of likelihood ratios/scores based on different vowel allophones from the same recordings, nor for an automatic system could MFCCs from multiple frames across a recording be assumed to be statistically independent). To produce a valid likelihood ratio for each comparison (each pair of recordings) on a single phoneme, the scores were therefore calibrated using logistic regression **[99.290]**.

There were parallel sets of scores from each phoneme, i.e, for each pair of recordings there was a score from each of the five phonemes, and these were fused and calibrated using logistic regression **[99.290]**. Only the results for the fused system are reported below. Calibration and fusion were conducted within the cross-validation procedure.

Results

[99.820] Morrison (2009c) compared different procedures for describing the shape of formant trajectories using a small set of numbers – third-order DCT was one of the most successful. Morrison (2010) compared the GMM procedure with another procedure previously used for the calculation of likelihood ratios from acoustic-phonetic data. The GMM procedure was found to be much better than the old procedure both in terms of accuracy and precision. The C_{llr} value of 0.030 obtained for the GMM system was less than one sixth of that for the system using the old procedure, and the estimated 95% credible interval was much narrower.

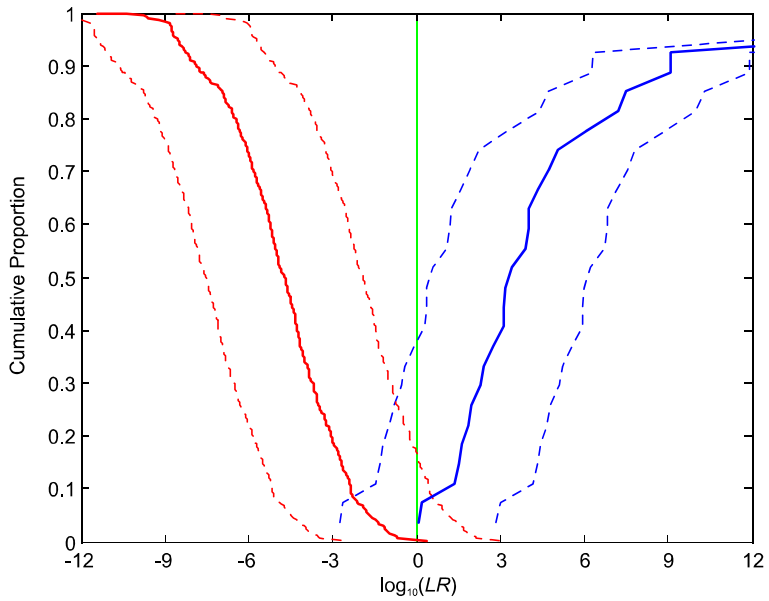
Figure 23 shows the Tippett plot of the likelihood-ratio results from Morrison, Thiruvanran, & Epps (2010a), including the estimated 95% credible interval shown as the dashed lines to the left and right of the different-speaker and same-speaker solid lines. The different-speaker solid line is the mean log-likelihood-ratio values over multiple non-overlapping pairs of samples taken from the pairs of speakers. Since only two non-contemporaneous recordings of each speaker were available, the 95% credible interval was estimated only using data from different-speaker comparisons using a parametric estimation procedure. In a $\log_{10}$ scale, the estimated 95% credible interval was ±2.81.

If the background and test data were consistent with the conditions in a case, and the comparison of the known- and questioned-voice samples resulted in a likelihood ratio of, say, 1/1 000 000 ($\log_{10}(LR)$ of −6), then the forensic scientist could make a statement of the following sort:

> Based on my evaluation of the evidence, I have calculated that one would be one million times more likely to obtain the acoustic differences between the voice samples if the questioned-voice sample had been produced by someone other than the accused than if it had been produced by the accused. Based on my calculations, I am 95% certain that it is at least 1 500 times more likely and not more than 650 million times more likely. In tests of my

forensic-voice-comparison system, none of the twenty-seven same-speaker comparisons provided greater or equal support for the different-speaker hypothesis.

FIGURE 23. Tippett plot including 95% credible intervals.



Automatic example

[99.830] This example is based on research reported in Morrison, Thiruvaran, & Epps (2010b). More details can be found in that paper.

Data

[99.840] Data consisted of recordings of telephone conversations involving adult male American-English speakers. The background database consisted of 800 recordings of approximately 200 speakers. A calibration database consisted of two non-contemporaneous recordings of each of 32 different speakers. The test database consisted of four non-contemporaneous recordings of each of 100 different speakers. The use of three separate databases avoids the need for cross validation.

Although the recording conditions and speaking styles were more forensically realistic than for the acoustic-phonetic example **[99.790]**, and the sizes of the databases were relatively large, the selection of dialect spoken was not as careful as it would need to be for forensic casework. The source databases used included the "American-English" designations but no finer grained classification of dialect, and no attempt was made to subdivide the recording by sub-dialect.

Acoustic analysis

[99.850] The spectrum of the whole of the speech-active portion of each recording was measured every 10 ms. The spectra were quantified using mel-frequency cepstral coefficients (MFCCs). MFCCs are similar to the DCTs described in the acoustic-phonetic example **[99.800]**, but they are used to quantify the relationship of amplitude to frequency, i.e. they describe the shape of a spectrum. At every measurement point sixteen MFCCs were used plus some additional measures of short-term spectral change calculated on the basis of the MFCCs.

Likelihood ratio calculation

[99.860] Likelihood ratios were calculated for each possible same-speaker and different-speaker combination in the test database. The likelihood ratios were calculated using GMMs with the background model built using the background database. Logistic regression was used to calibrate the scores **[99.290]**, **[99.810]**. The weights used in the calibration were calculated using the calibration database. There was no overlap in the use of the background, calibration, and test databases. No statistical procedure for dealing with potential channel effects was applied in this simple system.

Results

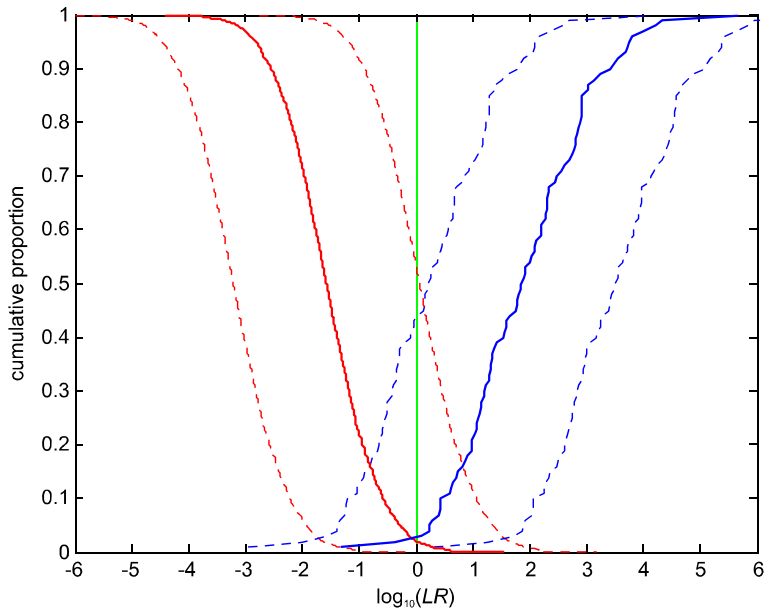
[99.870] Morrison, Thiruvaran, & Epps (2010b) compared the system under two test conditions: one using 40 ms of speech from each test offender, and one using 20 ms – the amount of speech available in questioned-voice recordings is often quite short. The accuracy of the likelihood ratios from the 40 ms and 20 ms was almost identical (to three figures $C_{llr} = 0.150$ for both), but the precision of the likelihood ratios from the 40 ms tests was a little better (±1.63 versus ±1.69 on a $\log_{10}$ scale using a parametric estimate of the 95% credible interval).

Figure 24 shows the Tippett plot of the likelihood-ratio results from the 40 ms tests. The estimated 95% credible intervals are shown as the dashed lines to the left and right of both the same-speaker and different-speaker solid lines. The solid lines are the mean log-likelihood-ratio values over multiple non-overlapping pairs of samples taken from the same speaker / same pairs of speakers (two for each same-speaker comparison and four for each different-speaker comparison).

If the background and test data were consistent with the conditions in a case, and the comparison of the known- and questioned-voice samples resulted in a likelihood ratio of, say, 100 ($\log_{10}(LR)$ of +2), then the forensic scientist could make a statement of the following sort:

> Based on my evaluation of the evidence, I have calculated that one would be 100 times more likely to obtain the acoustic differences between the voice samples if the questioned-voice sample had been produced by the accused than if it had been produced by someone other than the accused. Based on my calculations, I am 95% certain that it is at least 2.3 times more likely and not more than 4 300 times more likely. In tests of my forensic-voice-comparison system, of 19 800 different-speaker comparisons, 11 (0.056%) provided equal or greater support for the same-speaker hypothesis.

FIGURE 24. Tippett plot including 95% credible intervals.



COMPARISON OF TECHNICAL FORENSIC VOICE COMPARISON AND NON-TECHNICAL SPEAKER IDENTIFICATION

Introduction

[99.910] *Technical forensic voice comparison*, or simply forensic voice comparison, is performed by forensic scientists, and has been the topic of this chapter so far. In contrast *non-technical speaker identification* refers to the general ability of a person with no training in forensic voice comparison to recognise a voice and identify the speaker.

Sometimes non-technical speaker identification is performed by an *earwitness* who is present at the scene of a crime, hears the offender speaking, and either immediately recognises the offender's voice as belonging to a particular person they already know, or later attempts to pick the speaker out of a voice lineup. If no audio recording of the crime being committed is available, a technical forensic voice comparison is not possible. (The topics of voice lineups and earwitness testimony *per se* are not covered in this chapter.)

In other instances audio recordings of both the offender and suspect are available and a forensic voice comparison performed by a forensic scientist is potentially possible. Despite this, in some jurisdictions police officers who are laypersons with respect to forensic voice comparison are allowed to listen to the audio recordings and testify in court as to the identity of the questioned voice.

This section describes the problems with allowing an untrained layperson to do the work of a forensic scientist. It begins by describing some basic differences between non-technical speaker identification and technical forensic voice comparison **[99.920]**, describes some factors affecting the accuracy of non-technical speaker identification **[99.970]**, and finally relates these factors to an example of testimony provided by a police officer **[99.1040]**.

Note that (as will be apparent in some of the quotations below) "reliability" has often been used in the literature on non-technical speaker identification in place of "validity", although it is usually clear that it is validity rather than reliability which is being discussed (see **[99.290]**).

Non-technical speaker identification versus technical forensic voice comparison

[99.920] Rose (2002, p. 92) distinguishes two tasks which I will refer to as *non-technical speaker identification* and (technical) *forensic voice comparison*.

Non-technical speaker identification

[99.930] Non-technical speaker identification refers to the ability of almost all humans to identify a speaker simply by listening to their voice. If a relative or friend phones us and immediately starts a conversation without identifying themself, we do not normally need to resort

to caller ID or challenge the speaker to identify themself, we simply recognise them on the basis of their voice.

There are also instances when we do not recognise the voice as belonging to a speaker who is known to us, in which case we identify the voice as belonging to some unidentified person, perhaps someone we do not know well but has some business phoning us, or perhaps someone who has dialled the wrong number.

Note that sometimes our non-technical identification of the speaker may be incorrect, we may have mistaken the voice of one relative for the voice of another, or we may have mistaken the voice of a stranger for the voice of a friend. We may also be aware that we are not 100% certain as to the identity of the speaker.

Non-technical speaker identification is generally holistic in character. The listener simply states who they think the voice belongs to, and has difficulty identifying properties of the voice that allow them to distinguish it from other voices. Even if the listener can identify particular properties of the voice which have lead them to a particular identification, as a lay person they typically have a very limited vocabulary with respect to describing acoustic, phonetic, or other properties of voices.

Technical forensic voice comparison

[99.940] In contrast, technical forensic voice comparison is performed by forensic scientists with extensive training in phonetic science or speech processing, and training in forensic science. Typically the level of training is such that they have obtained a doctorate in a relevant field. A forensic scientist analyses multiple properties of the voice, typically making use of acoustic measurements and statistical analysis to quantify the results. A forensic scientist working within the new paradigm provides the court with a statement as to the strength of the evidence considering both the prosecution's proposition that the known and questioned voices were produced by the same speaker, the defendant, and the defence's proposition that the questioned voice was produced by someone other than the defendant. A forensic scientist working within the new paradigm will also test and report the validity and reliability of the methods by which they arrived at this strength-of-evidence statement.

Mistaken beliefs about non-technical speaker identification

[99.950]

> Because of the common experience of readily recognizing the voices of relatives, friends, and fellow workers ... and identifying the voices of many politicians, actors, and radio and television personalities ..., a myth exists that all voice recognition is accurate and reliable. (Yarmey et al., 2001, p. 284).

It is perhaps because they are aware of their own capacity to perform non-technical speaker identification that police officers, lawyers, judges, and jury members may believe that they themselves or other lay people can determine whether a questioned voice on a recording is the same as the voice on a recording known to be that of a suspect or defendant. In contrast, few police officers, lawyers, judges, or jury members would believe that a lay person could evaluate DNA

evidence or fingermark evidence and would immediately call upon a trained forensic scientist for assistance.

> When it comes to admitting tape-recorded evidence, judges sometimes seem to assume that law enforcement officers will be particularly good at this task, but little evidence supports this assumption. Interestingly, experimental evidence suggests that police officers are no better at eyewitness identification than lay witnesses. (Solan & Tiersma, 2003, p. 403)

> The trial court found that there was "no extensive scientific basis that 'earwitness' identification is as susceptible to the same misidentification as eyewitness identification." [However] . . ., voice identification is probably even more problematic than eyewitness testimony. We see no reason for refusing to give an instruction that could help jurors decide more analytically how much weight to give an identification. . . . Courts Should Allow Expert Witnesses to Testify on the Reliability of Earwitness Identification. (Solan & Tiersma, 2003, p. 432)

See also Bull & Clifford (1999b, pp. 202–203).

> Trial judges have assumed that jurors are adequately informed about the reliability of person identification. In contrast, empirical evidence suggests that there are wide discrepancies between lay people's opinions and scientific findings about the reliability of face and voice identification. (Yarmey et al., 2001, p. 286)

True earwitnesses versus listening to audio recordings

[99.960] It is important to distinguish between non-technical speaker identification performed by a so called *earwitness*, i.e., someone present at the scene of the crime who heard the voice of the offender while the crime was being committed and there are no audio recordings of the crime, and non-technical speaker identification performed by someone listening to an audio recording of the voice of the offender recorded while the crime was being committed and comparing that with audio recordings of the voice of a suspect. As will be explained below, non-technical speaker identification is of unknown validity unless the individual listener can be tested under circumstances similar to those of their identification of the questioned voice; however, if earwitness testimony is all that is available and the court is apprised of the problems with its validity then arguments can be made that earwitness testimony is likely to be of assistance to the court and should therefore be admitted. In contrast, when audio recordings of the questioned and known voices are available there is no need to rely on non-technical speaker identification with its problematic validity. Instead technical forensic voice comparison of demonstrable validity and reliability can be performed by a trained forensic scientist. When the lay person performing the non-technical speaker identification is not actually an earwitness, but instead someone who has listened to the audio recordings for the specific purpose of making a speaker identification, I cannot see any logical argument for why this should be admitted in court in place of technical forensic voice comparison.

Validity of non-technical speaker identification

[99.970] There are multiple factors which are reported in the scientific literature as being related to listeners' abilities to correctly identify voices:

Variability between listeners

[99.980] Some listeners are good at identifying speakers from their voices and some listeners are poor at identifying speakers from their voices. In experiments under different conditions, different listeners have been found to range from 100% correct to at-chance performance.

Age has been identified as a factor which is related to individual listener differences: Bull & Clifford (1984) found that older listeners do not tend to perform as well as younger adult listeners (this may be due to general age-related hearing loss). However, it is not the case that a particular older listener will necessarily perform worse than a particular younger listener.

There is substantial between-listener variation which, at least at present, can only be accounted for as idiosyncratic.

In order to assess the validity of an individual listener's non-technical speaker identification, that listener would have to be tested: They would have to identify a large number of voice samples so that their correct-identification rate could be obtained. In order for such a test to be meaningful to the court, it would have to be conducted under conditions similar to the conditions under which the listener made the identification of the questioned voice. The number of voice samples which would have to be identified would depend on a trade off between the desired level of precision and practicality, 10 would be very practical but likely have insufficient precision, 1000 would probably satisfy everyone's concerns about precision but would likely be very impractical.

Listener certainty

[99.990] Listeners may vary in their certainty that their speaker identification is correct. Bull & Clifford (1999a, pp. 217–218) summarise studies on the relationship between listeners' certainty and their correct-identification rates. Most such studies focussed on differences between listeners, i.e., whether a listener who expresses a greater degree of certainty in their identification is actually more likely to be correct than a listener who expresses less certainty. In voice-lineup situations a positive correlation has been found between certainty and correct-identification, but only when the offender voice was included in the lineup and not when the offender voice was not included, even though the listeners were told that the offender voice may or may not be in the lineup. This is to say that when the offender voice is not included in the lineup a listener may with relatively high certainty incorrectly identify one of the voices in the lineup as the offender voice.

This is rather worrying if one considers the situation where the police are mistaken and their suspect is not the offender but the police have focussed on this suspect in part because he or she sounds superficially similar to the offender. The listener may pick out the suspect's voice because it is the voice in the lineup which sounds most similar to that of the offender.

> At the present time the conclusion has to be that earwitness confidence should not be taken to indicate that one witness is more likely to be correct than another witness. (Bull & Clifford, 1999a, p. 218).

This opinion echoes that of Yarmey (1995, pp. 802–803) and was again echoed by Rose (2002, p. 101).

What may be more relevant than the between-listener relationship between certainty and correct-identification is the within-listener relationship; whether, for example, on tests on which an individual listener says they are 50% certain they are actually correct 50% of the time, and on tests on which they say they are 75% certain they are actually correct 75% of the time, etc., or whether, for example, instead when they say they are 50% certain they are actually correct 25% of the time, and when they say they are 75% certain they are actually correct 50% of the time. Clifford, Bull, & Rathbom (1980, cited in Bull & Clifford, 1999a, p. 218) and Bull & Clifford (1984, pp. 121–123) found a positive correlation between individual listeners' certainty and their correct-identification rates, but again only in cases in which the offender voice was included in the lineup. Also, note that in both the examples above (50-50, 75-75 versus 50-25, 75-50) the correlation between certainty and correctness is 100%, but it would not be appropriate to give the same weight to a listener's certainty statement in each case.

Solan & Tiersma (2003, p. 412) comment that:

> People rely on an identifier's level of confidence in judging how accurate the identification is likely to be. But that level of confidence correlates only slightly with the likelihood of accuracy. The result is that people tend to place too much credence in an identification.

For a listener's stated degree of certainty in their identification of the questioned voice to be truly of value to the court, the listener would have to be tested to find the relationship between their degree of certainty and their correct-identification rate.

Listener's familiarity with speaker's voice

[99.1000] Listeners are generally better at identifying the voices of familiar speakers such as relatives, friends, acquaintances, and media personalities, whereas they are poorer at identifying the voices of less familiar speakers.

"Familiar" in the literature is usually used to describe a voice which a listener has heard on many occasions over large stretches of time (typically years) summing to a long duration (at least many hours) of exposure to the voice. This includes substantial exposure to the within-speaker variability of the voice in different contexts (different degrees of formality and tiredness, different interlocutors etc.). Familiarity is not, of course, a binary construct and listeners may be more familiar with some voices and less familiar with others (Yarmey *et al.*, 2001).

Another sense in which "familiar" could be used applies in the situation where a listener repeatedly listens to audio recordings of a speaker's voice in order to deliberately "familarise" themself with the voice. Such a procedure is unlikely to expose the listener to the same duration and variety of the speaker's voice over the same period of time as would be the case for relatives and friends, and even possibly as would be the case of media personalities. Speaker-identification performance on

"familiarised" speakers would therefore be expected to be poorer than on highly-familiar speakers such as relatives and friends.

> Police, interpreters and even experts may have some exposure or limited familiarity with a suspect but not enough familiarity to match the 'specialised knowledge' of a family member, partner or good friend. (Edmond & San Roque, 2009, p. 31).

In Clarke & Becker (1969, cited in Ladefoged & Ladefoged, 1980, p. 45) listeners familiarised themselves with audio recordings of 20 previously-unknown speakers over a period of several weeks, but correct-identification rates only increased from 63% to 67%.

Although the identification of familiar voices has been found to be better than that of unfamiliar voices it has not been found to be perfect. Even listeners who one would expect to be highly proficient can make mistakes even with very familiar voices: On several tests, world-famous phonetician, Peter Ladefoged, failed to correctly identify the voice of his own mother (Ladefoged & Ladefoged, 1980, p. 49).

Foulkes & Barron (2000, pp. 182–183), Rose (2002, pp. 98–99), and Solan & Tiersma (2003, p. 411) are of the opinion that the ability of listeners to identify familiar voices is generally overestimated, including by members of the legal profession; Yarmey (1995, pp. 807–808) summarised empirical research demonstrating that this is indeed the case for undergraduate psychology students.

Typicality of speaker's voice

[99.1010] Some speakers have voices which are very distinctive, their speech patterns or the acoustic properties of their voice are unusual in that they are atypical in the population at large. Such speakers are usually easier to identify than speakers who have typical voices.

If two speakers are selected at random from the population, then it is more likely that they have relatively typical voices than that one or both have atypical voices. Recordings of two speakers with typical voices will sound similar to each other, not because the voices are produced by the same speaker, but simply because the two speakers both have typical voices. To illustrate: Amongst the population of Hollywood actors, Sean Connery has a very atypical voice in that his accent is Scottish rather than the more typical General American and he has a lisp rather than the typical lack of a lisp. Sean Connery's voice would therefore be very easy to identify in a selection of recordings of Hollywood actors. In contrast, the voices of Hollywood actors with typical voices, actors who speak General American English and have no speech impediments, would be harder to identify. Note also that if we change the population from Hollywood actors to Scotsmen with lisps, then Sean Connery's voice becomes more typical and he will be harder to identify in a selection of audio recordings of Scotsmen with lisps.

There are exceptions to the general pattern of less typical voices being easier to correctly identify: If two speakers have atypical voices which are both atypical in the same way (this may occur if, from the listeners's perspective, they both speak the same unusual dialect, or they are close relatives, etc.) then this might increase the likelihood that a listener will misidentify one for the other and be overconfident in their identification compared to if both speakers had more typical or differently atypical voices (for evidence of this see Ladefoged & Ladefoged, 1980, p. 47).

Duration, content, and quality of speech material

[99.1020] Listeners are better at identifying speakers when more speech material is available. For example, Ladefoged & Ladefoged (1980) and Rose & Duncan (1995) reported correct-identification rates for familiar voices ranging between 31% for single words to 95% for multi-sentence stretches of speech (see also Bull & Clifford, 1999b, p. 198; Solan & Tiersma, 2003, pp. 397–399).

Although experimental results were not conclusive, Bull & Clifford (1999a, p. 217) suggested that exposure to variability in the speaker's voice rather than pure duration of speech may be important for increasing correct-identification rates (see also Bull & Clifford, 1984, pp. 105–106).

Loss of acoustic information due to poor recording or playback quality, or due to transmission of the voice via a telephone system, and mismatches between the quality of the known and questioned samples, may reduce correct-identification rates (Rose, 2002, p. 102).

> Because telephones distort voices, can they reduce the likelihood of voice identification compared to the direct hearing of a voice not involving the telephone? Our study of this (Rathbom, Bull and Clifford, 1981) suggests that they do. (Bull & Clifford, 1999b, p. 197)

See also Bull & Clifford (1984, pp. 114–116).

Prior expectations

[99.1030] If a listener expects to hear a particular voice or one of a restricted number of voices, then they are more likely to identify the voice they hear as the voice / one of the voices they expected to hear. It is common in experiments on familiar voice recognition for listeners to identify an unknown voice as the voice of a person they know (Rose, 2002, p. 104).

> the fact that many of the *identifications* made by ad hoc experts are contaminated by the circumstances in which the identification is made should not be overlooked. (Edmond & San Roque, 2009, p. 31).

Ladefoged & Ladefoged (1980, p. 47) described the case of *People v Kalkin*: Mr. Kalkin rented a hotel room. Narcotics agents phoned that hotel room and arranged a narcotics deal with the person who answered the phone. On the basis of his voice the narcotics agents identified the speaker as Mr. Kalkin. The defence were able to demonstrate the Mr. Kalkin had not been in the room at the time, and an associate of Mr. Kalkin admitted to being the person who had spoken to the narcotics agents. The narcotics agents appeared to have misidentified the voice because they had expected Mr. Kalkin to answer the phone.

Obviously there must be at least some degree of similarity between the expected voice and the voice heard (had a woman answered the phone the narcotics agents would have been unlikely to identify her as Mr. Kalkin), but prior expectation can have a powerful effect on a listener's identification. To illustrate with some extreme examples:

> (1) If a listener is 100% certain that the voice they will hear is the voice of a particular person, then whatever the properties of the voice they will still be 100% certain of their identification after hearing the voice – hearing the voice will have had no effect on their identification.

(2) Say the listener is unable to distinguish two brothers on the basis of their voices, as far as the listener is concerned they both sound the same; however, for other reasons the listener is 90% certain that the voice they will hear will be that of brother A rather than brother B – after hearing the voice they will still be 90% certain that it is the voice of brother A. Again hearing the voice will have had no effect on their identification.

In these extreme examples the listener's identification is dictated entirely by their prior expectations; usually the situation will be less extreme but prior expectations can still have a very large influence on the listener's identification of the voice. It is for this reason that forensic scientists using acoustic-phonetic and automatic approaches within the new paradigm base their forensic comparison on acoustic measurements rather than on listening, and do not make statements as to the probability that two voice recordings were made by the same speaker, but rather provide a strength-of-evidence statement as to the relative probability of observing the acoustic differences between the voice samples under the same-speaker proposition versus under the different-speaker proposition (the posterior probability can only be calculated if some prior probability is assumed and it will change if the value of the prior probability is changed, the calculation of a likelihood ratio does not include consideration of prior probabilities, see **[99.160]**).

Example

[99.1040] In *State of Western Australia v Cameron James Mansell* [WA Dist Ct, No 665 of 2008], a police officer listened to a series of telephone intercepts and was subsequently part of a team conducting a search of a suspect's office. She claimed that while conducting the search she heard the voice of someone talking with one of her colleagues and immediately recognised it as the same as the voice on the telephone intercepts. Audio recordings of the suspect talking during the search and on subsequent occasions were available and a technical forensic voice comparison would therefore have been potentially possible. The prosecution did not have a forensic voice comparison conducted but instead put the police officer on the stand to give her non-technical speaker identification. The defence called an expert witness (the author) to summarise research on the validity of non-technical speaker identification in general (not specifically related to the police officer's testimony).

Below, **[99.1050]**–**[99.1100]**, the factors listed above as being relevant to the validity of non-technical speaker identification, **[99.980]**–**[99.1030]**, are related to the police officer's written statements and oral testimony (only the oral testimony was presented to the jury).

Variability between listeners

[99.1050] Some listeners are good at identifying speakers from their voices, some listeners are poor. In theory it would be possible to have a listener participate in an experiment which would test their ability to identify speakers from their voices under similar conditions to those in the case.

No such tests were conducted and no evidence as to the validity of the police officer on this task was presented. It is therefore unknown whether the police officer is good, average, or poor at identifying speakers from their voices. The probability that her identification was correct or incorrect is

unknown. The validity of the police officer's identification of the questioned voice on the telephone intercepts as being the voice of the defendant is therefore unknown.

Listener certainty

[99.1060] In her written statements and oral testimony the police officer appeared to state with absolute certainty that she believed that the questioned voice was the voice of the defendant. She made definitive statements such as "I recognised this voice as the same voice I heard on the phone." (Statement of 17 September 2008, point 10; Statement of 3 April 2009, point 23). At no point did she add modifiers such as "I'm almost completely sure" or "I'm 95% certain". Even when given the opportunity to express less than 100% certainty she did not take it: Defence Attorney: "You simply believe that it *might* be Mansell?" Police Officer: "I *do* believe it is Mr. Mansell." (transcript of oral testimony, p. 189, emphasis added).

In line with the research findings on the relationship between a listener's certainty in their identification and their actual correct-identification rate, the police officer's certainty in her identification should not be equated with the actual probability of her identification being correct.

Listener's familiarity with speaker's voice

[99.1070] The police officer stated that she "attended the Telephone Intercept Unit daily to listen to the calls which had been intercepted" (Statement of 3 April 2009, point 12), that she "listened to all the phone calls as listed in Annex A" (Statement of 3 April 2009, point 13), and that "Whilst listening to the calls [she] became familiar with the accuseds [*sic*] voice." (Statement of 3 April 2009, point 10). Annex A listed 33 calls over a period of 7 days, estimated as having a total duration of around 40 minutes. In her oral testimony the police officer stated that she listened to each recording at least ten times (transcript of oral testimony, p. 135). She did not state why she listened to the recordings so many times, and did not say how much attention she paid to the characteristics of the voice or voices while listening.

This amount of exposure to a voice would fall far below the exposure necessary to make a voice highly familiar as would be the case for the voice of a relative or friend, at best it might be approximated with the low to moderate familiarity one might have with a media personality. This suggests that the validity of the police officer's identification of the questioned voice was likely to be substantially less than the relatively good (but not perfect) validity of the identification of highly familiar speakers such as relatives and friends, but substantially greater than a voice which had only previously been heard for a few seconds or minutes.

There is however a flaw in the logic above – it assumes that all of the telephone intercepts in question included recordings of the same questioned speaker. If in fact within this set of recordings there are recordings of two or more questioned speakers (whose voices are sufficiently similar that a listener may not realise that they come from two or more different speakers) then the police officer would have been familiarising herself with multiple voices erroneously assuming that they were the same voice. This would be detrimental to her ability to identify any one of these voices, and especially detrimental to her ability to distinguish these voices from each other.

When performing a forensic voice comparison a forensic scientist working within the new paradigm would not assume that the voices on different questioned-voice recordings belong to the same speaker (it would be inappropriate for a forensic scientist to make such an assumption), and they would instead separately compare the voice on each questioned-voice recording with the known voice.

Typicality of speaker's voice

[99.1080] The police officer stated that "The accused [*sic*] general [*sic*] spoke in a calm voice which was quieter and of a lower pitch than the other males that spoke." (Statement of 3 April 2009, point 11). Note that the police officer was actually referring to the questioned voice not the voice of the defendant..

Speaking in a calm voice, speaking quietly, and, for a male, speaking with a relatively low fundamental frequency are not unusual.

The police officer did not identify any properties of the questioned voice which would make it atypical with respect to the population in general, and when questioned on this agreed that there were no atypical features in either the questioned voice or the defendant's voice (transcript of oral testimony, pp. 184–185). Her identification of the questioned voice as the same as the known voice was therefore likely to be of poorer validity than if she had noted properties of the questioned voice which would make it atypical.

If a forensic scientist working in the new paradigm were to perform a forensic voice comparison, they would statistically quantify the typicality of the acoustic properties of the questioned and known voices with respect to the relevant population (estimated from a sample of audio recordings of a large number of voices from the relevant population).

Duration, content, and quality of speech material

[99.1090] The recordings of the questioned speaker were made via intercepts of a mobile telephone. The quality of the transmission of speech via mobile telephone systems is often relatively poor and the recording quality of telephone intercepts is also often poor. In addition there was a mismatch between the quality of the recording of the questioned voice heard by the police officer and her hearing of the known voice which was in the physical presence of the speaker. These factors are likely to make the police officer's identification less valid than if she had heard both questioned and known voices live or as high quality recordings.

The duration of time over which the police officer was exposed to the known voice was very short because she made her identification immediately on hearing the voice: "Prior to entering an office I heard the accused [*sic*] voice talking to [my colleague]. I recognised this voice as the same voice I heard on the phone."(Statement of 3 April 2009, points 22–23). "the voice I heard, before I entered that room, I believed to be the same as the same voice that I heard on those phones."(transcript of oral testimony, p. 183).

Her identification was therefore likely to be less valid than if she had spent more time comparing the properties of the known and questioned voices before coming to her definitive conclusion.

As discussed below, however, such an immediate identification could have been influenced by a prior expectation bias which is not likely to be remedied by hearing the known voice for a longer duration of time or hearing it on subsequent occasions (Solan & Tiersma, 2003, p. 391).

> A witness may become more confident through repeated exposure without any corresponding improvement in accuracy. (Edmond & San Roque, 2009, p. 31)
>
> The opinions of investigating police are not sanitised through repeatedly listening to tapes or repeatedly observing incriminating images. In reality, this may introduce confirmation bias, contaminate the evidence and endanger the accused. Repeated listening or watching alone should not provide grounds for the admission of identification evidence or evidence of similarity. (Edmond & San Roque, 2009, p. 22).

Prior expectations

[99.1100] The police officer stated that she assisted at the execution of a search warrant, and that "Prior to entering an office [she] heard the accused [*sic*] voice talking to [her colleague, and] recognised this voice as the same voice [she] heard on the phone."(Statement of 3 April 2009, points 22–23). In her oral testimony, she stated that prior to the execution of the search warrants she knew that Mr. Mansell was the target of the investigation, that she believed that Mr. Mansell was the speaker on the telephone intercepts, and that she knew she would be searching the office and home of Mr. Mansell (transcript of oral testimony, pp. 183–184).

It would seem reasonable to assume that when a police officer executes a search warrant they have a high expectation of encountering a suspect, and, if they have already been working on a particular case, have a high expectation of encountering a particular suspect. The key issue is the prior expectation that a person encountered at the scene of a search will be the suspect, rather than the expectation that the suspect will be at that particular location at that particular time. It may therefore be reasonable to assume that before arriving at Mr. Mansell's office the police officer had a high expectation of encountering the owner of the voice which she had heard on the telephone intercept recordings. If this is the case, then the fact that she immediately identified Mr. Mansell's voice as the same as the questioned voice will have been heavily influenced by her prior expectation that she would hear the same voice.

There is ample evidence in the police officer's written statements to suggest that she did have a bias towards identifying Mr. Mansell as the speaker of the questioned voice: She continually referred to the questioned voice on the telephone intercept recordings as the voice of the accused, e.g., "These telephone calls included conversations between [third party] and the accused." (Statement of 17 September 2008, point 5). By listening to the recordings "I became familiar with the voice of a male person who I now know to be the accused, Cameron James MANSELL." (Statement of 17 September 2008, point 6). "Whilst listening to the calls I became familiar with the accuseds [*sic*] voice." (Statement of 3 April 2009, point 10). "The accused general [*sic*] spoke in a calm voice which was quieter and of a lower pitch than the other males that spoke." (Statement of 3 April 2009, point 11). Annex A also inappropriately identified the person making and receiving the intercepted telephone calls as Mr. Mansell. In her oral testimony, the police officer affirmed that prior to making her speaker identification, she already believed the voice on the telephone intercepts to be the voice of Mr. Mansell (transcript of oral testimony, p. 186).

Solan & Tiersma (2003, pp. 381–382) summarise a number of similar cases in the United States, for example an earwitness is asked to come to the police station and upon arrival hears a single suspect being interviewed by a detective and identifies the voice of the suspect as the same as the offender whom they had heard earlier. A number of US courts have found such identifications to be overly tainted by suggestion (prior expectations instilled in the earwitness by the police), and have ruled them inadmissible. Illogically, US courts have typically allowed such suggestively-tainted identifications when made by a police officer presented with an audio recording of the voice of a single suspect (Solan & Tiersma, 2003, pp. 388–393).

Outcome

[99.1110] The police officer was either correct or incorrect in her identification of the voice of Mr. Mansell as the same as the voice she had heard on the telephone intercept recordings. The discussion above merely provides a guide to factors which should be considered in coming to a subjective opinion as to the likelihood that the police officer's identification was correct. It is also important to note that other evidence and legal arguments were presented in court and it is not known how much weight the jury assigned to the non-technical-speaker-identification evidence. One should also heed Evett's (2009) advice that success in forensic science should not be linked to verdicts, and that the forensic scientist should be neutral, not a partisan of either the prosecution or the defence. These caveats with respect to the interpretation of the influence of the expert testimony on verdict should be borne in mind; however, the curious reader may wish to know that the jury found the defendant not guilty.

I have no opinion as to whether the voice on the telephone intercepts was or was not that of the accused, but I believe some good may arise if the verdict in this case leads to prosecutors making greater use of the services of trained forensic scientists whose forensic-voice-comparison systems produce logically-correct strength-of-evidence statements of demonstrated degrees of validity and reliability, rather than relying on lay persons whose statements are of unknown validity and reliability and potentially tainted by prior expectations.

APPENDIX A: INTERNATIONAL PHONETIC ALPHABET

[99.1150] A reproduction of the International Phonetic Alphabet (2005) appears on the following page.

The International Phonetic Alphabet may be freely copied on condition that acknowledgement is made to the International Phonetic Association (Department of Theoretical and Applied Linguistics, School of English, Aristotle University of Thessaloniki, Thessaloniki 54124, GREECE). http://www.langsci.ucl.ac.uk/ipa/

THE INTERNATIONAL PHONETIC ALPHABET (revised to 2005)

CONSONANTS (PULMONIC)

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

CONSONANTS (NON-PULMONIC)

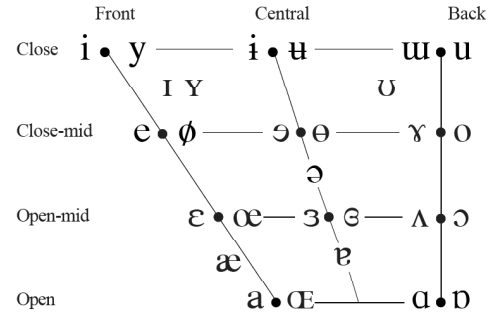
Clicks		Voiced implosives		Ejectives	
ʘ	Bilabial	ɓ	Bilabial	’	Examples:
ǀ	Dental	ɗ	Dental/alveolar	pʼ	Bilabial
ǃ	(Post)alveolar	ʄ	Palatal	tʼ	Dental/alveolar
ǂ	Palatoalveolar	ɠ	Velar	kʼ	Velar
ǁ	Alveolar lateral	ʛ	Uvular	sʼ	Alveolar fricative

OTHER SYMBOLS

ʍ	Voiceless labial-velar fricative	ɕ ʑ	Alveolo-palatal fricatives
w	Voiced labial-velar approximant	ɺ	Voiced alveolar lateral flap
ɥ	Voiced labial-palatal approximant	ʢ	Simultaneous ʃ and x
ʜ	Voiceless epiglottal fricative		
ʡ	Voiced epiglottal fricative		
ʡ	Epiglottal plosive		

Affricates and double articulations can be represented by two symbols joined by a tie bar if necessary. k͡p t͡s

VOWELS



Where symbols appear in pairs, the one to the right represents a rounded vowel.

SUPRASEGMENTALS

ˈ	Primary stress		ˌfoʊnəˈtɪʃən
ˌ	Secondary stress		
ː	Long	eː	
ˑ	Half-long	eˑ	
˘	Extra-short	ĕ	
|	Minor (foot) group		
‖	Major (intonation) group		
.	Syllable break	ɹi.ækt	
‿	Linking (absence of a break)		

DIACRITICS Diacritics may be placed above a symbol with a descender, e.g. n̥

◌̥	Voiceless	n̥ d̥	◌̤	Breathy voiced	b̤ a̤	◌̪	Dental	t̪ d̪
◌̬	Voiced	s̬ t̬	◌̰	Creaky voiced	b̰ a̰	◌̺	Apical	t̺ d̺
◌ʰ	Aspirated	tʰ dʰ	◌̼	Linguolabial	t̼ d̼	◌̻	Laminal	t̻ d̻
◌̹	More rounded	ɔ̹	◌ʷ	Labialized	tʷ dʷ	◌̃	Nasalized	ẽ
◌̜	Less rounded	ɔ̜	◌ʲ	Palatalized	tʲ dʲ	◌ⁿ	Nasal release	dⁿ
◌̟	Advanced	u̟	◌ˠ	Velarized	tˠ dˠ	◌ˡ	Lateral release	dˡ
◌̠	Retracted	e̠	◌ˤ	Pharyngealized	tˤ dˤ	◌̚	No audible release	d̚
◌̈	Centralized	ë	◌̴	Velarized or pharyngealized	ɫ			
◌̽	Mid-centralized	e̽	◌̝	Raised	e̝	(ɹ̝ = voiced alveolar fricative)		
◌̩	Syllabic	n̩	◌̞	Lowered	e̞	(β̞ = voiced bilabial approximant)		
◌̯	Non-syllabic	e̯	◌̘	Advanced Tongue Root	e̘			
◌˞	Rhoticity	ɚ a˞	◌̙	Retracted Tongue Root	e̙			

TONES AND WORD ACCENTS

LEVEL			CONTOUR	
e̋ or ˥	Extra high		ě or ˩˥	Rising
é ˦	High		ê ˥˩	Falling
ē ˧	Mid		e᷄ ˧˥	High rising
è ˨	Low		e᷅ ˩˧	Low rising
ȅ ˩	Extra low		e᷈ ˥˩˥	Rising-falling
↓	Downstep		↗	Global rise
↑	Upstep		↘	Global fall

APPENDIX B: TRAINING AND ASSOCIATIONS

Training

[99.1190] Until recently there have been no formal programmes for training forensic scientists specialising in forensic voice comparison. Rather, individuals have usually obtained advanced degrees specialising in acoustic phonetics or speech processing and have received additional training in forensic voice comparison by working with someone who is already an expert in this area, often as part of doctoral or postdoctoral research.

That the validity and reliability of a forensic voice comparison system has been tested (including decisions and actions taken by a forensic scientist as part of the design or operation of the system) is more important than the qualifications of the forensic scientist, but one might expect a forensic scientist specialising in forensic voice comparison to have a doctorate in a relevant technical area and training in forensic science. Knowledge of forensic science is essential, having a doctorate in acoustic phonetics or speech processing alone is not a sufficient qualification for performing forensic voice comparison.

I know of two Masters degrees providing training in forensic voice comparison:

> The Master of Science Programme in Forensic Speech Science, Department of Language and Linguistic Science, University of York includes training in auditory-acoustic-phonetic forensic voice comparison. http://www.york.ac.uk/depts/lang/prospective/postgrad/forensic/
>
> The Judicial Phonetics Specialisation of the Master in Phonetics and Phonology Programme of the Consejo Superior de Investigaciones Científicas [Spanish National Research Council] / Universidad Internacional Menéndez Pelayo includes training in acoustic-phonetic and automatic forensic voice comparison, and training on the evaluation of forensic evidence within the new paradigm. http://www.estudiosfonicos.cchs.csic.es/

The quality of formal training programmes in forensic voice comparison should be assessed according to whether they produce graduates who are proficient and knowledgeable practitioners of forensic voice comparison conducted within the new paradigm.

Associations

[99.1200] As is the case with training, that the validity and reliability of a forensic voice comparison system has been tested (including decisions and actions taken by a forensic scientist as part of the design or operation of the system) is more important than the membership of the forensic scientist in any association, but one might expect a forensic scientist specialising in forensic voice comparison to be a member of relevant forensic science associations and to follow their codes of ethics and practice.

As far as I am aware, the only professional association focussing on forensic voice comparison and currently active is the *International Association for Forensic Phonetics and Acoustics* (IAFPA) http://www.iafpa.net/.

Professional associations covering forensic science in general in English-speaking countries include:

- *The American Academy of Forensic Sciences* (AAFS) http://www.aafs.org/
- *The Association of Forensic Science Providers* (AFSP), UK and Republic of Ireland. An association of forensic laboratories rather then of individual forensic scientists.
- *The Australian and New Zealand Forensic Science Society* (ANZFSS) http://www.anzfss.org.au/
- *The Canadian Society of Forensic Sciences* (CSFS) http://www.csfs.ca/
- *The Forensic Science Society* (FSSoc), UK http://www.forensic-science-society.org.uk/

In addition:

- *The Audio Engineering Society* (AES) http://www.aes.org/ has an Audio Forensics Technical Committee with forensic voice comparison included in their mandate.
- *The Australasian Speech Science and Technology Association* (ASSTA) http://www.assta.org/ has a Forensic Speech Science Committee (FSSC).
- *The European Network of Forensic Science Institutes* (ENFSI) http://www.enfsi.eu/ has a Forensic Speech and Audio Analysis Working Group (FSAAWG).

Abbreviations

ABRE	American Board of Recorded Evidence
AFSP	Association of Forensic Science Providers
bit	binary digit
CAI	Case Assessment and Interpretation
C_{llr}	log-likelihood-ratio cost
DNA	deoxyribonucleic acid
E	evidence (measured differences between the samples of known and questioned origin
f0	fundamental frequency
F1	first formant
F2	second formant
F3	third formant
FBI	Federal Bureau of Investigation
ff	following
g	gramme(s)
GMM	Gaussian mixture model
H_{do}	different-origin hypothesis
H_{so}	same-origin hypothesis
Hz	hertz
IAFPA	International Association for Forensic Phonetics and Acoustics
IAI	International Association for Identification
IAVI	International Association of Voice Identification
ID	identification
IPA	International Phonetic Association
k	kilo-
log	logarithm
$\log_{10}$	log base ten
$\log_2$	log base two
LR	likelihood ratio
LR_{ds}	likelihood ratios derived from different-speaker comparisons
LR_{ss}	likelihood ratios derived from same-speaker comparisons

MFCC	mel-frequency cepstral coefficients
ms	millisecond(s)
N_{ds}	number of different-speaker comparisons
NRC	National Research Council
N_{ss}	number of same-speaker comparisons
p	probability
p.	page
PhD	Doctor of philosophy
pp.	pages
s	second(s)
UK	United Kingdom
US	United States
v	versus

Glossary

accuracy — extent to which a measurement or estimate approximates the true value **[99.290]**

activity level — a level in the Case Assessment and Interpretation (CAI) model **[99.140]**

alveolar ridge — part of the vocal tract near the front of the roof of the mouth **[99.480]**

anti-resonance — frequency at which a sound is reduced in amplitude by a side-resonator such as part of the vocal tract **[99.480]**

approach — methodology for extracting information from voice samples **[99.650]**

arytenoid cartilages — part of the larynx **[99.540]**

aspiration — turbulent airflow between vocal folds at the beginning off a plosive **[99.520]**

background database — a sample taken from a large number of members of the relevant population **[99.180]**

background model — model of the distribution of properties of the background database **[99.230]**

bandpass — range of frequencies which are transmitted by a transmission system **[99.610]**

basis function — a simple function, a series of which are combined to build a complex function **[99.800]**

Bayes' Theorem — statement of the logical relationship between beliefs about the hypotheses before and after the presentation of the strength of evidence **[99.160]**

bilabial — speech sound made with a closure or constriction of the lips **[99.520]**

bit rate — multiplication of sampling frequency and word size, representing the amount of information used to encode a signal **[99.600]**

blade — part of the front of the tongue **[99.480]**

breathy voicing — voicing combined with turbulent airflow **[99.540]**

broad transcription — written representation of speech sounds providing relatively little detail **[99.460]**

calibration — a procedure for converting scores to likelihood ratios **[99.240]**

channel effect — changes to a signal introduced by recording or transmission conditions **[99.600]**

clipping — truncation of the high-amplitude portions of a signal **[99.600]**

closed quotient — portion of a vocal fold vibration during which the vocal folds are closed **[99.540]**

coarticulation — overlap of articulation (pronunciation) of one speech sound with the articulation of earlier or later speech sounds **[99.480]**

codec — signal compression and decompression algorithm **[99.610]**

coefficient estimate — the value calculated when fitting a model to data **[99.800]**

constriction — a narrowing in a tube such as a vocal tract **[99.460]**

creaky voicing — irregular low-frequency voicing **[99.540]**

cross validation — a procedure which allows a single database to be used in place of two databases **[99.810]**

Daubert — US Supreme Court ruling establishing criteria for the admissibility of expert evidence **[99.80]**

defence attorney's fallacy — error in the interpretation of a likelihood ratio **[99.390]**

diacritic — small symbol used to modify a phonetic symbol and give a more detailed transcription of a speech sound **[99.460]**

diphthong — a vowel with substantial formant movement over its duration **[99.460]**

discrete cosine transform — numeric description of the shape of a complex curve **[99.800]**

dorsum — part of the back of the tongue **[99.480]**

earwitness — a person who is present at the scene of a crime, hears the offender speaking, and either immediately recognises the offender's voice as belonging to a particular person they already know, or who later attempts to pick the speaker out of a voice lineup **[99.960]**

error rate — proportion of test results which exceed a specified threshold **[99.300]**

formant transitions — changes in formants as the vocal tract changes from the shape needed to make a consonant to the shape needed to make a vowel or vice versa **[99.520]**

formant — a resonance frequency of a vocal tract **[99.460]**

framework — procedure for evaluating evidence **[99.650]**

fricative — speech sound made with turbulent airflow **[99.500]**

fundamental frequency — the rate at which the vocal folds vibrate during voicing **[99.540]**

fusion — a procedure for combining multiple parallel sets of scores or likelihood ratios **[99.240]**

Gaussian distribution — a simple probability density function **[99.220]**

Gaussian mixture model — a complex probability density function **[99.230]**

harmonic — a multiple of the fundamental frequency **[99.610]**

histogram — a model of the distribution of a discrete variable **[99.210]**

idiolect — pronunciation peculiarities of an individual, finer grained than the peculiarities of a dialect **[99.470]**

intonation — a long-term fundamental-frequency pattern which signals linguistic (or paralinguistic) information such as whether the utterance is a question versus a declaration **[99.540]**

jitter — degree to which the duration of individual voicing cycles vary **[99.540]**

larynx — structure at the bottom of the vocal tract which includes the vocal folds **[99.450]**

likelihood-ratio framework — the logically correct framework for the evaluation of forensic-comparison evidence **[99.140]**

likelihood ratio (*LR*) — the numeric expression of the strength of evidence **[99.150]**

log likelihood ratio — likelihood ratio expressed on a logarithmic scale **[99.300]**

log-likelihood-ratio cost (C_{llr}) — a measure of the accuracy of a forensic-comparison system **[99.300]**

logistic regression — a method for performing calibration and fusion **[99.240]**

mel-frequency cepstral coefficients (MFCC) — numeric values encoding measurements of spectra **[99.720]**

monophthong — a vowel with negligible formant movement over its duration **[99.460]**

naïve Bayes — a procedure for combining likelihood ratios which does not take correlation into account **[99.390]**

narrow transcription — a relatively detailed written representation of speech sounds **[99.460]**

nasal cavities — internal structure of the nose, part of the vocal tract **[99.450]**

nasal — speech sound produced with voicing, a closure in the oral cavity, and an open the velopharyngeal port such that air flows through the nasal cavities **[99.480]**

nasalised vowel — vowel made with the velopharyngeal port open and an open oral cavity such that air flows through both the oral and nasal cavities **[99.480]**

nasopharyngeal tube — pharyngeal cavity plus nasal cavity **[99.480]**

non-stressed vowel — relatively short vowel with some degree of neutralisation of formant values **[99.460]**

non-technical speaker identification — the general ability of a person with no training in forensic voice comparison to recognise a voice and identify the speaker **[99.930]**

offence level — a level in the Case Assessment and Interpretation (CAI) model **[99.140]**

open quotient — portion of a vocal fold vibration during which the vocal folds are open **[99.540]**

oral cavity — mouth, part of the vocal tract **[99.450]**

oropharyngeal tube — pharyngeal cavity plus oral cavity **[99.450]**

packet — portion of a signal transmitted as a unit **[99.610]**

paradigm shift — a change in the culture of science **[99.70]**

pharyngeal cavity — throat, part of the vocal tract **[99.450]**

phoneme — a speech sound which contrasts with other speech sounds in a given language or dialect **[99.460]**

phonetic symbol — symbol used to transcribe a speech sound **[99.460]**

phonetics — the study of the physical aspects of the production, transmission, and perception of human speech **[99.440]**

plosive — speech sound made by creating complete oral and velopharyngeal-port closure in the vocal tract, compressing the lungs so as to increase air pressure in the vocal tract, then releasing the pressure by rapidly opening the oral closure **[99.520]**

precision — extent to which multiple measurements or estimates of the same value differ from each other **[99.290]**

probability density function — a model of the distribution of one or more continuous variables **[99.220]**

prosecutor's fallacy — error in the interpretation of a likelihood ratio **[99.380]**

reliability — extent to which multiple measurements or estimates of the same value differ from each other **[99.290]**

resonance frequency — frequency at which a sound is amplified by a resonator such as a vocal tract **[99.460]**

rounded lips — lips held in a configuration such as when saying the vowel sound of "who" **[99.460]**

sampling frequency — number of times per second a measurement is taken when digitising a signal **[9960.0]**

score — a value which quantifies the degree of similarity of samples of known and questioned origin including consideration of their typicality with respect to the relevant population, but which is not interpretable as a likelihood ratio **[99.240]**

shimmer — degree to which the amplitude of voicing varies across cycles **[99.540]**

similarity — the extent to which the properties of samples of known and questioned origin do not differ from each other **[99.150]**

source level — a level in the Case Assessment and Interpretation (CAI) model **[99.140]**

spectrogram — a graphical representation of frequency and its changes over time of, for example, a speech utterance **[99.460]**

spectrum — the frequency properties of, for example, a speech sound at a point in time **[99.460]**

spread lips — lips held in a smiling-like configuration **[99.460]**

strength of evidence — the value of the evidence with respect to the competing same-origin and different-origin hypotheses **[99.150]**

stressed vowel — relatively long vowel with well defined formant values **[99.460]**

suspect model — model of the distribution of properties of the suspect sample(s) **[9923.0]**

tip — part of the front of the tongue **[99.480]**

Tippett plot — a graphical representation of the results of a test of a forensic-comparison system **[99.330]**

tone — a fundamental-frequency pattern which distinguishes one phoneme from other phonemes **[99.540]**

Tosi Extrapolation — assertion that performance will be better under real-world conditions than under controlled laboratory-test conditions **[99.680]**

transposition of conditionals — error in the interpretation of a likelihood ratio, commonly known as the prosecutor's fallacy **[99.380]**

turbulent airflow — irregular movement of air which causes a noise **[99.500]**

typicality — the extent to which the properties of samples of known and questioned origin do not differ from those of the relevant population **[99.150]**

validity — extent to which a measurement or estimate approximates the true value **[99.290]**

velopharyngeal port — opening between the pharyngeal and nasal cavities **[99.450]**

velum — soft palate, part of the vocal tract **[99.450]**

vocal tract — the mouth, throat, and nose used to make speech sounds **[99.450]**

voice-onset time (VOT) — time between release of the closure of a plosive and the beginning of voicing **[99.520]**

voiced — a sound make with vibrating vocal folds **[99.460]**

voicegram identification — spectrographic approach **[99.680]**

voiceless — a sound make without vibrating vocal folds **[99.460]**

voiceprinting — spectrographic approach **[99.680]**

word size — number of binary digits (bits) used to encode the amplitude of a signal **[99.600]**

References

Aitken C.G.G., & Taroni F., 2004. *Statistics and the evaluation of forensic evidence for forensic scientist*. 2nd ed. Chichester, UK: Wiley.

Alexander,, A., Dessimoz D., Botti, F., & Drygajlo A., 2005. Aural and automatic forensic speaker recognition in mismatched conditions. *International Journal of Speech, Language and the Law*, 12, pp. 214–234.

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

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

Balding D.J., 2005. *Weight-of-evidence for forensic DNA profiles*. Chichester, UK: Wiley.

Becker T., Jessen,M., & Grigoras C., 2008. Forensic speaker verification using formant features and Gaussian mixture models. In: *Proceedings of Interspeech 2008 Incorporating SST 2008*. International Speech Communication Association, pp. 1505–1508.

Becker T., Jessen M., & Grigoras C., 2009. Speaker verification based on formants using Gaussian mixture models. In: Proceedings of NAG/DAGA International Conference on Acoustics, Rotterdam.

Brümmer N., Burget L., Cernocký J.H., Glembek O., Grézl F., Karafiát M., van Leeuwen D.A., Matejka P., Schwarz P., & Strasheim A., 2007. Fusion of heterogenous speaker recognition systems in the STBU submission for the NIST SRE 2006. *IEEE Transactions on Audio, Speech, and Language Processing*, 15, pp. 2072–2084. doi:10.1109/TASL.2007.902870

Brümmer N., & du Preez J., 2006. Application independent evaluation of speaker detection. *Computer Speech and Language*, 20, pp. 230–275. doi:10.1016/j.csl.2005.08.001

Buckleton J., 2005. A framework for interpreting evidence. In: Buckleton J., Triggs C.M., & Walsh S.J. eds. *Forensic DNA evidence interpretation*. Boca Raton (FL): CRC, pp. 27–63.

Buckleton J., Triggs C.M., & Walsh S.J. eds., 2005. Forensic DNA evidence interpretation. Boca Raton (FL): CRC.

Bull R., & Clifford B., 1984. Earwitness voice recognition accuracy. In Wells G.L., & Loftus E.F. eds. *Eyewitness testimony*. Cambridge (UK): Cambridge University Press, pp. 92–123.

Bull R., & Clifford B., 1999a. *Earwitness testimony*. *New Law Journal Expert Witness Supplement*, February 12, pp. 216–220. [The identical text was also published in: *Medicine, Science and the Law*, 39, pp. 120–127.]

Bull R., & Clifford B., 1999b. Earwitness testimony. In Heaton-Armstrong A., Shepherd E., & Wolchover D. eds. *Analysing witness testimony: A guide for legal practitioners and other professionals*. London: Blackstone Press, pp. 194–206.

Cambier-Langevald T., 2007. Current methods in forensic speaker identification: Results of a collaborative exercise. *International Journal of Speech, Language and the Law*, 14, pp. 223–243. doi:10.1558/ijsll.2007.14.2.223

Champod C., & Meuwly D., 1998. The inference of identity in forensic speaker recognition.. In: *Proceedings of RLA2C Workshop: Speaker Recognition and its Commercial and Forensic Applications*, pp. 125–135. doi:10.1016/S0167-6393(99)00078-3

Clark J., Yallop C., & Fletcher J., 2007. *An introduction to phonetics and phonology*. 3rd ed. Oxford: Blackwell.

Cole S.A., 2006. Is fingerprint identification valid: Rhetorics of reliability in fingerprint proponents' discourse. *Law & Policy*, 28, pp. 109–135. doi:10.1111/j.1467-9930.2005.00219.x

Cole S.A., 2009. Forensics without uniqueness, conclusions without individualization: The new epistemology of forensic identification. *Law, Probability and Risk*, 8, pp. 233–255. doi:10.1093/lpr/mgp016

Cole S.A., 2010. Who speaks for science? A response to the National Academy of Sciences Report on forensic science . *Law, Probability and Risk*, 9, pp. 25–46. doi:10.1093/lpr/mgp032

Cook R., Evett I.W., Jackson G., Jones P.J., & Lambert J.A., 1998a. A hierarchy of propositions: deciding which level to address in casework. *Science & Justice*, 38, pp. 231–239. doi:10.1016/S1355-0306(98)72117-3

Cook R., Evett I.W., Jackson G., Jones P.J., & Lambert J.A., 1998b. A model for case assessment and interpretation. *Science & Justice*, 38, pp. 151–156. doi:10.1016/S1355-0306(98)72099-4

Curran J.M., 2005. An introduction to Bayesian credible intervals for sampling error in DNA profiles. *Law, probability and Risk*, 4, 115–126. doi:10.1093/lpr/mgi009

Donnelly P., 2005. Appealing statistics. *Significance*, 2, pp. 46–48. doi:10.1111/j.1740-9713.2005.00089.x

Edmond G., Kemp R., Porter G., Hamer D., Burton M., Biber K., & San Roque M., 2010. Atkins v The Emperor: The 'cautious' use of unreliable 'expert' opinion. *International Journal of Evidence and Proof*, 14, pp. 146–166. doi:10.1350/ijep.2010.14.2.349

Edmond G., & San Roque M., 2009. Quasi-justice: Ad hoc expertise and identification evidence. *Criminal Law Journal*, 33, pp. 8–33.

Elliot J.R., 2002. *Okay, what are the odds?* Master's thesis. Canberra: Australian National University.

Evett I.W., 1998. Towards a uniform framework for reporting opinions in forensic science case-work. *Science & Justice*, 38, pp. 198–202. doi:10.1016/S1355-0306(98)72105-7

Evett I.W., 1991. Interpretation: A personal odyssey. In: Aitken C.G.G., & Stoney D.A. eds. *The use of statistics in forensic science*. Chichester (UK): Ellis Horwood, pp. 9–22.

Evett I.W., 2009. Evaluation and professionalism. *Science & Justice*, 49, pp. 159–160. doi:10.1016/j.scijus.2009.07.001

Evett I.W., & Buckleton J.S., 1996. Statistical analysis of STR data. In: Carraredo A., Brinkmann B., & Bär W., eds. *Advances in forensic haemogenetics*. Heidelberg: Springer-Verlag, vol. 6 pp. 79–86.

Faigman D.L., Saks M.J., Sanders J., & Cheng E.K. eds. 2008. Modern *Scientific Evidence: The Law and Science of Expert Testimony*. Eagan (MN): Thomson Reuters/West.

Foreman L.A., Champod C., Evett I.W., Lambert J.A., & Pope S., 2003. Interpreting DNA evidence: A review. *International Statistical Review*, 71, pp. 473–495.

Found B., & Rogers D., 2008. The probative character of Forensic Handwriting Examiners' identification and elimination opinions on questioned signatures. *Forensic Science International*, 178, pp. 54–60. doi:10.1016/j.forsciint.2008.02.001

Foulkes P., & Barron A., 2000. Telephone speaker recognition amongst members of a close social network. *Forensic Linguistics*, 7, pp. 180–198.

French J.P., & 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, pp. 137–144. doi:10.1558/ijsll.v14i1.137

French J.P., Nolan, F., Foulkes, P., Harrison P., & 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, pp. 143–152. doi:10.1558/ijsll.v17i1.143

González-Rodríguez J., Drygajlo A., Ramos-Castro D., García-Gomar M., & Ortega-García, J., 2006. Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. *Computer Speech and Language*, 20, pp. 331–355. doi:10.1016/j.csl.2005.08.005

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

Gruber J.S., & Poza F., 1995. Voicegram Identification Evidence. In: *American Jurisprudence Trials*. Westlaw. Vol. 54.

Guillemin B.J., & Watson C., 2008. Impact of the GSM mobile phone network on the speech signal: Some preliminary findings. *International Journal of Speech Language and the Law*, 15, pp. 193–218. doi : 10.1558/ijsll.v15i2.193

Hollien H., 1990. *The acoustics of crime*. New York: Plenum

Hollien H., 2002. *Forensic voice identification*. San Diego: Academic.

Jessen M., 2008. Forensic phonetics. *Language and Linguistics Compass*, 2, pp. 671–711. doi:10.1111/j.1749-818x.2008.00066.x

Johnson K., 2003. *Acoustic and auditory phonetics*. 2nd ed. Maldon (MA): Blackwell.

Joos M., 1948. *Acoustic phonetics*. Language Monograph No. 23. Supplement to *Language*, 24(2).

Kaye D.H., 2010. Probability, individualization, and uniqueness in forensic science evidence: Listening to the academies. *Brooklyn Law Review*, 75(4).

Kaye D.H., & Sensabaugh Jr. G.F. ,2008. DNA typing: II. Scientific status. In: Faigman D.L., Saks M.J., Sanders J., & Cheng E.K. eds. *Modern Scientific Evidence: The Law and Science of Expert Testimony*. Eagan (MN): Thomson Reuters/West, §30:21–30:58.

Kersta L.G., 1962. Voiceprint identification. *Nature*, 196, pp. 1253–1257. doi:10.1038/1961253a0

Kinnunen T., & Li H., 2010. An overview of text-independent speaker recognition: From features to supervectors. Speech Communication, 52, 12–40. doi:10.1016/j.specom.2009.08.009

Kinoshita Y., Ishihara S., & Rose P., 2009. Exploring the discriminatory potential of F0 distribution parameters in traditional forensic speaker recognition. *International Journal of Speech, Language and the Law*, 16, pp. 91–111. doi:10.1558/ijsll.v16i1.91

Kinoshita Y., & Osanai T. (2006). Within speaker variation in diphthongal dynamics: What can we compare? In: Warren P., & Watson C.I. eds. Proceedings of the 11th Australasian International Conference on Speech Science & Technology, Auckland, New Zealand. Canberra: Australasian Speech Science & Technology Association, pp. 112–117.

Kirkland J.A., 2003. *Forensic speaker identification using Australian English* fuken*: A Bayesian likelihood ratio-based auditory and acoustic phonetic investigation*. Bachelor's thesis. Canberra: Australian National University.

Koehler J.J., 2010. Forensic science reform in the 21st century: A major conference, a blockbuster report and reasons to be pessimistic. *Law Probability and Risk*, 9, pp. 1–6. doi:10.1093/lpr/mgp029

Koenig B.E., 2002. Review of Hollien (2002) Forensic voice identification. *Journal of Forensic Identification*, 52, pp.762–766.

Kuhn T.S., 1962. *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Künzel H.J., 2001. Beware of the 'telephone effect': The influence of telephone transmission on the measurement of formant frequencies. *Forensic Linguistics*, 8, pp. 80–99.

Ladefoged J., & Ladefoged P., 1980. The ability of listeners to identify voices. *UCLA Working Papers in Phonetics*, 49, pp. 43–51.

Ladefoged P., 2001. *Vowels and consonants: An introduction to the sounds of language*. Malden (MA): Blackwell.

Ladefoged P., 2006. *A course in phonetics*. Boston: Thomson Wadsworth.

Law Commission of England & Wales, 2009. *The admissibility of expert evidence in criminal proceedings in England and Wales: A new approach to the determination of evidentiary reliability*. London. Available at: http://www.lawcom.gov.uk/expert_evidence.htm [Accessed April 2009]

Lucy D., 2005. *Introduction to statistics for forensic scientists*. Chichester (UK): Wiley.

Meuwly D., 2001. *Reconnaissance de locuteurs en sciences forensiques: l'apport d'une approche automatique*. PhD dissertation. University of Lausanne.

Meuwly D., 2006. Forensic individualisation from biometric data. *Science & Justice*, 46, pp. 205–213. doi:10.1016/S1355-0306(06)71600-8

Meuwly D., & Drygajlo A., 2001. Forensic speaker recognition based on a Bayesian framework and Gaussian mixture modelling (GMM). In: *Proceedings of 2001: A Speaker Odyssey. The Speaker Recognition Workshop*. International Speech Communication Association.

Morrison G.S., 2008. Forensic voice comparison using likelihood ratios based on polynomial curves fitted to the formant trajectories of Australian English /aɪ/. *International Journal of Speech, Language and the Law*, 15, pp. 249–266. doi:10.1558/ijsll.v15i2.249

Morrison G.S., 2009a. Comments on Coulthard & Johnson's (2007) portrayal of the likelihood-ratio framework. *Australian Journal of Forensic Sciences*, 41, pp. 155–161. doi:10.1080/00450610903147701

Morrison G.S., 2009b. Forensic voice comparison and the paradigm shift. *Science & Justice*, 49, pp. 298–308. doi:10.1016/j.scijus.2009.09.002

Morrison G.S., 2009c. Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs. *Journal of the Acoustical Society of America*, 125, pp. 2387–2397. doi:10.1121/1.3081384

Morrison G. S. (2010). A comparison of procedures for the calculation of forensic likelihood ratios from acoustic-phonetic data: Multvariate kernel density (MVKD) versus Gaussian mixture model – universal background model (GMM-UBM). Manuscript submitted for publication.

Morrison G. S. (2011). Theories of vowel inherent spectral change: A review. In: Morrison G.S., & Assmann P.F., *Vowel inherent spectral change*. Heidleberg, Germany: Springer-Verlag.

Morrison G.S., & Kinoshita Y., 2008. Automatic-type calibration of traditionally derived likelihood ratios: Forensic analysis of Australian English /o/ formant trajectories. In: *Proceedings of Interspeech 2008 Incorporating SST 2008*. International Speech Communication Association, pp. 1501–1504.

Morrison G.S., Thiruvaran T., & Epps J., 2010a. An issue in the calculation of logistic-regressoin calibration and fusion weights for forensic voice comparison. *Proceedings of the 13th Australasian International Conference on Speech Science and Technology*, Melbourne, Australia.

Morrison G.S., Thiruvaran T., & Epps J., 2010b. Estimating the precision of the likelihood-ratio output of a forensic-voice-comparison system. *Proceedings of Odyssey 2010: The Language and Speaker Recognition Workshop*, Brno, Czech Republic.

Morrison G.S., Zhang C., & Rose P., 2010. An empirical estimate of the precision of likelihood ratios from a forensic-voice-comparison system. Manuscript submitted for publication.

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

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

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

Pigeon S., Druyts P., & Verlinde P., 2000. Applying logistic regression to the fusion of the NIST'99 1-speaker submissions, *Digital Signal Processing*, 10, pp. 237–248. doi:10.1006/dspr.1999.0358

Potter R.K., Kopp A.G., & Green H.C., 1947. *Visible Speech*. New York: Van Nostrand.

Poza F., & Begault D.R., 2005. Voice identification and elimination using aural-spectrographic protocols. In: *Proceedings of the Audio Engineering Society 26th International Conference: Audio Forensics in the Digital Age*. Paper No. 1-1.

Ramos Castro D., 2007. *Forensic evaluation of the evidence using automatic speaker recognition systems*. PhD dissertation. Universidad Autónoma de Madrid.

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

Roberts H., 2004. Statistical evaluation in forensic DNA typing. In: Freckelton, I. & Selby, H. eds. *Expert evidence*. Sydney: Thomson Reuters. Ch. 80A.

Robertson B., & Vignaux G.A., 1995. *Interpreting Evidence*., Chichester (UK): Wiley.

Robertson B., & Vignaux G.A., 2000. Interpreting scientific evidence. In: Freckelton I., & Selby H. eds. *Expert evidence*. Sydney: Thomson Reuters. Ch. 28.

Rogers H., 2000. *The sounds of language: An introduction to phonetics*. Harlow (UK): Pearson Education.

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

Rose P., 2003. The technical comparison of forensic voice samples. In: Freckelton I., & Selby H. eds. *Expert evidence*. Sydney: Thomson Reuters. Ch. 99.

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

Rose P., & Duncan S., 1995. Naive auditory identification and discrimination of similar voices of familiar speakers. *Forensic Linguistics*, 2, pp. 1–17.

Rose P., & Morrison G.S., 2009. A response to the UK position statement on forensic speaker comparison. *International Journal of Speech, Language and the Law*, 16, pp. 139–163. doi:10.1558/ijsll.v16i1.139

Rose P., Osanai T., & Kinoshita Y., 2003. Strength of forensic speaker identification evidence: Multispeaker formant- and cepstrum-based segmental discrimination with a Bayesian likelihood ratio as threshold. *Forensic Linguistics*, 10, pp. 179–202.

Saks M.J., & Faigman D.L., 2008. Failed forensics: How forensic science lost its way and how it might yet find it. *Annual Review of Law and Social Science*, 4, pp. 149–171. doi:10.1146/annurev.lawsocsci.4.110707.172303

Saks M.J., & Koehler J.J., 2005. The coming paradigm shift in forensic identification science. *Science*, 309, pp. 892–895. doi:10.1126/science.1111565

Saks M.J., & Koehler J.J., 2008. The individualization fallacy in forensic science. *Vanderbilt Law Review*, 61, pp. 199–219.

Shriberg E., & Stolke A., 2008. The case for automatic higher-level features in forensic speaker recognition. In: *Proceedings of Interspeech 2008 Incorporating SST 2008*. International Speech Communication Association, pp. 1509–1512.

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

Thiruvaran T., Ambikairajah E., & Epps J., 2008. FM features for automatic forensic speaker recognition. In: *Proceedings of Interspeech 2008 Incorporating SST 2008*. International Speech Communication Association, pp. 1497–1500.

Thompson W.C., & Schumann E.L., 1987. Interpretation of statistical evidence in criminal trials: The prosecutor's fallacy and the defence attorney's fallacy. *Law and Human Behaviour*, 11, pp. 167–187. doi: 10.1007/BF01044641

Tosi O., Oyer H., Lashbrook W., Charles P., Nicol J., & Nash E., 1972. Experiment on voice identification. *Journal of the Acoustical Society of America*, 51, pp. 2030–2043. doi:10.1121/1.1913064

van Leeuwen D.A., & Brümmer N., 2007. An introduction to application-independent evaluation of speaker recognition systems. In: Müller C. ed. *Speaker Classification I: Fundamentals, Features, and Methods*. Heidelberg: Springer-Verlag, pp. 330–353. doi:10.1007/978-3-540-74200-5_19

Yarmey A.D., 1995. Earwitness speaker identification. *Psychology, Public Policy and Law*, 1, pp. 792-816. doi: 10.1037/1076-8971.1.4.792

Yarmey A.D., Yarmey A.L., Yarmey M.J., & Parliament L., 2001. Commonsense beliefs and the identification of familiar voices. *Applied Cognitive Psychology*, 15, pp. 283–299. doi:10.1002/acp.702