99
Forensic Speech Science

Geoffrey Stewart Morrison

BSc, MTS, MA, PhD

Ewald Enzinger

MPhil, PhD

Cuiling Zhang

BSc, MSc, PhD

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

A webpage related to this chapter can be found at:

http://expert-evidence.forensic-voice-comparison.net/

This includes access to previous editions of Chapter 99:

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

Rose P.J., 2003. The technical comparison of forensic voice samples. In Freckelton I., Selby H. (Eds.), *Expert Evidence* (Ch. 99). Sydney, Australia: Thomson.

The present version was prepared by the authors. The publisher distributes the content in different formats. Please cite by section number, not by page number.

Comments on the previous edition of Chapter 99 (Morrison, 2010, Forensic voice comparison):

Morrison has a very nice writing style and I think he has phrased some of the fundamental matters in a way that is more clearly put than I have ever seen. I think he has done a masterly job.

Dr John S Buckleton, Principle Scientist, ESR Forensics, Auckland, New Zealand

It is very informative and at the same time easy to read – a rare combination. It's a great book.

Dr Michael Jessen, Senior Scientist, Department of Speaker and Audio Analysis, Federal Criminal Police Office, Wiesbaden, Germany

Author information

Dr Geoffrey Stewart Morrison is Associate Professor of Forensic Speech Science, Centre for Forensic Linguistics, Aston University. His previous appointments include: Director, Forensic Voice Comparison Laboratory, School of Electrical Engineering & Telecommunications, University of New South Wales; and Scientific Counsel, Office of Legal Affairs, General Secretariat, International Criminal Police Organization (INTERPOL). In 2016 he was a Simons Foundation Visiting Fellow in the Probability and Statistics in Forensic Science Programme at the Isaac Newton Institute for Mathematical Sciences. He is author of more than 50 refereed and invited publications, has been a Subject Editor and a Guest Editor for the journal *Speech Communication*, and a Guest Editor for the journal *Science & Justice*. He has had research collaborations with law enforcement agencies in Australia, Europe, and the United States. He has been involved in casework in Australia, the United States, the United Kingdom, and Canada, including advising the defence in relation to a 2015 US Federal Court *Daubert* hearing on the admissibility of forensic voice comparison testimony.

Dr Morrison's webpages: http://geoff-morrison.net/
http://forensic-evaluation.net/
http://forensic-voice-comparison.net /

Dr Ewald Enzinger is a Research Engineer at Eduworks Corporation. His previous appointments include Research Scientist, Acoustics Research Institute, Austrian Academy of Sciences. He is a Guest Editor for the journal *Speech Communication*. He has had research collaborations with law enforcement agencies in Australia and Germany. He has been involved in casework in Australia and Austria. He graduated PhD 2016 from the School of Electrical Engineering & Telecommunications, University of New South Wales. His doctoral dissertation *Implementation of forensic voice comparison within the new paradigm for the evaluation of forensic evidence*, has been called by Prof John HL Hansen (Associate Dean for Research & Professor of Electrical Engineering, University of Texas at Dallas) "a remarkable and significant step forward in the field" and "one of the strongest research advancements in this domain to date".

Dr Enzinger's webpage: https://entn.at/

Prof Cuiling Zhang (张翠玲) is Director of the Chongqing Institutes of Higher Education Key Forensic Science Laboratory, and Vice Dean in the School of Criminal Investigation, Southwest University of Political Science and Law. Her previous appointments include: Director of the Forensic Speech Science Section, Department of Forensic Science, National Police University of China; and Visiting Professorial Fellow, Forensic Voice Comparison Laboratory, School of Electrical Engineering & Telecommunications, University of New South Wales. She has more than 20 years of forensic casework experience in China, and has worked on more than 200 cases. Her work was featured in a 2012 episode of the TV documentary series, *Partners in Crime*.

Prof Zhang's webpage: http://cuiling-zhang.forensic-voice-comparison.net/

Preface

The current edition of Chapter 99 is a revised and expanded version of the previous edition (Morrison, 2010, Forensic voice comparison). Seven years have passed since the publication of the previous edition. Much has changed in the intervening years, including advances in research and technology, and evolution in our knowledge and understanding of the field.

The current edition maintains forensic voice comparison as its primary topic. The previous edition had a heavy focus on acoustic-phonetic statistical approaches to forensic voice comparison. Since the publication of the previous edition we have conducted a number of studies comparing the performance of acoustic-phonetic systems and automatic systems under increasingly more forensically realistic conditions. Automatic systems performed much better and required much less investment of human time. As a result, the current edition has a heavier focus on automatic approaches. The examples of forensic voice comparison from the previous edition have been replaced with examples of the use of the automatic approach in actual cases.

The current edition also updates the previous edition's coverage of speaker identification by laypeople (previously titled non-technical speaker identification).

Additions for the current edition include a short section on legal admissibility of forensic voice comparison, substantial coverage of disputed utterance analysis, and brief coverage of other branches of forensic speech science. With the expansion of coverage, we have changed the title from "Forensic voice comparison" to "Forensic speech science".

Despite the changes, the current edition is a revised edition of the previous edition rather than a completely new work. The revised edition still has a relatively long section on human voices, which is a brief introduction to phonetics. We think that this is useful background information for understanding forensic voice comparison, and especially for understanding disputed utterance analysis. Some text from the previous edition has been deleted, but most has been revised, replaced, or augmented. To maintain the same section numbering as in the previous edition, the sections have not been reordered. Where sections have been deleted entirely, their section numbers have been retired. Where sections have been added, they have been given previously unused section numbers. The exception is that within the sections on examples of forensic voice comparison, the section numbers have been reused for the new examples. The text of the previous edition was somewhat cluttered by references. For the revised edition some references from the previous edition have been culled, and others have been moved to further reading sections. New references have also been added.

Preparation of the current edition of Chapter 99 was proximal in time to the writing of three other works with partially overlapping content: Morrison & Thompson (2017), Morrison (2018), and Morrison & Enzinger (2018). Although partially overlapping in content, we have tried to write each as a standalone work and write the overlapping content differently to address the different intended audiences. Morrison & Thompson (2017) and Morrison (2018) review admissibility of forensic voice comparison in the United States and in England & Wales respectively, and are primarily addressed to legal audiences. Morrison & Enzinger (2018) gives a more technical introduction to forensic voice comparison than the present work, and is primarily addressed to phoneticians. Of the four works, the present one is the only one to include sections on speaker recognition by laypeople and on disputed utterance analysis. It also gives greater coverage to

misinterpretations of forensic likelihood ratios (logical fallacies) than the other works. We hope that readers will find the four works complementary rather than redundant.

Finally, we would like to dedicate the current revised edition in memory of Dr Bryan James Found, who died suddenly on 23 October 2016. Bryan was Chief Scientist at Victoria State Police, and also held research positions at La Trobe University and at the University of New South Wales. He was well known for his pioneering work on empirical validation, cognitive bias, and forensic analysis of handwriting and signatures. He was extremely knowledgeable and insightful, was dedicated to improving forensic science, and we count him as one of the giants in the field. He was incredibly generous, and will be deeply missed by all who knew and loved him.

Geoffrey Stewart Morrison
Ewald Enzinger
Cuiling Zhang
2017

CONTENTS

INTRODUCTION

[99.10] What is forensic voice comparison?

A *forensic voice comparison* is an analysis conducted in order to help a court of law decide w*ho is speaking* on an audio recording.

In forensic voice comparison, a recording of a speaker of questioned identity (a *questioned-speaker recording*) is compared with one or more recordings of a speaker of known identity (a *known-speaker recording*). The known speaker is often a suspect or defendant and the questioned speaker is often an offender. Other scenarios are possible, e.g., the issue may be whether the questioned speaker is a particular victim or not.

Here are two representative forensic voice comparison scenarios (the details are fictional):

- 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 practitioner conducts a forensic comparison of the two voice recordings. In court, the forensic practitioner testifies that one would be 2000 times more likely to observe the acoustic properties of the voice on the fraud recording had it been made by the defendant than had it been made by some other speaker. 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 Mr Y being involved is weak and will not likely lead to a conviction. The audio recordings are provided to a forensic practitioner for analysis. The forensic practitioner conducts a forensic voice comparison and reports that one would be 1000 times more likely to observe the acoustic properties of the voice on the Mr X recording had it been produced by some other speaker than had it been produced by Mr Y. The police and prosecutor decide to focus their resources on other suspects.

Sections **[99.770]**–**[99.890]** provide examples of two real forensic voice comparison cases.

[99.12] What is speaker recognition by laypeople?

Speaker recognition by laypeople refers to the ability of people without any special training to recognise speakers' voices. Listeners may recognise the voice of a speaker they know. Someone hearing a crime being committed may think they recognise the voice of an offender, or someone who knows a suspect may be played a questioned-speaker recording and asked if they recognise the speaker. An *earwitness* is someone who hears the voice of an offender in a situation where no audio recording is available for analysis. The listener usually does not recognise the voice of the offender as someone they know, but may be asked to listen to a voice lineup and see if they

recognise any of the speakers in the lineup. Sections **[99.910]ff** describe speaker recognition by laypeople and contrast it with forensic voice comparison performed by forensic practitioners.

[99.14] What is disputed utterance analysis?

A *disputed utterance analysis* is conducted in order to help a court of law decide *what was said* on an audio recording.

The words spoken on an audio recording may be indistinct, and hence there may be a dispute about what was said, because of the speaking style (e.g., the speaker may have been out of breath), because of poor quality recording conditions (e.g., background noise), because the words are acoustically similar and therefore intrinsically difficult to differentiate (e.g., "fifteen" versus "fifty"), or because of some combination of the above. Sections **[99.1500]ff** describe disputed utterance analysis.

[99.20] Audience

As part of the *Expert Evidence* series this chapter is aimed first at lawyers, judges, and police investigators, however, it is hoped that this chapter will also be of interest to forensic scientists, phoneticians, 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 the focus will be on the understanding of concepts and the provision of basic knowledge.

[99.30] Structure

This chapter is structured in an order suitable for reading from beginning to end, but some readers may wish to go straight to sections covering particular topics. To aid the reader, cross-references point both backwards and forwards.

The first four major sections after this introduction describe the *new paradigm for forensic science* **[99.70]ff**, including the *likelihood ratio framework* for the evaluation of forensic evidence **[99.140]ff** and **[99.370]ff**, and the *testing of the validity and reliability* of forensic-comparison systems **[99.290]ff**. These provide an introduction to evaluation of forensic evidence applicable across all branches of forensic science, not just forensic speech science. This should be considered foundational material for readers not already familiar with these topics, and some later sections will assume knowledge of these topics.

The next major section **[99.440]ff** provides an *introduction to phonetics*, which will help the reader understand human voices, which are the source of the data that are analysed in forensic voice comparison and disputed utterance analysis. The next major section **[99.600]ff** describes *how speech is recorded* and how various *factors commonly affecting forensic casework recordings* degrade the quality of the speech information in those recordings.

The next major section **[99.650]ff** describes different *approaches to forensic voice comparison*. These can be thought of as different ways of extracting information from speech recordings.

Sections **[99.750]**– **[99.760]** briefly discuss the *admissibility of forensic voice comparison* in several common-law jurisdictions.

The next major section **[99.770]ff** presents two *examples of forensic voice comparison* analyses conducted in actual cases. These assume knowledge of many of the preceding sections (with the primary exception that knowledge of phonetics **[99.440]ff** is not essential to understand these examples).

The next major section **[99.910]ff** discusses *speaker recognition by laypeople* as opposed to forensic voice comparison conducted by forensic practitioners. Speaker recognition by laypeople includes earwitnesses who hear an offender speaking while a crime is being committed. When there is no audio recording available, earwitnesses may be asked to listen to a speaker lineup. Speaker recognition by laypeople also includes when individuals such as police officers listen to questioned-speaker recordings and then claim to recognise the speaker. An example from a real case is provided. Note that in this scenario a questioned speaker recording exists, so the listener is not a true earwitness, and a forensic voice comparison could potentially be conducted by a forensic practitioner. A reader whose immediate interest is speaker recognition by laypeople should be able to read sections **[99.910]ff** without having to read the preceding sections of the chapter, although knowledge of many of the preceding sections would probably help.

The penultimate major section **[99.1500]ff** describes *disputed utterance analysis*. Whereas in forensic voice comparison the question the court wants to answer is "Who was speaking?" in disputed utterance analysis the question is "What was said?" Knowledge of the new paradigm **[99.70]ff**, **[99.140]ff**, **[99.370]ff**, validation **[99.290]ff**, phonetics **[99.440]ff**, and factors affecting the quality of speech recordings **[99.600]ff** is assumed. Examples based on real cases are included.

The final major section **[99.1700]ff** briefly describes other branches of forensic speech science:

- Lie detection
- Intoxication detection
- Voice disguise in forensic casework
- Speaker profiling
- Language analysis for determination of origin (LADO)

[99.40] Questions

There are a number of questions which investigators, prosecutors, defence attorneys, judges considering admissibility, and triers of fact should ask about forensic speech science. The exact form of the questions and which questions are most important will differ depending on the questioner's role in the justice system, but they are fundamentally the same questions addressing the same underlying issues. The questions below are phrased assuming that the task at hand is forensic voice comparison.

1. Has the voice evidence been evaluated using the logically correct framework for the evaluation of forensic evidence?
2. Was the forensic voice comparison analysis based on quantitative measurements of the acoustic properties of the voices on the recordings?

3. Has an adequate database of voice recordings of speakers representative of the relevant population been used to assess the typicality of the questioned-speaker recording?
4. Has the strength of evidence been assessed using an appropriate statistical model, and is the output of that model directly reported as the strength of evidence statement?
5. Have the validity and reliability (accuracy and precision) of the forensic voice comparison system been empirically evaluated under conditions reflecting those of the known- and questioned-speaker recordings in the present case?
6. Is the demonstrated degree of validity and reliability acceptable?
7. What is the strength of evidence from the comparison of the voices on the known- and questioned-speaker recordings?

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. Also, how to reword the questions to apply to other branches of forensic speech science (and to other branches of forensic science in general).

A PARADIGM SHIFT IN FORENSIC SCIENCE

[99.70] A paradigm shift

We are currently 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.

[99.80] The new paradigm

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 echoes the requirements for admissibility of scientific evidence set out in Federal Rule of Evidence 702 (FRE 702, as amended Apr. 17, 2000, eff. Dec. 1, 2000; Apr. 26, 2011, eff. Dec. 1, 2011) and the 1993 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 should consider the methodology's scientific validity, including whether it has been empirically tested and found to have an acceptable error rate. In 2014 a section on expert evidence was added to the *Criminal Practice Directions* (CPD) in England & Wales (current version: [2015] EWCA Crim 1567 Consolidated with Amendment No. 2 [2016] EWCA Crim 1714 at [19A]). CPD 19A has substantial parallels with FRE 702 - *Daubert* and stated that "It is essential to recall the principle which is applicable, namely in determining the issue of admissibility, the court must be satisfied that there is a sufficiently reliable scientific basis for the evidence to be admitted." (CPD at 19A.4).

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 2009 National Research Council (NRC) report 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.

More recently, the 2016 report by President Obama's Council of Advisors on Science and Technology (PCAST) on *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods* found that empirical demonstration of scientific validity under casework conditions was still lacking in a number of branches of forensic science. The report opined that:

> neither experience, nor judgment, nor good professional practices (such as certification programs and accreditation programs, standardized protocols, proficiency testing, and codes of ethics) can substitute for actual evidence of foundational validity and reliability. The frequency with which a particular pattern or set of features will be observed in different samples, which is an essential element in drawing conclusions, is not a matter of "judgment." It is an empirical matter for which only empirical evidence is relevant. Similarly, an expert's expression of *confidence* based on personal professional experience or expressions of *consensus* among practitioners about the accuracy of their field is no substitute for error rates estimated from relevant studies. For forensic feature-comparison methods, establishing foundational validity based on empirical evidence is thus a *sine qua non*. Nothing can substitute for it. (PCAST, 2016, p. 6, emphasis in original)

Empirical validation is also required by the Forensic Science Regulator of England & Wales as part of accreditation (Forensic Science Regulator, 2014, 2016), and recommended by the European Network of Forensic Science Institutes' (ENFSI) *Methodological guidelines for best practice in forensic semiautomatic and automatic speaker recognition* (Drygajlo et al., 2015), the latter specifically in the context of forensic voice comparison.

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 we 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.

The use of the likelihood-ratio framework is recommended in the Association of Forensic Science Providers' *Standards for the Formulation of Evaluative Forensic Science Expert Opinion* (AFSP, 2009); the Royal Statistical Society's *Fundamentals of Probability and Statistical Evidence in Criminal Proceedings: Guidance for Judges, Lawyers, Forensic Scientists and Expert Witnesses* (Aitken et al., 2010); ENFSI's *Guideline for evaluative reporting in forensic science* (Willis et al., 2015); ENFSI's *Methodological guidelines for best practice in forensic semiautomatic and automatic speaker recognition* (Drygajlo et al., 2015) the latter specifically in the context of

forensic voice comparison; and implicitly by PCAST's report on *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods* (PCAST, 2016; see also Morrison, Kaye, et al., 2017).

[99.90] Further reading

For a history of the adoption of the new paradigm in forensic-voice-comparison research and practice up to 2009, see Morrison (2009). For a review of calls from the 1960s onward for the validity and reliability of forensic voice comparison to be empirically tested under casework conditions, see Morrison (2014).

THE LIKELIHOOD-RATIO FRAMEWORK FOR THE EVALUATION OF FORENSIC EVIDENCE

[99.140] Introduction

The likelihood-ratio framework has already been described in **Interpreting Scientific Evidence [28]** (Berger et al., 2016), and its application to DNA in **Statistical Evaluation in Forensic DNA Typing [80A]** (Federle et al., 2017). Other descriptions are listed in the further reading section below **[99.250]**. Here, we describe the likelihood-ratio framework in the context of forensic voice comparison.

[99.150] The likelihood-ratio framework

In the likelihood-ratio framework, the task of the forensic practitioner is to provide the court with a *strength-of-evidence* statement in answer to the question:

> How likely are the observed properties of the voice on the questioned-speaker recording (the *evidence*), had it been produced by the known speaker (the *same-speaker hypothesis*) versus had it been produced by some other speaker selected at random from the relevant population (the *different-speaker hypothesis*)?

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

Formula 1

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

where *LR* is the likelihood ratio; E is the evidence, i.e., the measured properties of the voice on the questioned-speaker recording; $p(E|H)$ is "probability of E given H"; and H_s is the same-speaker hypothesis, and H_d is the different-speaker hypothesis (or more generally *same-origin* and *different-origin* hypotheses, or *prosecution* and *defence* hypotheses).

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 practitioner must consider not only the degree of similarity between the samples, but also their degree of typicality with respect to the relevant population (we discuss the relevant population in section **[99.180]** below). In fictional television shows, forensic practitioners 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 similar in terms of some physical properties, this is of little value if these physical properties are also very typical, because under such circumstances 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 similar in terms of properties which are 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.

If the evidence is more likely to occur under the same-speaker hypothesis than under the different-speaker hypothesis then the value of the likelihood ratio will be greater than 1, and if the evidence is more likely to occur under the different-speaker hypothesis than under the same-speaker hypothesis then the value of the likelihood ratio will be less than 1.

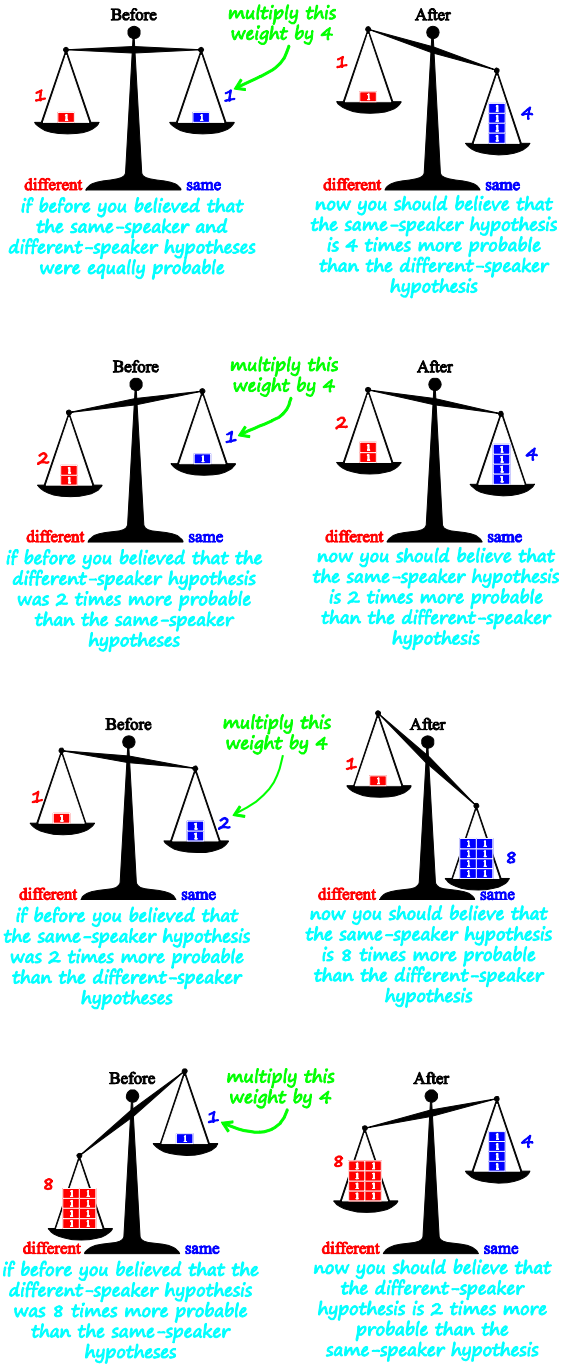
Likelihood ratios should not, however, be thought of as binary indicators – it matters how far the likelihood ratio is from 1. The value of the likelihood ratio is a numeric expression of the strength of the evidence with respect to the competing hypotheses. If the forensic practitioner testifies that one would be 100 times more likely to observe the evidence under the same-speaker hypothesis than under the different-speaker hypothesis ($LR = 100$), then whatever the trier of fact's belief about the relative probabilities of the same- and different-speaker hypotheses being true prior to hearing this, afterwards they should believe that the probability of the same-speaker hypothesis being true relative to the different-speaker hypothesis being true is 100 greater than they believed it to be before. Likewise, if the forensic practitioner testifies that one would be one thousand times more likely to observe the evidence under the different-speaker hypothesis than under the same-speaker hypothesis ($LR = 1/1000$), then whatever the trier of fact's belief about the relative probabilities of the same- and different-speaker hypotheses being true prior to hearing this, afterwards they should believe that the probability of the different-speaker hypothesis being true relative to the same-speaker hypothesis being true is 1000 greater than they believed it to be before.

Figure 1 shows a series of examples in which the likelihood ratio is 4, but the trier of fact's belief as to the relative probabilities of the same- and different-speaker hypotheses being true differ from example to example. The examples use the analogy of weights on a set of scales to represent the belief in the relative probabilities of the hypotheses being true.

In the first example, before hearing the likelihood ratio, the trier of fact believes that the probability that same-speaker hypothesis is true and the probability that different-speaker hypothesis is true are equal. This is represented by having an equal weight on each side of the set of scales. The evidence is 4 times more likely if the same-speaker hypothesis were true than if the different-speaker hypothesis were true, therefore the trier of fact should multiply the weight on the same-speaker side of the scale by 4. After doing this, the same-speaker side of the scales is 4 times heavier than the different-speaker side, i.e., the probability that same-speaker hypothesis is true is 4 times greater than the probability that different-speaker hypothesis is true.

In the second example, before hearing the likelihood ratio, the trier of fact believes that the probability that the different-speaker hypotheses is true is 2 times greater than the probability that the same-speaker hypotheses is true. This is represented by having 2 weights on the different-speaker side of the scales and one weight on the same-speaker side. The evidence is 4 times more likely if the same-speaker hypothesis were true than if the different-speaker hypothesis were true, therefore the trier of fact should multiply the weight on the same-speaker side of the scale by 4. After doing this, the same-speaker side of the scales is twice as heavy as the different-speaker side, i.e., the probability that same-speaker hypothesis is true is 2 times greater than the probability that different-speaker hypothesis is true.

Figure 1. Weights on scales as an analogy for the effect a likelihood ratio on beliefs about the relative probabilities of the same-speaker and different-speaker hypotheses being true.



Whatever the trier of fact's prior beliefs and whatever the value of the likelihood ratio, the same procedure applies to update those beliefs. The other two examples in Figure 1 show other prior beliefs, and hence other posterior beliefs. The likelihood ratio could have a different value. If the likelihood ratio were 10, this would require the weights on the same-speaker side of the scales to be multiplied by 10. If the likelihood ratio were 1/4, this would require the weights on the different-speaker side of the scales to be multiplied by 4. If the likelihood ratio were 1/10, this would require the weights on the different-speaker side of the scales to be multiplied by 10. The further the likelihood ratio from 1, the greater the strength of the evidence, and the greater the change in beliefs.

[99.160] Why the forensic practitioner must present the probability of evidence, and must not present the probability of hypotheses

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

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. What the trier of fact requires from a forensic practitioner, however, is a statement of the strength of a specific piece of evidence. It is not the role of a forensic practitioner to consider all the evidence. One forensic practitioner may present the strength of evidence related to specific DNA samples, another may present the strength of evidence related to specific fingermark / fingerprint 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 practitioner 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. This is in fact just a different expression of the concepts we described in section **[99.150]** using the analogy of weights on a scale.

Formula 2

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

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

In order to calculate the *posterior odds* (the relative probability of the same-origin versus the different-origin hypothesis, given the evidence), the forensic practitioner would need to know both

the *likelihood ratio* and the *prior odds*. The prior odds would represent the trier of fact's belief in the relative probabilities of the two hypotheses being true prior to the evidence being presented. When conducting their analysis, the forensic practitioner does not know the trier of fact's prior belief.

What might be reasonable prior probabilities?

If a crime were committed on an island and there are known to have been 100 people on the island at the time, then reasonable prior odds could be as follows:

- Assume that prior to hearing any evidence we assume that the suspect is no more or less likely to be guilty than any other individual on the island, and that in general no particular individual is more or less likely to be guilty than any other individual.
- There are 100 people on the island, therefore the prior probability for the suspect is 1/100.
- Similarly, the prior probability for each other individual on the island is 1/100.
- There are 99 other individuals on the island, hence the total probability for the other individuals is 99×1/100 = 99/100.
- Dividing the former probability by the latter, the prior odds are therefore: (1/100) / (99/100) = 1/99
- This can be represented as 99 weighs on the different-speaker side of the scale and 1 weight on the same-speaker side.

The reasoning above includes the assumption that, prior to hearing any evidence, no individual on the island is believed to be more or less likely to have committed the crime than any other individual. Although it may be appropriate for the trier of fact to make such an assumption, it is not appropriate for the forensic practitioner to do so. The trier of fact may take into consideration that some people live in parts of the island remote from where the crime was committed, or that some portion of the population are children who could not have physically committed the crime, and therefore the trier of fact may have prior odds different from 1/99. Also, if other evidence has already been presented in the trial, it is unlikely that the trier of fact's belief as to same-origin versus different-origin hypotheses 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 practitioner 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 practitioner were to present posterior odds then they would have to supply their own prior odds. If one forensic practitioner used a high value for the prior odds and another practitioner used a low one, and otherwise acted the same, the difference in the prior odds would make the value of the first scientist's posterior odds higher and that of the second lower, but this difference has nothing to do with the materials they were asked to compare. The forensic practitioner's choice of prior odds could be influenced by their own conscious or unconscious opinion as to the guilt or innocence of the defendant. *Cognitive bias* was a major concern in the NRC report (NRC, 2009, pp. 122–124).

[99.170] Terminology

Although the likelihood ratio is a component of Bayesian analysis, we have used the term *likelihood-ratio framework* rather than *Bayesian framework* since the latter, unlike the former, could imply that the forensic practitioner makes use of priors and calculates posteriors.

The fact that forensic practitioners present likelihood ratios in court does not imply that the trier of fact must assign numeric values to evidence which is not forensic comparison evidence, nor that they must arrive at their decision via the rigid application of a Bayes' Theorem.

Another terminological point is that in the likelihood-ratio framework the forensic practitioner does not perform *recognition*, *identification*, or *individualisation*, because these terms could imply making a categorical decision, which logically would require imposing a threshold on a posterior probability. A neutral term such as *comparison* is more appropriate. We therefore use the term "forensic voice comparison" rather than either of the traditional terms *forensic speaker identification* or *forensic speaker recognition*. We do not use *speaker comparison* since that would be akin to calling fingermark comparison *toucher comparison*. A term such as *forensic comparison of voice recordings* would be more accurate (it is the properties of the recordings which are actually compared, not the voices themselves), but, since the "of" construction has the potential to interfere with the understanding of sentence structure, we use the somewhat less exact term *forensic voice comparison*.

[99.180] A database representative of the relevant population

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 practitioner must have access to a database of samples which are representative of the relevant population. Such a database (sometimes called a *background database*) is necessary in order to calculate a quantitative estimate of the typicality of the voice on the questioned-speaker recording. A database representative of the relevant population is also needed to implement the validity and reliability testing requirements of the new paradigm **[99.290]ff**.

The relevant population is the population to which the questioned-speaker 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 from listening to the questioned-speaker recording. For example, if it were apparent that the speaker on the questioned-speaker 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.

Known- and questioned-speaker recordings are often (but not always) sent for forensic comparison after a police officer has listened to them and decided that they are sufficiently similar sounding that it is worth sending them for forensic comparison. If the voices on the two recordings had sounded very different, they would not have been sent for forensic comparison. In this scenario, a reasonable relevant population would be speakers who sound sufficiently similar to the voice on the questioned-speaker recording that a non-expert listener would think they were worth sending for forensic comparison. In practice, this would at least exclude speakers who sound

very different from the voice on the questioned speaker recording. For example, if the speaker had a very deep voice, speakers with high pitched voices would be excluded. This approach also provides a solution for cases such as when it is not clear whether the speaker is male or female. An appropriate relevant population could be speakers who sound similar to the voice on the questioned speaker recording, irrespective of their sex. This could include females plus males with high pitched voices, or males plus females with low pitched voices.

The defence could posit a more specific relevant population. An extreme case could be that the relevant population is the defendant's sister. Because the properties of the sister's voice would likely be more similar to those of a female defendant than most of the speakers in a larger relevant population, the denominator (the typicality part) of the likelihood ratio may be expected to be larger and the likelihood ratio therefore smaller. But this will not necessarily help the defence. If the size of the relevant population were in the thousands or millions, then the trier of fact might start out with prior odds of one over thousands or millions, but if the size of the relevant population were one, the trier of fact would be more likely to choose prior odds closer to one (equal probability for the defendant and for the sister). Also, the trier of fact considers all the evidence in the case. There may be lots of other evidence pointing towards the defendant but no other evidence pointing towards the sister.

Note that when they conduct the forensic voice comparison, the forensic practitioner is unlikely to be aware of the exact nature of the defence's hypothesis, and will usually have to anticipate what it may be. Whatever hypothesis the forensic practitioner adopts, they should clearly document what it is so that the judge at an admissibility hearing and/or the trier of fact at trial can decide whether it is appropriate or not.

[99.190] Differences between DNA data and voice data

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 recordings. These differences may lead to differences in the way the results of forensic DNA comparison and forensic voice comparison are presented, which may superficially give the impression that the two are not evaluated using the same framework. In fact, both DNA evidence and voice comparison evidence can and should be evaluated using the likelihood ratio framework.

This section includes a simplified account of forensic DNA comparison. Our purpose is to highlight some basic differences between DNA and voice data, not to discuss issues in the interpretation of DNA evidence.

A DNA profile consists 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 (locations and height of peaks on an electropherogram), but they have traditionally been converted back to discrete values to provide discrete DNA profiles for statistical analysis. To a first approximation it is assumed 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 1, and the probability of obtaining non-matching DNA profiles given the same-origin hypothesis is 0. The numerator of the likelihood ratio is therefore either 1 or 0.

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 prior odds are irrelevant since anything multiplied by 0 is still 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 came from a source other than the known source.

Often when the samples match, the *match probability* rather than the likelihood ratio is reported in court. The match probability is simply the denominator of the likelihood ratio, or equivalently the inverse of the likelihood ratio, i.e., it is the probability of obtaining the matching DNA profile under the different-origin hypothesis versus under the same-origin hypothesis.

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. (Another example of a continuous measurements is height: People do not have to be exactly 174 cm tall or 175 cm tall or 176 cm tall, they can be any value including 174.5, 174.9, 174.999999, 175.0000001 cm.) 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 at 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 speaking style and recording conditions (see **[99.600]ff**).

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 the difference not being "statistically significant", or in terms of the difference between the two not exceeding some pre-determined threshold, suffers from a cliff-edge effect (Robertson et al., 2016, pp. 148–150). 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. Approaches using a "match"/"non-match" threshold also fail to fully exploit the information available in the measurements made on the known- and questioned-origin samples, thus leading to poorer performance than would be obtained using statistical models which work directly with the continuously valued data (Morrison, Kaye, et al., 2017).

"Match" is therefore not a useful concept for the acoustic properties of voices, and forensic voice comparison results should be reported in the form of a full likelihood ratio. The same can be said for many other branches of forensic science.

In fact, even for DNA, problems have emerged with the "match"/"non-match" approach, leading to the use of full likelihood ratios rather than random match probabilities. With technological advances and increased sensitivity of equipment for extracting DNA, smaller and smaller amount of DNA can be analysed, but the probability of error in the measurement process increases. Thus

two measured profiles from the same individual may not have an exact match and it makes sense to use models in the numerator of the likelihood ratio that can provide probabilities between 0 and 1. A greater proportion of cases now involve mixtures of DNA from different individuals. Again, the solution requires models in the numerator of the likelihood ratio that can provide probabilities between 0 and 1.

[99.200] Calculating a forensic likelihood ratio

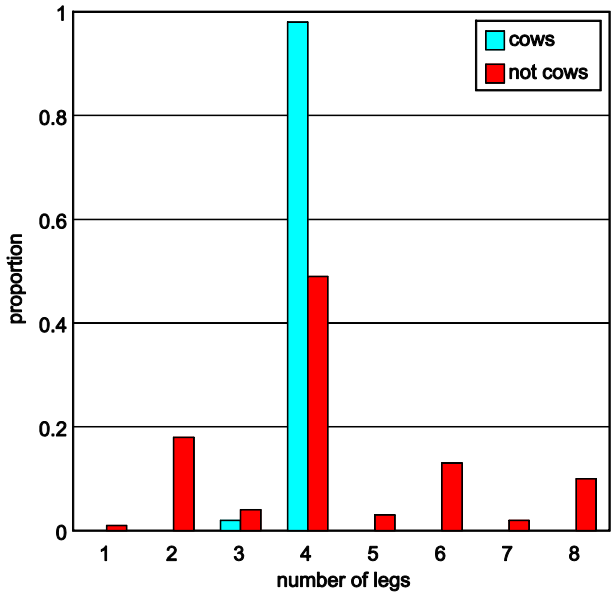
This section describes how to calculate a forensic likelihood ratio at a general conceptual level. At a detailed mathematical level there are multiple different procedures for calculating forensic likelihood ratios, many of which are much more complicated than those presented here. The aim of this section 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, they are not intended to be realistic.

[99.210] Calculating a forensic likelihood ratio from discrete data

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 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. 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 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 *bar graphs* in Figure 2. It turns out that 2% of the cows I saw had three legs and the other 98% had four legs (or in proportions 0.02 and 0.98). Note that, since each is a proportion of the whole, the heights of the red bars in Figure 2 sum to 1. 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). Note that, since each is a proportion of the whole, the heights of all the blue bars in Figure 2 sum to 1.

Now I am told that the evidence is that the animal in question has four legs. How do I calculate the likelihood ratio $p(\text{4 legs} \mid \text{cow}) / p(\text{4 legs} \mid \text{not cow})$? In Figure 2 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 would be twice as likely if the animal were a cow than if it were not a cow. Whatever one believed before hearing this evidence, one should now believe that the probability that the animal is a cow relative to the probability that it is not a cow is two-times greater than one believed it to be before.

FIGURE 2. Bar graph of discrete data.



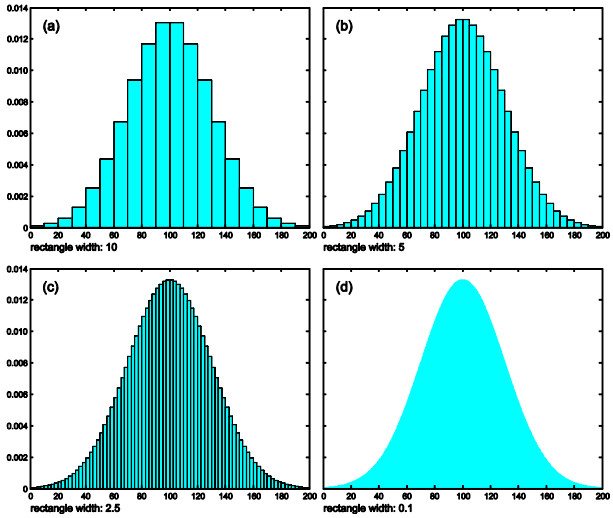
[99.220] From discrete data to continuous data

As noted in section **[99.190]** voice data are normally continuous, not discrete. For calculations based on continuous data, *bar graphs* are replaced by *histograms* or *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 3a). The *area* of a rectangle represents the proportion of the data that falls 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 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 3b). 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 3. Transition from histogram to probability density function for continuous data.



As the widths of the rectangles are reduced (Figures 3a through 3d), the size of the steps between rectangles decrease, not just the widths of the steps but also their height differences. Eventually the tops of the rectangles will look like a smooth curve rather than a series of steps (see Figure 3d). If we make some assumptions about the shape of this curve, such as that it is a *Gaussian distribution* (also called a *normal distribution*), then even with relatively small amounts of data we can skip straight to an estimate of the shape of the curve. The curve is the calculated *probability density function* trained on the data. To train a Gaussian distribution we only need to estimate the *mean* and *standard deviation*. Note that the total area under the curve is still equal to 1.

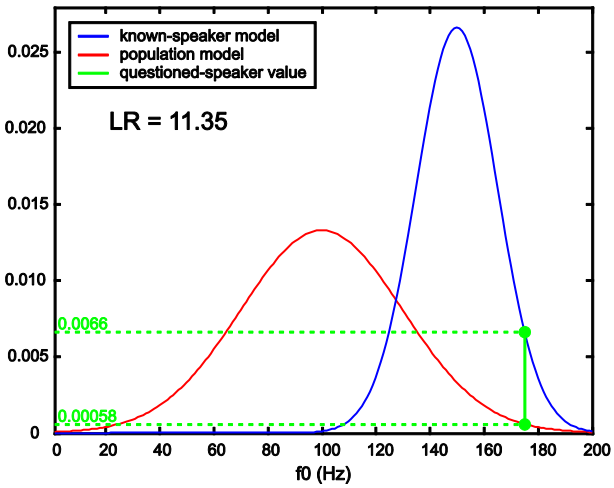
[99.230] Calculating a forensic likelihood ratio for continuous data

As mentioned above **[99.220]**, for calculations based on continuous data, bar graphs 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 4. 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 4. The former probability density function we will call the *population model*, and the latter the *known-speaker model*.

FIGURE 4. Calculation of a likelihood ratio from a known-speaker model and a population model.



In this example the population model is a Gaussian distribution with a mean of 100 Hz and a standard deviation of 30 Hz, and the known-speaker 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 voice on the questioned-speaker recording, then, at that value find the relative heights of the curves of the known-speaker and population models, see Figure 4. If the questioned-speaker value is 175 Hz, the probability-density-function (aka likelihood) value of the known-speaker model at 175 Hz is 0.0066, the probability-density-function (likelihood) value of the population model at 175 Hz is 0.00058, and 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 of the voice on the questioned-speaker recording if it had been produced by the known speaker than if it had been produced by a speaker selected at random from the relevant population.

What if, instead of 175 Hz, the mean f0 of the questioned-speaker recording was 150 Hz, right at the mean value for the known-speaker recordings? In this case, as shown in Figure 5, instead of 11.35, the likelihood ratio would be 8.02.

What if the voice on the questioned-speaker recording were even more typical and had a mean f0 of 125 Hz or 100 Hz? In these cases, as shown in Figure 6, 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 voice on the questioned-speaker recording is atypical in the opposite direction to the atypicality of the voice of the known speaker, then the support for the different-speaker hypothesis is even higher, e.g., if the voice on the questioned-speaker recording has a mean f0 of 75 Hz then the likelihood ratio is 94,810 in favour of the different-speaker hypothesis.

FIGURE 5. Calculation of a likelihood ratio from a known-speaker model and a population model.

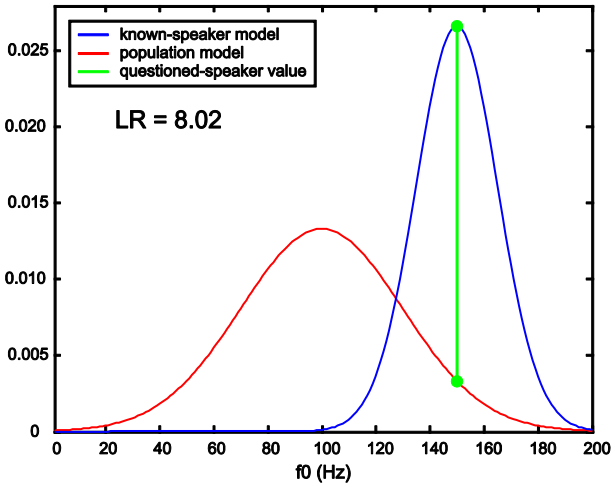
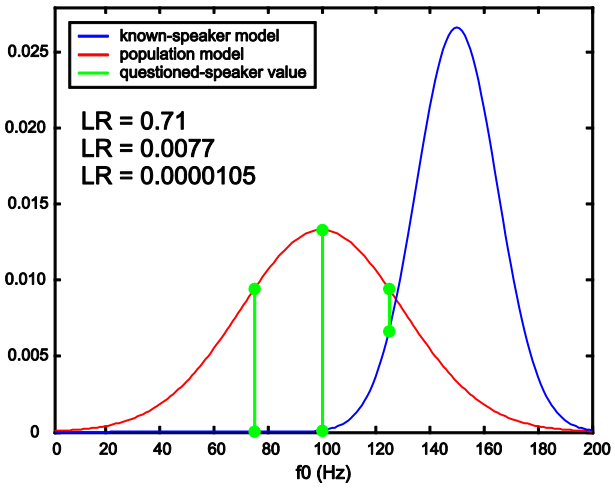


FIGURE 6. Calculation of likelihood ratios from a known-speaker model and a population model.



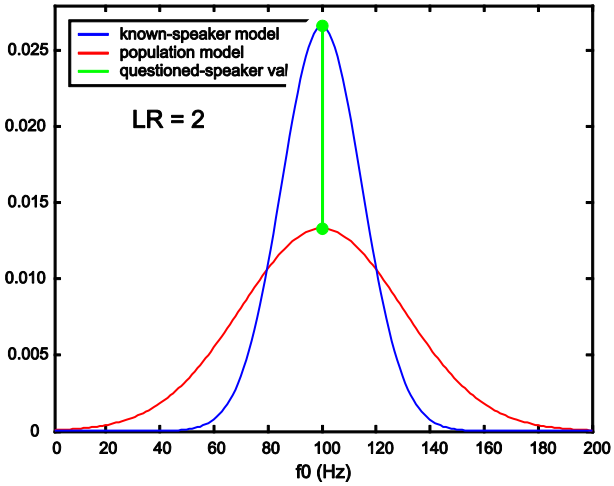
Note that at a value of approximately 128 Hz, the likelihood ratio would be 1 – one would be equally likely to obtain this value irrespective of whether the voice on the questioned-speaker recording had been produced by the known speaker or by another speaker from the relevant population.

In the previous examples the known-speaker model was relatively atypical. What if the voice of the known speaker were more typical? In Figure 7 the known-speaker model has the same mean as the population model (100 Hz), and is thus maximally typical. The standard deviations are unchanged from the previous examples. As was the case in Figure 5, the mean f0 value for the

questioned-voice sample is at the mean value for the known-speaker model, but instead of being 8.02, because the known-speaker model is now more typical, the likelihood ratio is only 2.

Note that even though in this example the known-speaker model is maximally typical and has the same mean as the population 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 voice on the questioned-speaker recording 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 being less likely for the population as a whole.

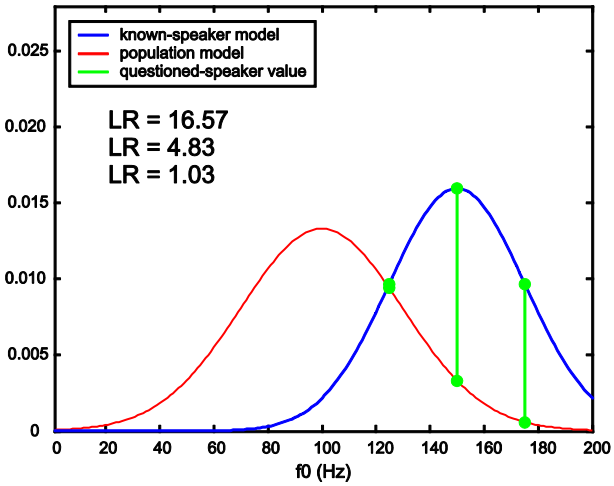
FIGURE 7. Calculation of a likelihood ratio from a known-speaker model and a population model.



What if the within-speaker variability were greater? The known-speaker model in Figure 8 has a standard deviation of 25 Hz as opposed to 15 Hz as was the case in the earlier examples. The mean for the known-speaker model is 150 Hz as was the case in Figures 4 through 6. 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-speaker values close to the mean of the known-speaker model now result in smaller likelihood ratios than before and questioned-speaker values relatively far from the mean of the known-speaker model now result in larger likelihood ratios than before.

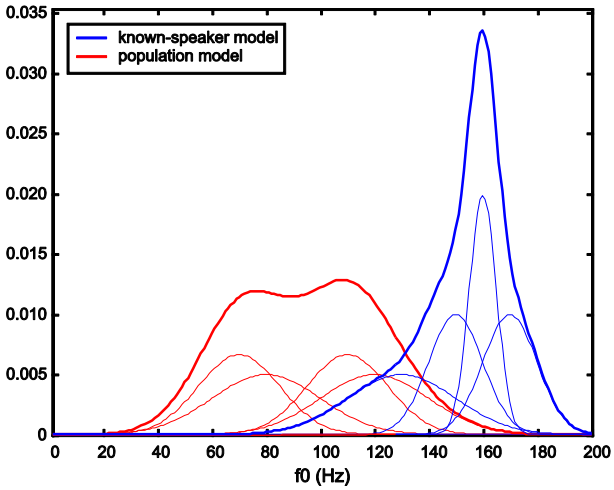
In general, the smaller the within-speaker variability relative to between-speaker variability, the better the performance of the forensic-comparison system (assessing system performance is discussed in **[99.290]**). Most speakers will be relatively typical (by definition) so most known-speaker models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

FIGURE 8. Calculation of likelihood ratios from a known-speaker model and a population model.



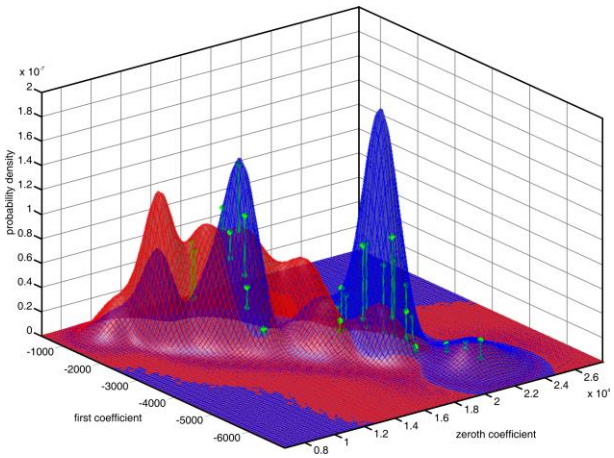
All the previous examples have used models consisting of a single Gaussian as a relevant-population model and a single Gaussian as a known-speaker model; however, more complex models are usually used in forensic voice comparison. For example, procedures based on *Gaussian mixture models* (GMMs) are common. Multiple Gaussians are used to fit a more complex distribution than can be achieved using a single Gaussian. Figure 9 provides an example of a relevant-population model and a known-speaker 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 was given equal weight).

FIGURE 9. One-dimensional Gaussian mixture models.



All the previous examples have been unidimensional, using measurements of a single acoustic property of the voice recordings. In practice, forensic voice comparison is usually based on measurements of multiple acoustic properties of voice recordings. This has the potential to lead to much larger likelihood ratios in favour of one hypothesis or the other. A recording of a voice 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 10 provides an example of a relevant-population model and a known-speaker model (each a Gaussian mixture model) in a two-dimensional space.

FIGURE 10. Two-dimensional Gaussian mixture models.



[99.240] Calibration and fusion

The models used in the description of the calculation of likelihood ratios above **[99.230]** are theoretically correct, but there may be a number of practical difficulties in using them. These could, for example, be related to whether the model is appropriate for the true distribution of the data, whether there are sufficient data to train 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, by different systems, e.g., an automatic system and an acoustic-phonetic system.

The practical solutions to these problems are called *calibration* and *fusion*, and a single procedure, *logistic regression*, can be used to do both (Brümmer & du Preez, 2006; González-Rodríguez et al., 2007; Pigeon et al., 2000). One way to view calibration is to consider the raw likelihood ratios calculated using the sorts of procedures described above, not as likelihood ratios per se, but rather simply as *scores*. Scores quantify the degree of similarity of pairs of samples while also taking account of their typicality, but their value are not directly interpretable as likelihood ratios answering the question posed by the same- and different-speaker hypotheses. Calibration converts

scores into likelihood ratios, or fusion converts parallel sets of scores from different systems into likelihood ratios.

[99.250] Further reading

General introductions to the likelihood-ratio framework can be found in numerous books and articles, including Robertson et al. (2016), Balding & Steele (2015) ch. 1–3 and 11, and Kaye et al. (2011). Introductions in the context of forensic voice comparison can be found in Rose (2002, 2003), Morrison & Thompson (2017), and Morrison & Enzinger (2018).

Selection of the relevant population in the context of forensic voice comparison is discussed in Morrison et al. (2012) and Morrison, Enzinger & Zhang (2016), see also Gold & Hughes (2014) and Hughes & Foulkes (2015).

A tutorial on logistic regression calibration and fusion is presented in Morrison (2013).

The problem of cognitive bias in forensic science is reviewed in Risinger et al. (2002), Saks et al. (2003), Found (2015), Stoel et al. (2015), and Edmond et al. (2017).

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

[99.290] Introduction

In judicial literature, the word *reliability* has often not been used without being explicitly defined. The *Daubert* ruling (at footnote 9) equates *evidential reliability* with *scientific validity*. *Daubert* advises the judge at an admissibility hearing to consider whether the forensic practitioner's methods have been empirically tested and found to have an acceptable error rate.

In statistics and scientific literature *validity* and *reliability* mean different things – *validity* is synonymous with *accuracy* and *reliability* 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, 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 their height is read off as the value marked on the pole.

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 fixed 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.

[99.300] Measuring the accuracy of a forensic-comparison system

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 common measure of accuracy 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 rate 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 rate is simply the inverse of correct-classification rate.

Classification-error rate (and correct-classification rate) 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, see section **[99.150]**.

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. Count the number of zeros in the likelihood ratio! A likelihood ratio of 1 corresponds to a log likelihood ratio of 0.

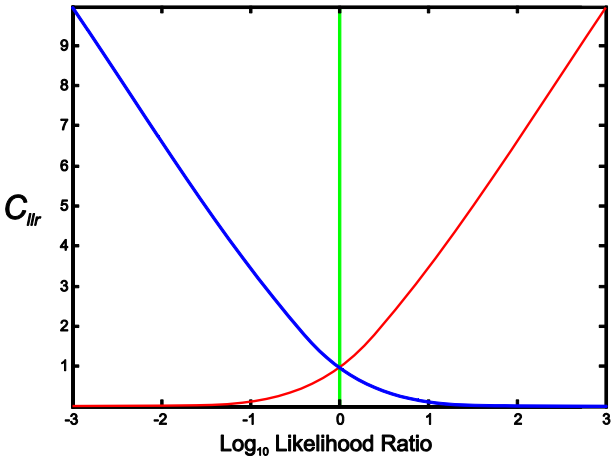
Ideally, for a same-origin pair the forensic-comparison system should produce a large positive log likelihood ratio, and for a different-origin pair it should produce a large negative log likelihood ratio. For a same-origin comparison, a small positive log likelihood ratio is not as good as a large positive 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 the trier of fact making an incorrect decision.

A measure of accuracy which is consistent with the likelihood-ratio framework is the *log-likelihood-ratio cost* (C_{llr}; Brümmer & du Preez, 2006). 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). 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 penalty value for the likelihood ratio from each test pair. Figure 11 provides a plot of the function for calculating a penalty 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 penalty values, log likelihood values close to zero provide little support for either the same-origin or different-origin hypothesis and are assigned moderate penalty values, and negative log likelihood values which contrary-to-fact

support the different-origin hypothesis are assigned high penalty values. The size of the penalty values increase rapidly as the log likelihood values become more negative and provide stronger support for the contrary-to-fact different-origin hypothesis. The function for calculating a penalty value when the input to the system is a different-origin pair (red line in Figure 11) is a mirrored version of the same-speaker function.

FIGURE 11. Plot of the function for calculating a C_{llr} penalty value for same-origin test pairs (blue line) and different-origin test pairs (red line).



To calculate C_{llr}, one finds the mean of all the penalty values from same-origin test pairs, the mean of all the penalty values from different-origin test pairs, and then takes the mean of the latter two means. The function for calculating C_{llr} is given in Formula 3, where N_s and N_d are the number of same-speaker and different-speaker test pairs, and LR_s and LR_d are the likelihood ratios derived from same-speaker and different-speaker test pairs. A same-origin penalty value is $\log_2(1 + 1/LR_s)$, and a different-origin penalty value is $\log_2(1 + LR_d)$.

Formula 3

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

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 data should therefore be samples which are representative of the relevant

population **[99.180]**, and the conditions of each member of each test pair should reflect as closely as possible to the conditions of the known- and questioned-speaker recordings (e.g., speaking style, recording quality, and recording duration, see section **[99.600]ff**). In this way, the results of the tests will reflect the expected performance in the system under the conditions of the case.

[99.310] Measuring the precision of a forensic-comparison system

It is important to consider the precision of a forensic-comparison system as well as its accuracy. All else being equal, a system that outputs a more precise value is better than a system that outputs a less precise value. There is, however, disagreement among forensic statisticians as to the best way to handle imprecision. Some provide a best estimate for the value of the likelihood ratio and a range of values within which they believe the value is likely to lie, e.g., a best estimate of 1000 with a 90% probability of being in the range 900 to 1100. An alternative to this is to only report the bound closest to a likelihood ratio of 1, e.g. a 95% probability that the likelihood ratio is at least 900. Others have philosophical objections to this approach, and instead calculate a single value, but that single value would usually be closer to a likelihood ratio of 1 than the first group's best estimate. The first group could be called *frequentists* and the second group *subjective Bayesians*, but in reality the situation is more complex and there are a variety of subgroups with different nuanced thinking on the issue. Part of the debate on this issue appears in a virtual special issue of the journal *Science & Justice*, and can be accessed at <http://www.sciencedirect.com/science/journal/13550306/vsi>. Given the lack of agreement on the issue, we do not here go into details as to how to calculate the precision of a forensic analysis system.

Imagine that a forensic practitioner has calculated a likelihood ratio using a sample of 200 speakers from the relevant population and presents the resulting value in court. Curran (2016, p. 380) points out that:

> An astute lawyer would also ask "If I took another sample of size 200, would this figure change?" The single most effective response to this question is "Yes, and my method for assessing this probability has already taken this into account." I believe that an expert witness who has used a statistically justifiable method for quantifying and adjusting for sampling uncertainty in his or her evaluation will be well-equipped to respond to the sample size question.

We recommend that lawyers engaging the services of a forensic practitioner ask the practitioner how the practitioner's method for assessing probability takes precision into account.

[99.330] Tippett plots

A graphical method for presenting the results of testing a forensic analysis system is a *Tippett plot*. Tippett plots were introduced in Meuwly (2001) (inspired by the work of C.F. Tippett), and are now a standard method for presenting results in 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 12 through 14 provide a series of Tippett plots drawn on the basis of hypothetical sets of output from forensic-comparison systems. The blue lines rising to the right represent the results from same-speaker test pairs. The value on the *y* axis is the cumulative proportion of log likelihood

ratios less than or equal to the value indicated on the *x* axis. The red lines rising to the left represent the results from different-speaker test pairs. The value on the *y* axis is 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 12. Tippett plot of hypothetical test results.

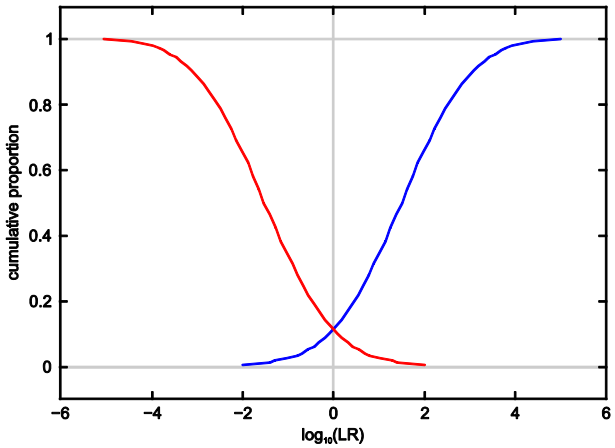


FIGURE 13. Tippett plot of hypothetical test results.

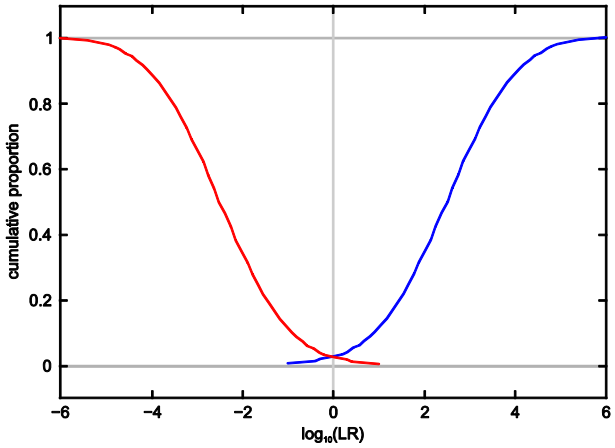
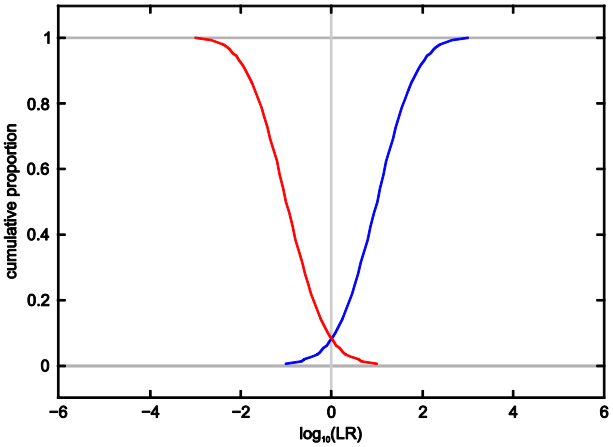


FIGURE 14. Tippett plot of hypothetical test results.



As discussed in section **[99.300]**, a forensic-comparison system with good performance would produce a large positive log likelihood ratio for a same-origin test pair, and a large negative log likelihood ratio for a different-origin test pair. 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 Tippett 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 performance. The results presented in the Tippett plot in Figure 13 therefore indicate a system with better performance than the system whose results are presented in the Tippett plot in Figure 12.

Note, however, that (as 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 reduces support for contrary-to-fact hypotheses is preferable even if this leads to some reduction in its strength of support for consistent-with-fact hypotheses. The results presented in the Tippett plot in Figure 14 are therefore better than those presented in the Tippett plot in Figure 12, but worse than those in Figure 13.

[99.340] Further reading

For more detailed descriptions of procedures for empirically testing the validity and reliability of forensic analysis systems (including forensic voice comparison systems), and metrics and graphics for communicating the results, see Morrison (2011), Meuwly et al. (2016), and Morrison & Enzinger (2016).

MISINTERPRETATIONS OF FORENSIC LIKELIHOOD RATIOS

[99.370] Introduction

A problem for the presentation of a likelihood ratio as a strength of evidence statement is potential misinterpretation of the meaning of the likelihood ratio. Misinterpretations can occur in the mind of a lawyer, judge, or jury member, and can be inadvertently caused by a forensic practitioner misphrasing their strength of evidence statement. Forensic practitioners should be careful not to inadvertently cause a misinterpretation via misphrasing, and as much as 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. Common misinterpretations include the *prosecutor's fallacy*, the *defence attorney's fallacy*, and the *trier of fact's fallacy*.

[99.380] The prosecutor's fallacy

Forensic Practitioner:

> One would be one thousand times more likely to obtain the measured acoustic properties of the voice on the questioned-speaker recording if it had been produced by the accused than if it had been produced by some other speaker from the relevant population.

Prosecutor:

> So, to simplify for the benefit of the jury, what you are saying is that the probability that the defendant is the questioned speaker is one thousand times greater than the probability that someone else from the relevant population is the questioned speaker.

The forensic practitioner's statement above is an expression of a likelihood ratio (relative probabilities of evidence given hypotheses). It does not include consideration of prior odds. In contrast, the prosecutor's statement above is an expression of posterior odds (relative probabilities of hypotheses given evidence), which logically must depend on both a likelihood ratio and prior odds, see **[99.160]**. The posterior odds and the likelihood ratio would only have the same value if the prior odds were 1 (if a priori the same- and different-speaker hypothesis were equally probable), which is seldom the case. If the prior odds are not explicitly 1, then equating the value of the posterior odds with the value of the likelihood ratio is fallacious. It is called the *prosecutor's fallacy*, or more generally the *transposition of the conditionals*.

To understand why the prosecutor's fallacy is such a serious mistake let us return to the cow example from section **[99.210]**. Imagine that we tell you we have a cow somewhere out of sight and we 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 we reported in **[99.210]** we said that 98% of cows had four legs, which corresponds to a probability of 0.98. In reality the probably may be much closer to 1.

Now let us 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 close to 1 – lots of animals including sheep, pigs, horses, dogs, cats, giraffes, and elephants usually 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:

> The probability of the animal having four legs given that it is a cow is very high.

And misphrase or misinterpret it as:

> Given that the animal has four legs, the probability of it being a cow is very high.

Or similarly take the statement:

> The probability of the occurrence of the acoustic properties of the voice on the questioned-speaker recording is much higher had it been produced by the known-speaker than had it been produced by some other speaker.

And interpret it as:

> Given the acoustic properties of the voice on the questioned-speaker recording, the probability that it was produced by the known-speaker is much higher than the probability that it was produced by some other speaker.

What is missing is the prior probability or the prior odds.

For sake of argument, assume the likelihood ratio is 99 and the trier of facts' prior odds are 1/1000. If the prosecutor's fallacy were committed, the mistake would be to interpret the likelihood ratio of 99 as posterior odds of 99. This is equivalent to a 99% posterior probability that the defendant is the questioned speaker.

In fact, using Formula 2 from **[99.160]** (prior odds × likelihood ratio = posterior odds), the correct posterior odds would not be 99, but instead: 1/1000 × 99 = 99/1000 = 0.099. This is equivalent to a 9% posterior probability that the defendant is the questioned speaker.

9% is very different from 99%!

For the mathematically inclined, the equation for converting from coherent odds, $o(H_s) = p(H_s) / p(H_d)$, to probability, $p(H_s)$, is given in Formula 4 (the odds, and the probabilities, are *coherent* if and only if $p(H_s) + p(H_d) = 1$).

Formula 4

$$p(H_s) = \frac{o(H_s)}{1 + o(H_s)}$$

[99.385] Avoiding the prosecutor's fallacy

The term "prosecutor's fallacy" was coined for transposition of the conditional in a legal context since, assuming the prior odds are less than 1, it is more advantageous to the prosecution than to the defence. Although the term "prosecutor's fallacy" may suggest that a prosecutor would transpose the conditionals, it is in fact a mistake which can easily be unintentionally made by prosecutors, defence council, judges, jury members, journalists, and forensic practitioners. A way

to help avoid making this mistake is to always ask: What is the evidence and what are the hypotheses? Substitute "number of legs" for the evidence, and "cow" versus "not cow" for the hypotheses. Then decide whether the statement is of the form "probability of legs given cow versus not cow", or of the form "probability of cow versus not cow given legs" or "probability of cow given legs". If it is one of the latter two and it is what the forensic practitioner said or an interpretation of what the forensic practitioner said, then it is probably an example of the prosecutor's fallacy.

[99.390] The defence attorney's fallacy

Forensic Practitioner:

> One would be one thousand times more likely to obtain the measured acoustic properties of the voice on the questioned-speaker recording had been produced by the accused than if it had been 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, we multiply by one thousand and arrive at posterior odds of one over one hundred thousand (1/1,000,000 × 1000 = 1/100,000). 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. In fact, the likelihood ratio should have shifted the trier of fact's beliefs by a factor of 1000, which is not insubstantial. By itself it may not be enough to convince the trier of fact that the same-origin hypothesis is true, but when the trier of fact weighs all the evidence, it may make a substantial contribution.

Different types of evidence (e.g., DNA, fingerprints, voice recordings) can reasonably be assumed to be statistically independent, and if they all address the same same-origin versus different-origin hypotheses, the likelihood ratios can be multiplied together. The prior odds plus the likelihood ratios can be multiplied in any order, it makes no difference mathematically. If we started out with prior odds of 1/1,000,000 and then heard four pieces of evidence, each with a likelihood ratio of 1000, the defence attorney's fallacy would argue that each piece of evidence should be dismissed, but the posterior odds would be 1/1,000,000 × 1000 × 1000 × 1000 × 1000 = 1,000,000 (in $\log_{10}$: −6 + 3 + 3 + 3 + 3 = +6). Posterior odds of one million may well lead the trier of fact to conclude that the same-origin hypothesis is true (and that the different-origin hypothesis false).

The term "defence attorney's fallacy" is used since the outcome of committing the fallacy is usually advantageous to the defence, but the mistake can be unintentionally made by defence council, prosecutors, judges, jury members, journalists, and forensic practitioners.

[99.394] The trier of fact's fallacy

Forensic Practitioner:

> One would be one billion times more likely to obtain the DNA profile of the blood found at the crime scene had it come from the accused rather than from another individual in the country unrelated to the accused.

Trier of Fact (thinking):

> One billion is a very big number. The blood must have come from the accused. He must be guilty. I can ignore the other evidence.

There are several counts on which, in this example, the trier of fact's logic is fallacious. The defendant could be innocent and the true offender could be a relative of the defendant (a likelihood ratio given a relevant population of close relatives would be much smaller), or there could have been a mistake leading to contamination, miscalculation, or misreporting. The likelihood ratio quoted addressed source level propositions: same-origin versus different-origin. It did not address activity level (how the blood got to be at the crime scene) or offence level (whether the accused is guilty or a crime or not). The blood could be that of the defendant without the defendant having committed the crime. For example, the defendant may have been present at the crime scene, attempted to prevent the crime from being committed and ended up shedding blood in the process. These may be issues considered by the court, but they were not addressed by the forensic scientist's source-level conclusion.

Even if none of the above were true, there is still a fallacy in the trier of fact's reasoning. The likelihood ratio presented by the forensic practitioner is probabilistic, not definitive. Even though the value of the likelihood ratio is very large, it is not infinite. This means that other evidence could potentially outweigh even this very large likelihood ratio. What if eye witnesses stated that the defendant did not resemble the person they saw committing the crime, and the defendant had a very strong alibi? This is not forensic science evidence and it does not come with a numeric likelihood ratio attached, but the trier of fact should consider whether the other evidence outweighs even the very strong DNA evidence. The trier of fact may still decide that the defendant is the source of the blood, but they should do so after considering the weight of all the relevant evidence presented to them, and not prematurely jump to a conclusion.

We have coined the term "trier of fact's fallacy" for this mistake. It could also be called the "large number fallacy", and be considered the inverse of the defence attorney's fallacy which could be called the "small number fallacy". As with the prosecutor's and defence attorney's fallacies, it can potentially be made by various actors, not just triers of fact. Indeed, there is a version of this fallacy where the forensic practitioner obtains a very large number and rounds it up to a probability of 1 (100%) and declares an "identification", or obtains a very small number and rounds it down to a probability of 0 (0%) and declares an "exclusion". In such circumstances, it would be better called the "forensic practitioner's fallacy".

[99.398] Further Reading

The terms *prosecutor's fallacy* and *defence attorney's fallacy* were coined by Thompson & Schumann (1987). They are also described in Robertson et al. (2016, ch. 9) and Balding & Steele

(2015, ch. 11). Hicks et al. (2016) includes advice on how to avoid the prosecutor's fallacy. Koehler (2014) discusses instances of fallacies found in some actual legal rulings.

HUMAN VOICES
(A BRIEF INTRODUCTION TO PHONETICS)

[99.440] Introduction

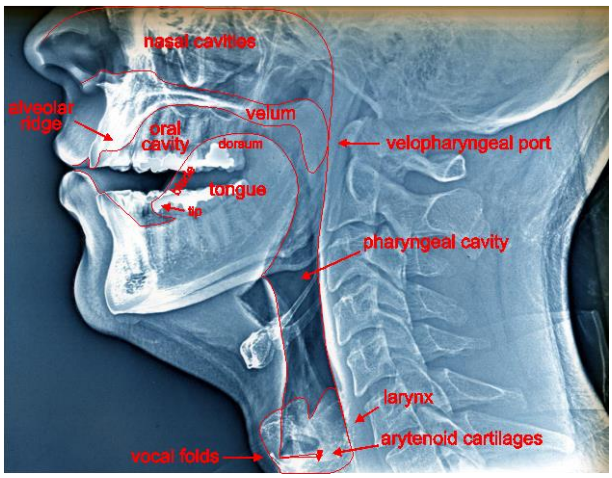
Phonetics is the study of the physical aspects of the production, transmission, and perception of human speech. This section provides a brief introduction to articulatory and acoustic phonetics, which cover the production and transmission of speech. The intent is to provide the reader with a basic understanding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

This introduction is not meant to be exhaustive. Suggestions for further reading are given in section **[99.560]**.

[99.450] Vocal tract

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 15 (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 15. The jaw can be lowered or raised and the tongue can be moved to change the shape of the oropharyngeal tube.

FIGURE 15. X-ray and tracing of a vocal tract.



[99.460] Vowels

[99.461] Description

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 section **[99.540]**). 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 "zzzzzzzzz". This is a voiced sound, you should be able to feel the vibration with your fingers. Next say "sssssss". This is a voiceless sound, you should not be able to feel vibration with your fingers. Now try saying "buzz" and "bus" – in both cases the vowel is voiced but the following consonant in "bus" isn't voiced 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 tubes and short 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 these words may depend on your accent). 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 15 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) <https://www.international phoneticassociation.org/content/ipa-chart> 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 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 vocal folds do not actually vibrate during the final consonant. "Bus" /bʌs/ would be [bʌs], without a long vowel, hence the actual difference in pronunciation of "buzz" and "bus" may be vowel length, not presence of absence of vocal fold vibration during the final consonant.

Figure 16 shows the *spectra* (singular: *spectrum*) of the vowels /i/, /ɛ/, and /u/ (spoken by Morrison). 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 measured formants, 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 17 (vowels spoken by Morrison). 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 vowel in "had" versus the first vowel in "father"), and ends up with a low F1 and a high F2, somewhere between /ɪ/ and /i/ (the vowel in "hid" versus the vowel in "heed"). In a broad Australian-English accent /aɪ/ may be realised as [ɔɪ]

rather than [aɪ], and in a Canadian-English accent "hide" may be realised as [haɪːd̥] ([̥] is the voiceless diacritic) but "height" as [hʌɪt].

FIGURE 16. Spectra of vowels /i/, /ɛ/, and /u/.

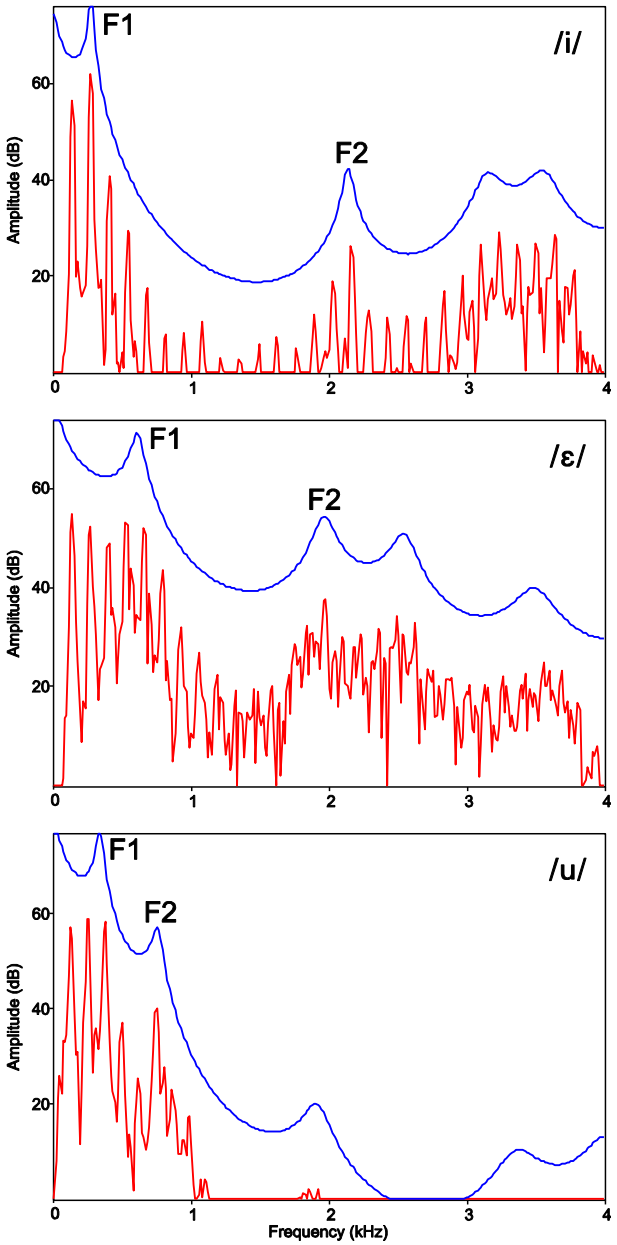
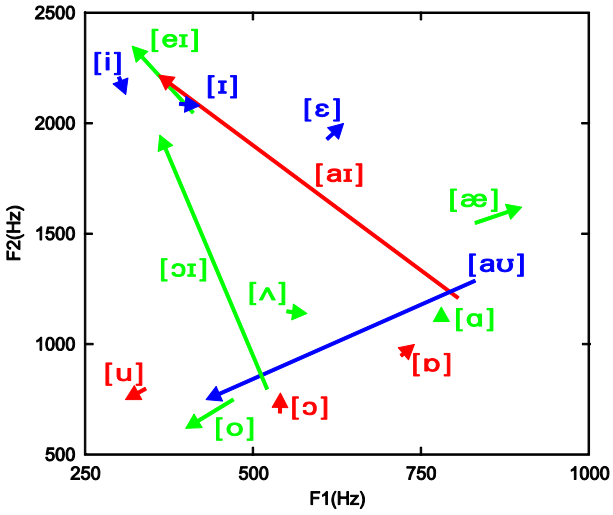


FIGURE 17. Plot of F1 and F2 measurements of a set of English vowels.

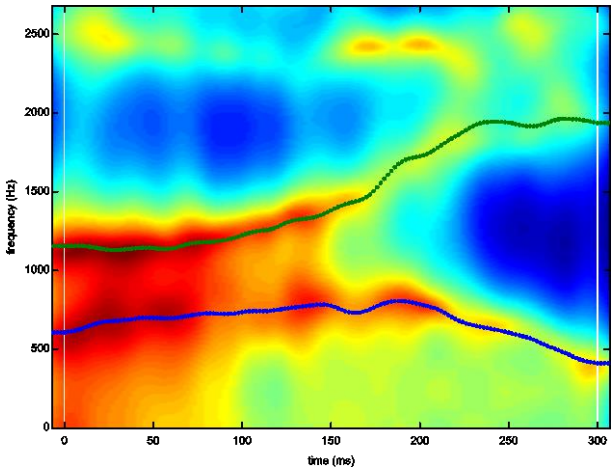


Another graphical method for representing the acoustic speech signal is a *spectrogram*. A spectrogram is made by measuring the spectrum of the speech signal every few milliseconds, then lining those spectra up in order so that time is on the *x* axis and frequency is on the *y* axis. On a three-dimensional 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 18 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 overlaid.

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 section **[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".

FIGURE 18. Spectrogram of /aɪ/.



[99.470] Potential forensic value

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 section **0**. Also, unlike DNA profiles or fingermarks, intrinsic within-speaker variability of many of the acoustic properties of speech may be very high.

[99.480] Nasals

[99.481] Description

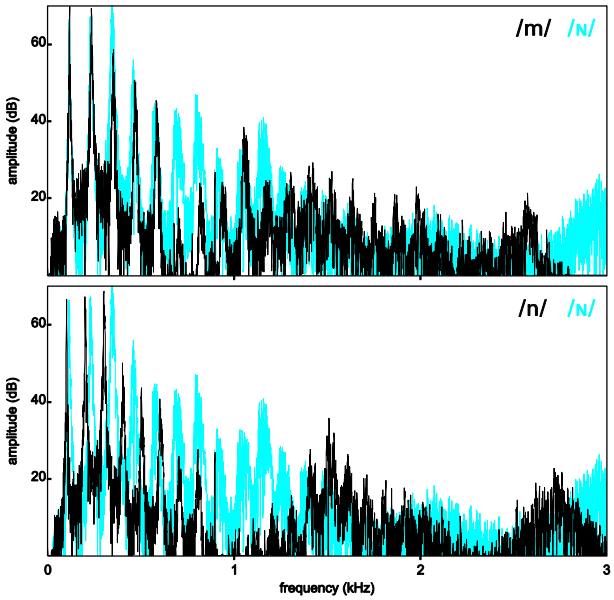
Nasals, such as /m/, /n/, and /ŋ/ (the last sound in “sum”, “sun”, and “sung” respectively) are made by producing voicing, opening the velopharyngeal port so that air can flow through the nasal cavities (see Figure 15), and making a closure in the oral cavity. The velopharyngeal port is also held open when one is breathing through one’s nose, but without making a speech sound. The

tracing of the nasal cavities in Figure 15 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 15) 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 15). For /ŋ/ the closure is made between the *dorsum* of the tongue and the velum (see Figure 15), 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 19 shows the raw spectra of /m/ and /n/ compared to /ɴ/. The latter 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 19 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.

FIGURE 19. Spectra of nasals /m/ and /n/ compared to /ɴ/.



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*.

[99.490] Potential forensic value

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 section **[99.610]**.

[99.500] Fricatives

[99.501] Description

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/ (the latter sound is more common in the middle of words like "pleasure"). Scottish English has an additional voiceless fricative /x/, the last sound in the Scottish word "loch".

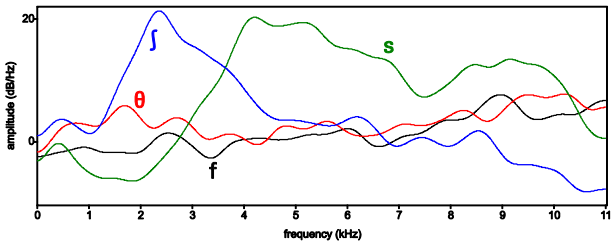
Fricatives are produced by making a constriction in the vocal tract, which is narrower than the constriction for making a vowel, and forcing air through the constriction quickly so that it makes a noise. The noise is the result of *turbulent airflow* – imagine a wide deep slow moving river, this has *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 lips. 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.

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 Morrison) are shown in Figure 20.

FIGURE 20. Smoothed spectra of English voiceless fricatives.



[99.510] Potential forensic value

With the exception of /s/ and /ʃ/, fricatives are relatively quiet. This can make them difficult to measure if there is any background or channel noise. /f/ and /θ/ have similar spectra (see Figure 20) 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 20) which are usually lost in telephone transmission, see section **[99.610]**. The spectra of fricatives may be too much affected by telephone transmission to make them particularly useful for forensic voice comparison.

[99.520] Plosives

[99.521] Description

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 before voicing begins, this is called *aspiration*. The phonetic details of utterance-initial English plosives are therefore, for example, /b/→[p] and /p/→[pʰ] ([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/ in section **[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.

[99.530] Potential forensic value

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.

[99.540] Laryngeal activity

[99.541] Description

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 (Mandarin), f0 is used to distinguish phonemes, e.g., 媽 (traditional character) 妈 (simplified character) "mā" /ma˥/ MOTHER with a steady high f0, and 罵 (traditional character) 骂 (simplified character) "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 if a wind blows through it. 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 15). If there is turbulent airflow through this gap, *breathy voicing* is produced. Some speakers may have a habitually breathy voice (Marilyn Monroe 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 actually 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}/ [mi^{32}] WHEAT and "mị" /mi^{21}/ [mḭ^{21}] COAX ([̰] is the diacritic for creaky voice and the numbers indicate the tones – historically the lowest tones have become creaky). Creaky voice has also been noted as a sociolinguistic marker for younger Australian English speakers and younger female American English speakers (including Scarlett Johansson when playing American, but not British, characters; Shaw & Croker, 2015).

[99.550] Potential forensic value

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, but be difficult to extract from anything other than high-quality audio recordings.

[99.560] Further reading

There are number of introductory phonetics textbooks available. Since its first edition in 1975, Ladefoged & Johnson (2014) has probably been the most widely read phonetics textbook. A different perspective from the same first author is given in another introductory textbook, Ladefoged & Disner (2012). Rogers (2000) includes a chapter describing a number of different dialects of English. Rose (2002) includes an introduction to phonetics as part of an introduction to forensic voice comparison.

VOICE RECORDING, TRANSMISSION, AND STORAGE

[99.600] Voice recording

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 *quantisation resolution* (sampling frequency multiplied by quantisation resolution gives *bit rate*). Sampling frequency refers to the number of times per second a measurement (a sample) is taken. Typical high-quality sampling rates are 44.1 kHz or 48 kHz, which can be used to record audio frequencies up to 22.05 kHz or 24 kHz respectively. 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 (for more on human hearing see **Hearing and the perception of sound [69]**, Alias, 2015). Quantisation resolution refers to the number of binary digits (bits) used to encode the intensity of the signal at each sample. The higher the quantisation resolution, the more detailed the representation of intensity can be and the better the recording quality. A common quantisation resolution is 16 bits which allows 65 536 different levels of intensity to be encoded. Each sample has an intensity value and moving from one sample to the next the intensity value usually changes. At low sampling frequencies and low quantisation resolutions, there can be big steps from sample to sample, but at high sampling frequencies and high quantisation resolutions the steps are very small and approximate smooth transitions.

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 quantisation 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 (a low *signal to noise ratio*). 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 values 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 the frequency envelope of the recorded signal (such differences are collectively known as *channel effects*). Since known-

and questioned-speaker recordings may be made on different recording systems under different recording conditions, this can introduce a source of variability which is conflated with within- and between-speaker variability.

Background noise is common on casework recordings. There are many different sources of background noise, e.g., ventilation systems, multiple other speakers speaking at once (*babble*), music, traffic, machinery. The particular type of background noise and signal to noise ratio often differs between the known- and questioned-speaker recordings. Known-speaker recordings are often recordings of interviews made in small rooms with hard walls – this results in substantial *reverberation* (echoes).

Degradation in speech signal information in the recording due to recording channel effects, background noise, and reverberation, and *mismatches* in recording conditions between the known- and questioned-speaker recordings leads to deterioration in the performance of forensic voice comparison systems.

[99.610] Voice transmission and storage

In forensic voice comparison casework, questioned- and/or known-voice recordings are often recordings of telephone conversations. Telephone systems introduce additional *channel effects*, which degrade the quality of the speech signal information and can be a source of variability between recordings.

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 (quantisation resolution is 8 bits, i.e., 256 levels of intensity). 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, f0 is recoverable from other parts of the spectrum (*harmonics*, which occur at multiples of the f0 value).

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 because 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 within a few seconds (in theory it could change every 20 ms), and, if the user is on the move, quality can change 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 reduces the quality of the transmitted signal. Bit rates can vary between 4.75 and 12.2 kbits/s, much lower than for a landline. Quality is dependent on the bit rate used to encode the signal. Occasional missing packets may be replaced by the immediately preceding packets.

Direct-to-satellite mobile-telephone systems compress the speech data more than terrestrial mobile-telephone systems leading to an even greater loss of information.

Voice-Over-Internet-Protocol (VoIP) systems, such as Skype, are broadly similar to mobile-telephone systems, but potentially have a wider bandpass and use higher bit rates. They can 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. In some parts of the world some user's landline telephones have been replaced by VoIP systems, especially in office environments. The service provider may provide dedicated bandwidth to avoid the worst quality problems that can sometimes plague free peer-to-peer systems such as Skype. Frequent users of such systems are aware that the quality can range from excellent to unintelligible.

Mobile telephone systems and VoIP systems use what are called *lossy codecs*. Information that was in the original signal is lost, and cannot be recovered. Various formats for saving or storing audio recordings also use lossy codecs. The most famous lossy codec for audio recordings is probably MP3, but there are many others in use. Video recordings which include an audio track may also compress the audio signal using a lossy codec. Lossy codecs are used so that recordings take up less storage space, or to allow files to be sent to or retrieved from remote servers more quickly, including live streaming.

Channel effects, especially lossy codecs used for transmission and/or storage, lead to deterioration of the speech signal information in the audio recording. Poor quality channels and mismatches in channels lead to deterioration in the performance of forensic voice comparison systems.

[99.620] Further reading

For a more detailed relatively non-technical description of different telephone systems, see Guillemin & Watson (2008).

APPROACHES TO FORENSIC VOICE COMPARISON

[99.650] Introduction

Historically there have been four basic approaches to forensic voice comparison: *auditory*, *spectrographic*, *acoustic-phonetic*, and *automatic*. What we mean by *approach* is a general method for extracting information from voice recordings for the purpose of conducting forensic voice comparison. To a greater or lesser extent the approach can be independent of the *framework* for inferring the strength of 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 a posterior-probability framework. In this section, each of the approaches is described and then evaluated with respect to its compatibility with the new paradigm.

[99.660] Auditory approach and auditory-acoustic phonetic approach

[99.661] Description

The *auditory approach* is practised by phoneticians with training and experience in auditory phonetics. This would include the ability to use phonetic symbols and diacritics to transcribe the speech sounds which they hear. The practitioner listens to the known- and questioned-speaker recordings and notes similarities and differences. The practitioner notes similarities which they would expect to observe if the recordings were of the same speaker but not if they were of different speakers, and differences that they would expect if the recordings were of different speakers but not if they were of the same speakers.

Audible features which are exploited could be the sorts of differences which distinguish dialects, e.g., consider how the word "height" (phonemically transcribed /haɪt/) would be pronounced by English speakers from the US Mid-West, Southern US, Central and Western Canada, and Australia (in 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 may 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" as "w" (/r/ as [w]). Again, although some of these features may be salient to untrained listeners, a practitioner trained in auditory phonetics may also be able to notice and systematically label smaller idiolectal differences.

In the auditory approach, the practitioner 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.

Use of an auditory-only approach is relatively uncommon. Many practitioners also consider some acoustic-phonetic measurements. They thus use an *auditory-acoustic-phonetic approach*. The measured acoustic-phonetic values are often used to make graphical plots, such as F1-F2 plots (see Figure 17 in section **[99.460]**), and the graphical plots are visually compared. In addition to

comparing the plotted values from the known- and questioned-speaker recordings, practitioners may also compare them with plots of values from recordings of *foil speakers*. Foil speakers are speakers who sound broadly similar to the voice on the questioned speaker recording (or to the known speaker). They could be a sample of the relevant population, but the number of speakers used may be too small to be considered a representative sample.

A conclusion as to the strength of evidence is based on consideration of auditory perception and of the measured acoustic-phonetic values. The *auditory-acoustic-phonetic* approach is a *non-statistical* approach: statistical models are not used to analyse the measured values. In section **[99.700]** we discuss the *acoustic-phonetic statistical approach*.

French et al. (2010) provided a list of features commonly considered in auditory-acoustic-phonetic analyses:

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

2. Intonation, ...

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

4. Articulation rate.

5. Rhythmical features.

6. Connected speech processes such as patterns of assimilation and elision [coarticulation].

7. A large set of consonantal features, including energy loci of fricatives and plosive bursts, durations of nasals, liquids [e.g., /r/ and /l/], and fricatives in specific phonological environments, voice onset time of plosives, presence/absence of (pre-)voicing in lenis plosives, and discrete sociolinguistic variables.

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

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

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

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

[99.670] Evaluation

The auditory and auditory-acoustic-phonetic approaches' reliance on experience-based subjective decisions mean that they are not well suited for use within the new paradigm.

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-speaker recordings, and also from speakers in a sample of the relevant population, then it would be possible to calculate a likelihood ratio. Elliot (2002) and Kirkland (2003) (unpublished Master's theses) both calculated likelihood ratios on the basis of frequencies of the occurrence of perceptually different pronunciations of a number of phonemes.

The majority of features used in the auditory approach are intrinsically subjective and qualitative – sounds are fitted into boxes according to the phonetician's perception of them – making it difficult to calculate quantitative measures of similarity and typicality. Even if relative frequencies for perceptually different pronunciations can be calculated, more fine-grained 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. The practitioner would either subjectively estimate values for the numerator and denominator of the likelihood ratio, or provide a qualitative verbal expression such as "based on my training and experience I believe that the properties I heard on the questioned-speaker recordings are much more likely to occur if they were produced by the known speaker than by some other speaker selected at random from the relevant population".

Theoretically it would be possible to measure the accuracy and precision of a practitioner of the auditory or auditory-acoustic-phonetic approach by having them provide strength of evidence responses to a large number of pairs of test samples (see section **[99.290]ff**).

A small-scale evaluation including practitioners of the auditory-acoustic-phonetic approach was reported in Cambier-Langeveld (2007). The evaluation included five auditory-acoustic-phonetic submissions, four acoustic-phonetic submissions, two automatic submissions, and one spectrographic or auditory-spectrographic submission. There were ten test pairs. The evaluation did not use performance metrics consistent with the new paradigm, and most participants did not report likelihood ratio values. Results were summarised in terms of whether they gave more support to the same-speaker hypothesis or more-support to the different-speaker hypothesis, without considering the strength of that support. The summary also indicated when a participant declined to give a strength of evidence statement, either because the recordings were of too poor quality or too short duration and therefore no analysis was attempted (labelled "reject"), or when a participant reported the results were "inconclusive". On average, as a group, practitioners of the auditory-acoustic-phonetic approach were willing to give responses other than "reject" or "inconclusive" to a larger percentage of the test pairs (88%, 35 of 40 opportunities to respond) than were practitioners of other approaches (65%, 39 of 60 opportunities to respond), but made a larger percentage of errors on the test pairs for which they did indicate support for either the same- or the different-speaker hypothesis (11%, 4 of 35 responses) than did practitioners of other approaches (3%, 1 of 39 responses). Indicating support for the contrary to fact hypothesis was counted as an error, irrespective of the strength of support indicated. Given the small size and other constraints on the evaluation, one should be cautious about generalising these results.

Human Assisted Speaker Recognition (HASR) evaluations were run by the National Institute of Standards and Technology (NIST) in 2010 and 2012 (Greenberg et al., 2010). The conditions of the evaluation were not intended to be forensically realistic. The results were not assessed using performance metrics appropriate for forensic application, and there were a small number of test pairs. Most participants only provided responses to a set of 15 test pairs, not to a larger set of 150

test pairs that was also made available. This was likely due to the practical difficulty of testing systems which require substantial human input for each test pair – Schwartz et al. (2011) reported taking about 8 hours to process each test pair. Automatic systems outperformed or performed at about the same level as naïve listeners (Ramos et al., 2011; Matějka et al., 2012; González-Hautamäki et al., 2013; also see Alexander et al., 2004; Hautamäki et al., 2010; Fernández Gallardo, 2014), and expertly operated auditory-acoustic-phonetic systems had about the same or a better level of performance compared to automatic systems (Schwartz et al., 2011; Saeidi & van Leeuwen, 2012). Given the small size of the test set, that the tests were not intended to be forensically realistic, and other constraints on the evaluation, one should be cautious about generalising these results.

[99.675] Further reading

Descriptions of auditory and auditory-acoustic-phonetic approaches can be found in Hollien (2002, 2016), Hollien et al. (2016), Jessen (2008, 2012), Nolan (1997, 2005), Rose (2002, 2006), and Zhang (2009).

[99.680] Spectrographic approach

[99.681] Description

The *spectrographic approach* is based on visual inspection of spectrograms (see Figure 18 in section **0**). The spectrographic approach is also known as *voicegram identification*, and as *voiceprinting*. It is typically combined with listening, resulting in an *auditory-spectrographic approach* (aka *aural-spectrographic approach*).

Protocols for performing auditory-spectrographic forensic voice comparison were developed by the *Federal Bureau of Investigation* (FBI), the *International Association of Voice Identification* (IAVI), the *International Association for Identification* (IAI) (the IAVI became part of the IAI in 1980), and the *American Board of Recorded Evidence* (ABRE) (ABRE was a later break-away from the IAI). The FBI ceased using the spectrographic approach in 2011 (personal communication from Hirotaka Nakasone, Senior Scientist, Digital Evidence Section, FBI, November 2011). 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. In the People's Republic of China, use of the spectrographic approach is recommended in the guidelines issued by the Ministry of Justice and the Ministry of Public Security (Cao et al., 2013; Zhang & Morrison, 2017).

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

The known 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 speaker on the questioned-speaker recording, but is someone who 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-speaker 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 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 (the same procedure may be used in an auditory-only approach). 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:

a. General formant shaping and positioning. ...

b. Pitch striations. ...

c. Energy distribution. ...

d. Word length. ...

e. Coupling. ... [i.e., nasality]

f. Other. Plosives, fricatives, and inter-formant features ... inhalation noise, repetitious throat clearing, or utterances like "um" and "uh" ...

And to auditorily compare:

a. Pitch. ...

b. Stress/Emphasis. ...

c. Rate. ...

d. Disguise. ...

e. Mode. ... [i.e., abruptness of voicing onset]

f. Psychological state. ...

g. Speech defects. ...

h. Vocal quality. ... [related to laryngeal activity]

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

In contrast, Poza & Begault (2005) recommend a gestalt approach (an immediate perception of the whole pattern).

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).

[99.690] Evaluation

From the 1960s to the 1990s there was a great deal of controversy over the spectrographic approach. Some proponents of the spectrographic approach made extravagant unproven claims about the accuracy of the procedure, claiming near infallibility. It was also claimed that real-world performance would be better than performance in controlled laboratory experiments (Tosi *et al*., 1972), the latter claim became known as the *Tosi Extrapolation*. This lead to criticism and an increasingly vociferous debate; see, for example, Hollien's criticisms of the spectrographic approach and its practitioners (Hollien, 1990, ch. 10; 2002, pp. 24–25, ch. 6), and Koenig's response (Koenig, 2002). 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, the International Association for Forensic Phonetics and Acoustics (IAFPA), which is dominated by practitioners of auditory-acoustic-phonetic approaches, passed a resolution in 2007 stating that "The Association considers this approach [voiceprinting] 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.

Gruber & Poza (1995, §6) summarised the objections to the spectrographic approach as follows:

> 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.

And added (§8):

> 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.

Many of the criticisms of the auditory-spectrographic approach could also be directed at the auditory-acoustic-phonetic approach (see Morrison, 2014).

The procedures for collecting the known-voice recording have also been criticised, specifically that a known-speaker recording of an innocent speaker who is a good mimic could end up being identified as the source of the voice on the questioned-speaker recording (Gruber & Poza, 1995, §63, §70; Rose, 2002, pp. 112–113). It should be noted, however, that in the ABRE protocol the known-speaker is never asked to directly mimic the questioned-speaker recording, and if there are foil speakers they are recorded under the same conditions. Practitioners of other approaches usually do not make recordings of the known-speaker specifically for the forensic voice comparison, and instead rely on existing recordings, such as police interviews or telephone calls made from jail, without requiring that they contain the same phrases spoken in the same way as on the questioned-voice recording.

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 or early 1960s using an analogue technology which was itself developed in the 1940s (Potter et al., 1947; Joos, 1948). The analogue devices were superseded by specialised digital hardware in the 1980s, which were in turn superseded 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 that information.

From the perspective of the new paradigm, the spectrographic approach suffers from the same drawbacks as the auditory and auditory-acoustic-phonetic approaches: Conclusions are subjective and experience-based rather than based on quantitative 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 an empirical evaluation of a practitioner's performance before allowing them to testify in court (a recommendation to do this was made by Gruber & Poza, 1995, §61).

[99.695] Further Reading

The spectrographic / auditory-spectrographic approach is described in Kersta (1962), Tosi (1979), and National Research Council (1979). Reviews of the controversy around its use include Gruber & Poza (1995), Solan & Tiersma (2003), Meuwly (2003a,b), Morrison (2014), and Lindh (2017).

[99.700] Acoustic-phonetic statistical approach

[99.701] Description

The *acoustic-phonetic statistical* 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 are used in empirical studies of speech production and speech perception, for example formant frequencies, fundamental frequency, and VOT.

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

Usually, comparable phonetic units are identified in both known- and questioned-speaker recordings 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 allophones preceding nasals /m/, /n/, /ŋ/ – 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". 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-speaker recording is shorter than the known-speaker recording, so tokens of potentially usable phonetic units are first identified and marked in the questioned-speaker recording. If the questioned-speaker 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-speaker 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 usually applied to multiple phonetic units from the same voice recordings.

It is also possible to use acoustic-phonetic-type measurements without explicit use of phonetic units. For example, formant frequencies or 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.

[99.710] Evaluation

The acoustic-phonetic statistical approach is suited for use within the new paradigm: The acoustic measurements made are relatively objective, the numeric data generated can be used to calculate likelihood ratios, and tests of validity and reliability can be 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 to produce posterior probabilities for a closed set of speakers.

A disadvantage of the acoustic-phonetic statistical approach 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 recordings included in a sample of the relevant population. Since test pairs also need to be measured, empirical testing is practically difficult (see section **[99.662]**).

In a series of comparisons of the performance of acoustic-phonetic statistical systems and automatic systems under condition approaching or reflecting casework conditions, we found that the acoustic-phonetic statistical systems performed poorly compared to the automatic systems, and that combining the two generally led to limited improvement (sometimes no improvement) over an automatic system alone. The automatic systems also required much less investment of human labour. The acoustic-phonetic features tested were formant trajectories, fundamental frequency, and nasal spectra. See Enzinger et al. (2012), Zhang et al. (2013), Zhang & Enzinger (2013), Enzinger (2014, 2016 ch. 5), Enzinger & Kasess (2014).

[99.715] Further reading

Descriptions of the acoustic-phonetic statistical approach can be found in Rose (2006, 2013, 2017) and Lindh (2017).

[99.720] Automatic approach

[99.721] Description

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., intelligence and security applications. 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 whereas in a security system, e.g., a telephone banking systems that uses voice recognition instead of a password, the task is to automatically make a same-speaker or different-speaker decision, for forensic applications the task is to produce an interpretable likelihood ratio.

As with the acoustic-phonetic approach, the automatic approach is based on quantitative measurements of acoustic properties of speech. The resulting numbers are invariably used as input to statistical models. Because the measurements and statistical models operate automatically, they can quickly process large numbers of files, and therefore can easily process hundreds or thousands of test pairs.

Although the system is automatic, there is still a very important role for the forensic practitioner. The forensic practitioner has to decide what question will be addressed, including what constitutes the relevant population, and the forensic practitioner has to select appropriate data for training and testing the automatic system. The forensic practitioner also makes sure that what is entered into the system is actually recordings of the speakers of interest, and not other speakers or noises. An automatic system is a tool to help a forensic practitioner perform a forensic voice comparison analysis, not a replacement for the forensic practitioner. Naïve use of an automatic system could lead to highly misleading results.

Common features measured in an automatic system are *mel-frequency cepstral coefficients* (MFCCs). MFCCs characterise the shape of the spectrum. The spectrum is usually measured within a frame of length 20 ms, and this frame is advanced in steps of 10 ms. Measurements are made over all parts of the recording during which the speaker of interest is speaking. This results in a long series of sequential MFCC measurements. Around 14 coefficients are obtained in every frame (the exact number varies from system to system). This results in a much more detailed characterisation of the speech spectrum than measurements of fundamental frequency plus two or three formants. MFCCs are usually supplemented with derivative measurements, called *deltas*, that characterise how fast the coefficient values are changing over time. 14 MFCCs plus 14 deltas would mean 28 values every 10 ms. These values are then used as input to statistical models. Common statistical models currently in use are *Gaussian mixture model - universal background model* (GMM-UBM) and *i-vectors* plus *probabilistic linear discriminant analysis* (PLDA). Automatic systems usually also incorporate statistical techniques for addressing mismatches in speaking styles and recording conditions between the known- and questioned-speaker recordings. Poor quality recordings and substantial mismatches between known- and questioned-speaker recordings still lead to poorer performance, but this is substantially mitigated by *mismatch compensation techniques*.

Traditionally, automatic systems do not explicitly exploit acoustic-phonetic information such as phoneme categories, but some systems exploit acoustic phonetic information, for example, by using an automatic speech recognition system to divide the speech signal up into phoneme size units. Some practitioners of acoustic-phonetic statistical approaches use MFCCs instead of f0 and formant measurements. The boundary between acoustic-phonetic statistical and automatic approaches is therefore fuzzy.

[99.730] Evaluation

The automatic approach is well suited for use within the new paradigm: The acoustic measurements made are objective, the numeric data generated can be used to calculate likelihood ratios, tests of validity and reliability can be conducted, and testing is practically easier than for acoustic-phonetic statistical systems.

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 and an automatic speaker-recognition system developed for some other purpose. The latter sort of system usually makes a yes/no decision. Most researchers working on automatic speaker recognition are not knowledgeable about forensic science. Some people may mistakenly believe that non-forensic automatic speaker recognition systems can be used as-is for forensic work.

A great advantage of the automatic approach to forensic voice comparison is that it is automatic, and can therefore analyse large amounts of data with little concern about human labour costs. In evaluations of automatic forensic voice comparison systems, the primary constraint on testing is the amount of suitable test data available.

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.

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. 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 population sample and test data.

Papers reporting the results of testing forensic voice comparison systems under (relatively) forensically realistic conditions include: Solewicz et al. (2012), van der Vloed et al. (2014), Enzinger & Morrison (2015), Enzinger (2016) Ch. 4, Enzinger et al. (2016), van der Vloed (2016), Zhang et al. (2016), Enzinger & Morrison (2017), Marks (2017), Silva & Medina (2017). Evaluations for different automatic systems under conditions reflecting those of one real forensic case are being published in a virtual special issue of the journal *Speech Communication* <http://www.sciencedirect.com/science/journal/01676393/vsi/10KTJHC7HNM>.

[99.735] Further reading

Descriptions of the automatic approach to forensic voice comparison can be found in Ramos Castro (2007), Becker (2012), Enzinger (2016), Morrison & Enzinger (2018). For a review of automatic speaker recognition in general, see Hansen & Hasan (2015).

LEGAL ADMISSIBILITY OF FORENSIC VOICE COMPARISON

[99.750] Review

We would argue that admissibility should be considered on a case by case basis, and that just because one forensic report using a particular approach is ruled admissible does not mean that another forensic report using the same approach (in the same or a different case) should necessarily also be admissible. *Mutatis mutandis*, if one is ruled inadmissible the other should not necessarily be ruled inadmissible. However, it simplifies matters to discuss admissibility in terms of different approaches, and although courts sometimes explicitly state that their ruling applies to the particular instance only, courts have often considered admissibility at an approach level.

Below we briefly summarise the situation regarding admissibility of forensic voice comparison in several common law jurisdictions.

In the United States, from the 1960s to the 80s testimony based on spectrographic or auditory-spectrographic approach was often proffered in court proceedings, but the number of cases in which it was proffered declined and fell to a trickle by the 1990s. Based on published rulings, the rate of admission appears to have been a little greater than the rate of exclusion (see Faigman et al., 2015). In *United States v Robert N. Angleton*, 269 F.Supp. 2nd 892 (S.D. Tex. 2003), in an admissibility hearing held under Federal Rule of Evidence (FRE) 702 and the criteria established in *William Daubert et al. v Merrell Dow Pharmaceuticals Inc.*, 509 US 579 (1993), the court ruled an auditory-spectrographic approach inadmissible. The court found that demonstration of an adequate level of scientific validity was lacking. Based on published rulings, no attempt to admit a spectrographic or auditory-spectrographic approach appears to have survived an FRE 702 - *Daubert* challenge since then. In *United States v Ali Ahmed, Madhi Hashi, & Muhamed Yusuf*, No. 12-661 (E.D.N.Y.), testimony was proffered which was in part based on an automatic approach, but which was combined with auditory and acoustic-phonetic approaches. In 2015 an FRE 702 - *Daubert* admissibility hearing was held, but before the judge ruled on the matter the case was resolved via a negotiated plea deal.

In New South Wales, Australia, in *R v Gilmore* [1977, 2 NSWLR 935] the court ruled testimony based on an auditory-spectrographic approach admissible. The decision was based in part on the fact that such testimony had been ruled admissible by a number of courts in the United States in the early to mid 1970s. In 2012 the admissibility of testimony based on an auditory-spectrographic approach was challenged in *R v Ly* [NSW District Court, 2010/295928]. Notwithstanding the passage of time, changes in available technology, and the changes in the US regarding admissibility of the spectrographic approach, the court ruled that *Gilmore* was precedential and the testimony was therefore admissible. Testimony based on an acoustic-phonetic statistical approach was admitted in *R v Hufnagl* [NSW District Court, 2008], see Rose (2013).

In Northern Ireland, the appeal court in *R v O'Doherty* [2002] NICA 20 / [2003] 1 Cr App R 5, ruled an auditory-only approach inadmissible, but auditory-acoustic-phonetic approaches admissible. It was reported that most practitioners considered auditory-only approaches to be unreliable.

In England & Wales, the appeal court in *R v Robb* [1991] 93 Cr App R 161 ruled an auditory-only approach admissible. In *R v Flynn and St John* [2008] EWCA Crim 970, the appeal court opined

that *Robb* was still precedential and that courts in England & Wales should not follow the example set in Northern Ireland in *O'Doherty*. The opinion in *Flynn* was echoed in the appeal court ruling in *R v Slade et al.* [2015] EWCA Crim 71. The appeal court in *Slade* ruled testimony based on an automatic approach inadmissible (it had been proffered as new evidence). The court was not satisfied with the quantity and quality of the data used to train and test the automatic system. Nor was it satisfied with the empirically demonstrated level of performance of the system. Ironically, testimony based on auditory-only and auditory-acoustic-phonetic approaches had been admitted at trial despite the fact that they had not been empirical tested (admissibility of these approaches does not appear to have been challenged at any point in the proceedings).

[99.760] Further reading

Admissibility of forensic voice comparison in the United States is reviewed in Morrison & Thompson (2017). Admissibility of forensic voice comparison in the England & Wales (with an excursion to Northern Ireland) is reviewed in Morrison (2018).

EXAMPLES OF FORENSIC VOICE COMPARISON

[99.770] Introduction

This section provides two examples of forensic voice comparison conducted within the new paradigm. Each example describes the analysis performed in an actual case.

The first example is based on a case in which the questioned speaker was one of two possible speakers. The second example is based on a more common sort of case in which the questioned speaker was either the known speaker or another speaker from a large relevant population.

[99.780] Example 1: Speaker *A* versus Speaker *B*

In a 2016 civil case in China, the plaintiff had spoken on her mobile telephone to a woman, and had recorded the conversation on her telephone. The recording, which was about 25 minutes long, became a questioned-speaker recording. The plaintiff stated that she believed she had spoken with the respondent (hereinafter Speaker *A*), whereas the respondent stated that the plaintiff had spoken with the respondent's sister (hereinafter Speaker *B*).

This case and the analyses conducted are described in greater detail in Zhang et al. (2016).

[99.790] Hypotheses

In this case we adopted the following hypotheses:

1. The voice on the questioned-speaker recording is Speaker *A*.
2. The voice on the questioned-speaker recording is Speaker *B*.

This is a relatively unusual situation: The hypotheses were overtly provided by the parties, and rather than the relevant population being a large group of speakers, it consisted of a single speaker.

[99.800] Relevant data

Both Speaker *A* and Speaker *B* were cooperative, as was the plaintiff. We were therefore able to have an officer of the court use the plaintiff's mobile telephone to record telephone conversations with each of Speaker *A* and Speaker *B*. Thus, we were able to obtain relevant data in the same speaking style and the same recording conditions as the questioned-speaker recording. Hence, there were no speaking style or recording condition mismatch. A limitation, however, was that we were only able to obtain 5 recordings of each speaker, each about 10 minutes long.

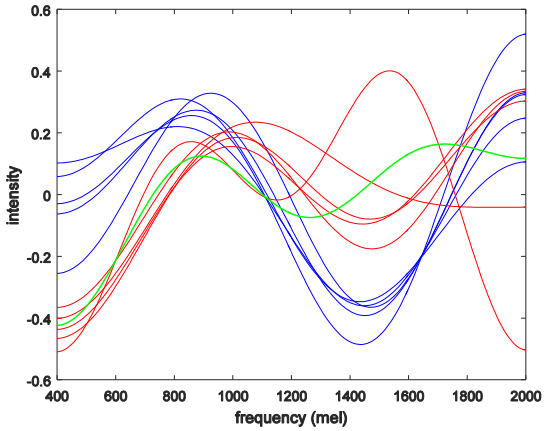
[99.810] Acoustic and statistical analysis

The acoustic and statistical analysis described below is only valid if one assumes that there is no substantial mismatch in recording condition and speaking style between the questioned-speaker recording and the Speaker *A* and Speaker *B* recordings.

In each recording, we manually located the speech of the speaker of interest. Over those portions of the recordings, we measured MFCCs once every 10 ms using 20 ms long windows (see section **[99.720]**). MFCCs were measured over the frequency range 300 Hz to 3.4 kHz (400 to 2000 mel).

Because there were only 5 recordings of each known speaker, we only used the 1st through 4th MFCC values. For each recording, we calculated the means of the 1st through 4th MFCC values for each recording. Each set of means is called a *mean vector*, in this case the mean vectors had 4 dimensions. The 4 dimensional mean vectors characterised the average smoothed spectrum of each recording, as plotted on Figure 21.

Figure 21. Average smoothed spectra of Speaker *A* recordings (blue lines), Speaker *B* recordings (red lines), and the questioned-speaker recording (thicker green line).



The 5 Speaker *A* mean vectors and the 5 Speaker *B* mean vectors can also be represented as clouds of points in a 4 dimensional space. We can't draw 4 dimensions, but Figure 22 plots the clouds of points in the first two dimensions (this type of plot is called a *scatterplot*). There is within-speaker variability within each cloud, and between-speaker variability between the two clouds.

We applied a statistical analysis technique, called *linear discriminant analysis* (LDA). This reduced the 4 dimensional MFCC space into a 1 dimensional line. This was a line in the direction which maximised the between-speaker variability in the 4 dimensional space. The direction of this line in the first two MFCC dimensions is plotted as the dashed line in Figure 22 (only the direction of the line matters, we could have drawn any line parallel to the one in Figure 22). The direction of the line is closer to the direction of the 1st MFCC dimension than the 2nd MFCC dimension because the ratio of between-speaker variability to within-speaker variability is greater in the 1st MFCC dimension. The 5 Speaker *A* mean vectors and the 5 Speaker *B* mean vectors are converted from points in the 4 dimensional space to points on the 1 dimensional line (see Figure 23). We will henceforth refer to the values of the points on the 1 dimensional line as *LDA values*.

Figure 22. Scatterplot of mean 1st and 2nd MFCC values from Speaker *A* recordings (blue circles), Speaker *B* recordings (red triangles), and the questioned-speaker recording (green cross). The dashed line is the 1 dimensional line created by linear discriminant analysis.

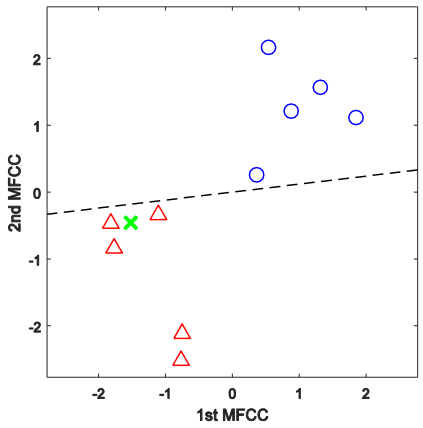
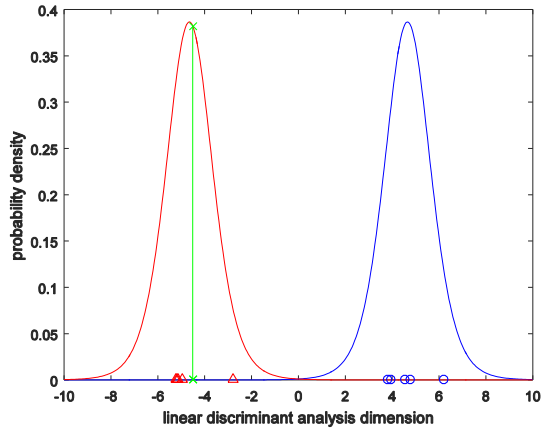


Figure 23. Speaker *A* model (blue curve) and Speaker *B* model (red curve). The green crosses indicate the likelihoods for the questioned-speaker recording.



Note that LDA was used only to reduce the number of dimensions, not to classify the speakers. In general, LDA is a technique which can be used to help compensate for mismatches in speaking styles and recording conditions. If one has multiple recordings of each speaker under different conditions, a substantial part of the within-speaker variability will be due to differences in speaking style and recording conditions. LDA minimises within-speaker variability and maximises between-speaker variability. Some between-speaker information may be lost in the process, but this is usually outweighed by the reduction in unwanted variability due to mismatched conditions.

Instead of fitting Gaussian distributions to the LDA values (see sections **[99.220]** and **[99.230]**), we fitted *t* distributions. This is recommended when the amount of data used to train the model is small. *t* distributions are similar to Gaussian distributions, but when the amount of data is small they are wider than Gaussian distributions trained on the same data. This leads to the calculated value of the likelihood ratio being closer to 1 than it would be if Gaussian distributions were used. To train the *t* distribution models, we calculated the mean of the Speaker *A* LDA values, and the mean of the Speaker *B* LDA values, and we used LDA values from both speakers to calculate a single pooled standard deviation which was used for both models. The probability density distributions of the models are shown in Figure 23.

To calculate the likelihood ratio, we first took the mean vector of the 1st through 4th MFCC values from the questioned-speaker recording, and transformed the 4 dimensional mean vector to a 1 dimensional LDA value. We evaluated the likelihood for the LDA value given the Speaker *A* model and the likelihood for the LDA value given the Speaker *B* model (see Figure 23), and divided the former likelihood value by the latter (see section **[99.230]**).

[99.820] Empirical testing

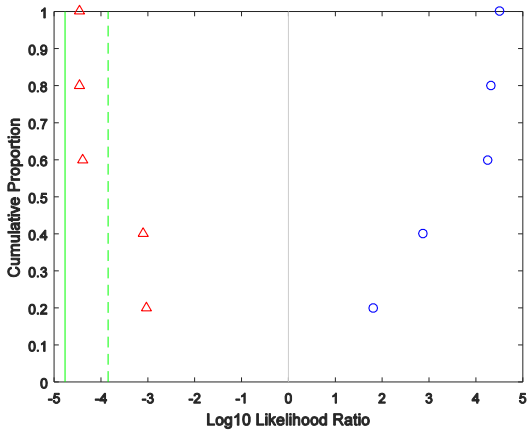
Before calculating the likelihood ratio for the questioned-speaker recording, we tested the performance of the system.

To assess validity, we used a *cross-validation* procedure. We held out one of the Speaker *A* recordings, and used the data from all the other Speaker *A* recordings and the Speaker *B* recordings to train the statistical models (linear discriminant analysis and *t* distributions). We then repeated this with each of the other Speaker *A* recordings, and with each of the Speaker *B* recordings. Thus we avoided training and testing on the same data. We calculated 5 likelihood ratio values for which we knew the test speaker was Speaker *A*, and 5 likelihood ratio values for which we knew the test speaker was Speaker *B*. These likelihood ratio values are shown in the Tippett plot in Figure 24 (we plotted symbols at each calculated likelihood ratio value rather than drawing a line between these points). The resulting C_{llr} value was very low: 0.003. Both the Tippett plot and the C_{llr} value indicate very good performance (see sections **[99.330]** and **[99.300]**).

To assess the reliability of the likelihood ratio value we calculated for the questioned-speaker recording (see section **[99.310]**), we used a *Monte Carlo simulation* procedure. We used the means of the 1st through 4th MFCC values of the 5 Speaker *A* recordings, and the means of the 1st through 4th MFCC values of the 5 Speaker *B* recordings to build Gaussian models which we then used in conjunction with a *random number generator* to create simulated data. We generated 1000 sets of 5 simulated Speaker *A* 1st through 4th MFCC mean vectors and 5 simulated Speaker *B* 1st through 4th MFCC mean vectors. Linear discriminant analysis models and *t* distribution models were trained using each set of simulated data, and a likelihood ratio calculated for the original mean vector of 1st through 4th MFCC values from the questioned-speaker recording. Using this procedure, we calculated 1000 likelihood ratio values. We then ranked the 1000 values, and found the 100th highest values. 90% of the 1000 simulated likelihood ratio values were further from a likelihood ratio of 1 than was the 100th value.

The latter is a frequentist approach to dealing with reliability, and the one we used for the actual case report. An appendix to Zhang et al. (2016) describes a Bayesian approach which we conducted for research purposes at a later date (see also Morrison & Poh, 2017).

Figure 24. Tippett plot. Speaker *A* likelihood ratio values shown as blue circles, and Speaker *B* likelihood ratio values shown as red triangles. The solid green line is at the likelihood ratio value calculated for the questioned-speaker recording, and the dashed green line is at the 100th likelihood ratio value from the Monte Carlo simulations.



[99.830] Conclusion

The calculated value for the likelihood ratio was approximately 1/60,000 (exact value 1/58,562). This is our best estimate for the strength of the evidence. Based on tests of the reliability of our system, we are 90% certain that the value of the likelihood ratio is at least approximately 1/7000 (exact value 1/7036).

Note that since we arbitrarily put the Speaker *A* hypothesis in the numerator of the likelihood ratio and Speaker *B* hypothesis in the denominator, and the likelihood of the evidence given the Speaker *B* hypothesis is larger than the likelihood of the evidence given the Speaker *A* hypothesis, the calculated likelihood ratio value is less than 1, and we have expressed it above as a fraction. Below we reverse the Speaker *A* and Speaker *B* hypotheses in the wording so as to be able to express the likelihood ratio as a value larger than 1.

We estimate that the probability of obtaining the measured acoustic properties of the voice on the questioned-speaker recording is 60,000 times higher if it were produced by Speaker *B* than if it were produced by Speaker *A*. We are 90% certain that it is at least 7000 times higher.

Let us assume the trier of fact is conservative and decides to use the value of 7000 rather than 60,000. Let us further assume that the trier of fact chooses to use the normative logic of Bayes Theorem (sections **[99.150]** and **[99.160]**). Whatever the trier of fact's prior odds, the trier of fact should multiply those prior odds by 7000 to obtain the posterior odds. Whatever the trier of fact's prior belief as to the relative probabilities that the questioned speaker was Speaker *B* versus that the questioned speaker was Speaker *A*, after hearing the strength of evidence statement, the trier of fact should believe that the relative probability that the questioned speaker is Speaker *B* as opposed to Speaker *A* is 7000 times higher than they believed it to be before.

Morrison & Poh (2017) reported on reanalyses using more conservative statistical procedures applied to the same data, which produced likelihood ratios in the range 1/5 to 1/2,500.

[99.840] Example 2: Known speaker versus a large relevant population

In a 2012–2013 criminal case in Australia, a fraud was perpetrated by a person who called a financial institution. The call was recorded, and became a questioned-speaker recording. A suspect was arrested and interviewed by the police. The recording of that interview became the known-speaker recording.

This case and the analyses conducted are described in greater detail in Enzinger et al. (2016). The statistical models used in the case and described below were GMM-UBM (see section **[99.720]**). Enzinger (2016) Ch. 4 also describes an i-vector PLDA analysis which was subsequently conducted as a research exercise.

[99.850] Hypotheses

The speaker on the questioned-speaker recording was clearly an adult male who spoke English with an Australian accent. In this case we adopted the following hypotheses:

1. The voice on the questioned-speaker recording is that of the known speaker (the suspect).
2. The voice on the questioned-speaker recording is not that of the known speaker. It is the voice of some other speaker from the relevant population. The relevant population is adult male speakers of Australian English.

We actually further restricted the relevant population by excluding speakers who sounded quite different from the voice on the questioned-speaker (see section **[99.180]**).

[99.860] Relevant data

Recordings for inclusion in the sample of the relevant population were taken from a database of recordings of 500+ Australian English speakers collected specifically for use in forensic voice comparison research and casework (the database is available from <http://databases.forensic-voice-comparison.net/>). Most speakers were recorded in multiple recording sessions, each session separated by about a week. In each session, the speakers performed different tasks which elicited different speaking styles. The speaking styles used for this case came from a task in which the speakers had to exchange information (including numbers and letters) over the telephone, and from a simulated police interview task. These speaking styles were similar to those of the questioned-speaker recording and the known-speaker recording respectively.

The questioned-speaker recording was of a landline telephone call made to the call centre of a financial institution. At the call centre end there was background office noise, including background speech from a large number of speakers (babble) and typing noised (see section **[99.600]**). The audio recording was also saved in a lossy compressed format (see section **[99.610]**).

The known-speaker recording was recorded in a small room with hard walls. This resulted in substantial reverberation (echoes) on the recording (see section **[99.600]**). There was also

background noise from the ventilation system. The audio was saved in a different compressed format.

The recordings from our database were high-quality recordings. We used signal processing techniques to simulate the conditions of the known- and questioned-speaker recordings. This included passing recordings through filters to simulate telephone transmission, compressing and decompressing recordings using the same codecs as had been used on the known- and questioned-speaker recordings, adding reverberation based on an estimate of the reverberation properties of the interview room (its *impulse response*), and adding background noise. The background noise was extracted from portions of the known- and questioned-speaker recordings when no one was speaking and added to the database recordings at the same signal to noise ratio as in the known- and questioned-speaker recordings. Audio recordings providing examples of the results of simulating the case conditions are available at <http://expert-evidence.forensic-voice-comparison.net/#audio>. This process gave us a set of recordings that came from a sample of speakers representative of the relevant population, and that reflected the speaking styles and recording conditions of the known- and questioned-speaker recordings in the case under investigation.

[99.870] Acoustic and statistical analysis

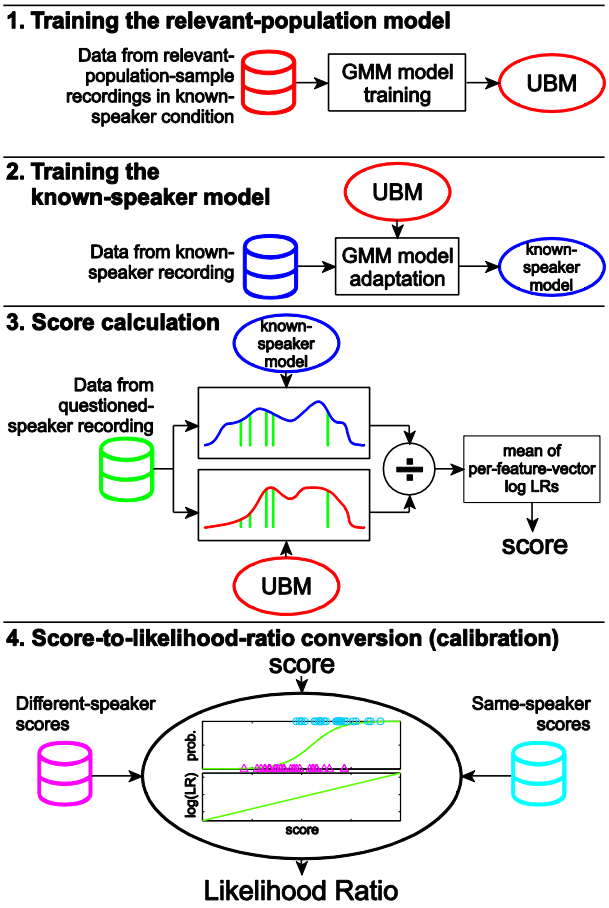
The portions of the known- and questioned-speaker recordings corresponding to the speech of the speaker of interest were manually located. The speech of the interlocutor (including when overlapping with the speech of the speaker of interest), periods of silence, and any transient noises were excluded. This left more than 14 minutes of known-speaker speech, but only 46 seconds of questioned-speaker speech. The database recording had been previously processed in a similar manner (except that an automated approach had been used initially and subsequently manually checked and corrected).

MFCCs were measured over the frequency range 300 Hz to 3.4 kHz. Since the questioned-speaker recording was a recording of a landline telephone transmission, it did not contain information outside this frequency range (see section **[99.610]**). Using 20 ms long windows, MFCCs + deltas were extracted every 10 ms during the speech of the speaker of interest in each recording (see section **[99.720]**). The 1st through 14th MFCC values plus their deltas were used to create 28 dimensional *feature vectors*.

A mismatch compensation technique known as *feature warping* was applied to the feature vectors. This statistical technique has been shown to improve system performance when there is a mismatch between the recording conditions of the known- and questioned-speaker recordings.

Figure 25 provides a schematic of the statistical modelling procedures, which consisted of GMM-UBM followed by logistic-regression calibration (see sections **[99.720]** and **[99.240]**).

FIGURE 25. Schematic of the statistical modelling approach.



Mismatch-compensated feature vectors from recordings of 44 speakers in the sample of the relevant population were used to train a relevant population model. The recordings used for training the UBM reflected the speaking style and recording conditions of the known-speaker recording. Data from these particular 44 speakers were used because they had only participated in one recording session, and we wanted to use data from the speakers with multiple recording sessions later for training the calibration model and for empirical testing. In the GMM-UBM procedure, the model of the relevant population is known as the *Universal background model* (UBM). The UBM was a Gaussian mixture model (GMM) with 512 Gaussian components.

A known-speaker model was trained using the data from the known-speaker recording. Because the amount of data from one speaker is relatively small, rather than training the model from scratch, it was adapted from the UBM. This is standard in the GMM-UBM procedure. Note that the known-speaker model and the UBM were both trained with data reflecting the known-speaker recording conditions and hence both have the same mismatch with the questioned-speaker recording.

A total of 4137 feature vectors could be extracted from the questioned-speaker recording. For each mismatch-compensated feature vector from the questioned-speaker recording, the likelihood of the known-speaker model and the likelihood of the UBM was assessed. The first likelihood was divided by the second to calculate a likelihood ratio. We then had 4137 per-feature-vector likelihood ratio values.

We calculated the log of each of the 4137 per-feature-vector likelihood ratio values, then calculated their mean. The latter value is known as a *score*. A score captures information about the similarity of the voice on the questioned-speaker recording with respect to the known speaker, and its typicality with respect to the relevant population, but the value of the score is not directly interpretable as a likelihood ratio. Scores have to be converted to likelihood ratios, a process which is also known as *calibration* (see section **[99.240]**). To train a calibration model, we took pairs of recordings from 61 speakers from the sample of the relevant population (speakers whose data had not been used in training the UBM). In each pair of recordings, one member of the pair reflected the speaking style and recording conditions of the known-speaker recording, and the other reflected the speaking style and recording conditions of the questioned-speaker recording. Mismatch-compensated feature vectors from the first member of the pair were used to train a speaker model, and a score was calculated using this speaker model, the UBM, and the mismatched-compensated feature vectors from the second member of the pair. A large number of pairs were same-speaker pairs, and a large number were different-speaker pairs. This resulted in a set of same-speaker scores and a set of different-speaker scores. The same-speaker scores and different-speaker scores were used to train a *logistic regression* model (this is a commonly used calibration model, see section **[99.240]**). The logistic regression model was then used to convert the score from the questioned-speaker recording to a likelihood ratio.

[99.880] Empirical testing

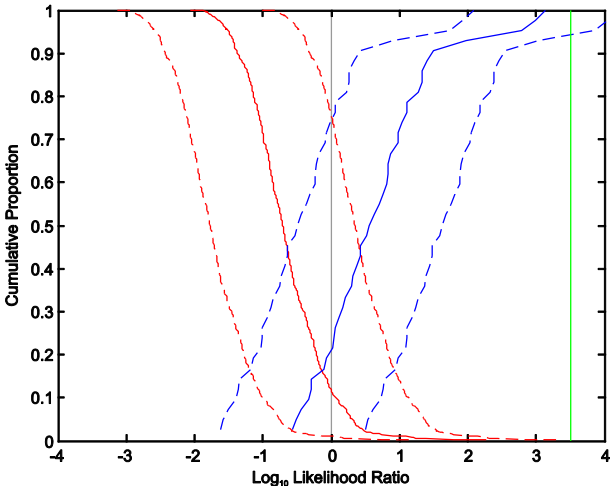
Recordings from another 61 speakers from the sample of the relevant population were used to test the system. No recordings from these speakers had been used to train the system. Pairs of recordings were used, one member of each pair reflecting the conditions of the known-speaker recording and the other reflecting the conditions of the questioned-speaker recording. A large number of pairs were known to be same-speaker pairs and a large number known to be different-speaker pairs. This resulted in a set of same-speaker likelihood ratio values and a set of different-speaker likelihood ratio values. These were used to make the Tippett plot in Figure 26 and to calculate a C_{llr} value, which was 0.523 (see sections **[99.330]** and **[99.300]**).

The empirical tests of system performance also included a frequentist assessment of precision (see section **[99.310]**). The 95% credible interval was ±1.06 orders of magnitude. This is shown on the Tippett plot as the dashed lines to the left and right of the solid lines.

Overall, system performance was reasonable but not very good. The likelihood ratio value calculated for the known- and questioned-speaker recording was, however, far from 1 (see section **[99.890]**), so even allowing for relatively poor overall performance leading to the use of a more conservative value than the best estimate for the likelihood ratio, we think the performance was adequate. Whether the level of performance would have been sufficient to satisfy a judge at an admissibility hearing we do not know, because, after we submitted our report and before going to trial, the case was settled via a plea deal.

In hindsight, we would criticise the size of sample used to train the relevant-population model as too small. In subsequent research activities we have used the same data in a different way, using recordings from 105 speakers to train the UBM and using cross-validation on the recordings from the 61 test speakers to train the logistic-regression calibration model. We have also tested additional mismatch-compensation techniques, and tested i-vector PLDA systems. The best performance was obtained using an i-vector PLDA system which included mismatch compensation at both the feature level and the i-vector level (see Enzinger, 2016, ch 4).

Figure 26. Tippett plot with 95% credible interval. The vertical green line indicates the value of the likelihood ratio calculated for the comparison of the voices on the known- and questioned-speaker recordings.



[99.890] Conclusion

The calculated value for the likelihood ratio was approximately 3300 (exact value 3287). This is our best estimate for the strength of the evidence. Based on tests of the reliability of our system, we are 95% certain that the value of the likelihood ratio is at least approximately 425 (exact value 424).

We estimate that the probability of obtaining the measured acoustic properties of the voice on the questioned-speaker recording is 3300 times higher if it were produced by the suspect than if it were produced by another speaker from the relevant population. We are 95% certain that it is at least 425 times higher.

Let us assume the trier of fact uses the best estimate of the strength of evidence. Let us further assume that the trier of fact chooses to use the normative logic of Bayes Theorem (see sections **[99.150]** and **[99.160]**). Whatever the trier of fact's prior odds, the trier of fact should multiply those prior odds by 3300 to obtain the posterior odds. Whatever the trier of fact's prior belief as to the relative probabilities that the questioned speaker was the suspect versus that the questioned speaker was another speaker from the relevant population, after hearing the strength of evidence

statement, the trier of fact should believe that the relative probability that the questioned speaker is the suspect, as opposed to another speaker from the relevant population, is 3300 times higher than they believed it to be before.

SPEAKER RECOGNITION BY LAYPEOPLE

[99.910] Introduction

Forensic voice comparison performed by forensic practitioners has been the topic of this chapter so far. In contrast *speaker recognition by laypeople* refers to the general ability of a person with no special training to recognise a voice and identify the speaker. This has also been referred to as *non-technical speaker identification* and as *naïve speaker recognition*.

Sometimes 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 forensic practitioner cannot perform a forensic voice comparison.

In other instances audio recordings of both the known and questioned speaker are available and a forensic voice comparison performed by a forensic practitioner is potentially possible. Despite this, in some jurisdictions police officers and interpreters 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 speaker.

Below we describe the problems with allowing a layperson to do the work of a forensic practitioner. We begins by describing some basic differences between speaker recognition by laypeople and forensic voice comparison **[99.930]**, discuss some mistaken beliefs about speaker recognition by laypeople **[99.950]**, and describe true earwitnesses and speaker lineups **[99.960]**. This is followed by discussion of some factors that have been found to affect the accuracy of speaker recognition by laypeople **[99.970]ff**, and finally these factors are related to an example based on the testimony of a police officer in an actual case **[99.1040]ff**.

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

[99.930] Speaker recognition by laypeople versus forensic voice comparison by forensic practitioners

Speaker recognition by laypeople 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 often recognise them from their voice. There are, however, also instances when we do not recognise the voice of a caller. Perhaps the caller is someone we do not know, or perhaps someone we do know but we do not recognise their voice.

Sometimes our identification of a speaker is 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 a speaker.

Speaker recognition by laypeople 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 led them to a particular identification, as a lay person they are unlikely to have the vocabulary be able to give a detailed description of the voice.

In contrast, forensic voice comparison is (ideally) performed by forensic practitioners with extensive training in phonetics or speech processing, and training in evaluation of forensic evidence. Often the level of training is such that the practitioner has obtained a doctorate in a relevant field. A forensic practitioner working within the new paradigm makes use of relevant data, quantitative measurements, and statistical models to quantify the strength of evidence. They also empirically test the validity and reliability of their system under conditions reflecting those of the case under investigation.

[99.950] Mistaken beliefs about speaker recognition by laypeople

It is perhaps because they are aware of their own capacity to identify speakers 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 fingerprint evidence and would immediately call upon a forensic practitioner for assistance.

Yarmey et al. (2001) state:

> 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. (p. 284)
>
> 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. (p. 286)

Solan & Tiersma (2003) state:

> 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. (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. (p. 432)

Bull & Clifford (1999b, pp. 202–203) express similar sentiments.

Laub (2010) reviewed literature and conducted experiments comparing eyewitness and earwitness identification. A number of differences between the visual and aural modalities have been found. A key finding is that the accuracy of earwitness identification is substantially worse than that of eyewitness identification. Mock juries, however, were found to be insensitive to the existence of this difference, apparently giving equal weigh to both. If mock juries were not told about factors that research has found to lead to better or worse accuracy of earwitness identification or they

were told about them only in jury instructions, they were insensitive to these factors. They were sensitive to these factors, however, when they were told about them in expert testimony or closing arguments. This suggests that the effects of these factors on earwitness accuracy are not common knowledge, and that this is therefore a proper topic for expert testimony.

> Mistaken witness identifications (voice and visual) are two of the greatest causes of actual or possible wrongful convictions, warranting the inclusion of courtroom safeguards that have been found to increase awareness. (Laub, 2010, p. 98)

[99.960] True earwitnesses and speaker lineups

It is important to distinguish between speaker identification performed by an *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 speaker identification performed by someone listening to an audio recording related to the crime and comparing that with a recording of a suspect or with their experience of having heard the suspect speak.

As will be explained below, speaker recognition by laypeople is of unknown validity unless the individual listener can be tested under circumstances similar to those under which they identified the questioned voice. If the circumstances are such that earwitness testimony is all that is possible, earwitness testimony may still be of assistance to the trier of fact. But, the court should be apprised of the problems with its validity and should keep the various caveats in mind when considering the weight of such testimony.

In contrast, when audio recordings of the questioned and known voices are available, forensic voice comparison of demonstrable validity and reliability should be performed by a forensic practitioner. If the layperson identifying the speaker is not actually an earwitness, but instead someone who has listened to the audio recordings for the specific purpose of making a speaker identification, we do not believe there is any reasonable argument for admitting this in court in place of forensic voice comparison.

When forensic voice comparison is not possible, an earwitness may be asked to listen to a voice lineup. Usually, the earwitness first describes the voice of the offender to the investigating officer or to a forensic practitioner, e.g., male Midwestern accent, deep voice. A forensic practitioner prepares one of more sets of audio recordings. Each set consists of foil speakers who match the description given by the earwitness, and may or may not also include a recording of the suspect. Care should be taken so that the linguistic content, speaking style, or recording conditions do not make the suspect stand out compared to the foil speakers. Where the suspect does not match the earwitness's description (e.g., the suspect has a Brooklyn accent but the earwitness described a Boston accent, or the suspect has a New Zealand accent but the earwitness described an Australian accent), at least a substantial proportion of the foil speakers should sound like the suspect. Listeners who were not earwitnesses to the crime and who have never heard the suspect or the foil speakers can be used to screen the recordings to make sure the suspect does not stand out. A set of recordings is presented to the earwitness. We recommend presenting one speaker at a time, rather than to allowing the earwitness to go backwards and forwards between speakers. The earwitness has to say whether they recognise the voice or not, and, if they recognise it, say who they think it is. Once they have given a response, they hear the next speaker. The earwitness is told that the voice of the offender may or may not be in the set of recordings. The earwitness is not told how many recordings or speakers will be included in the set (this avoids increasing the

probability of the earwitness identifying a speaker because they know they are getting towards the end of the list and haven't picked anybody yet). We recommend that the earwitness be asked to listen to more than one set of recordings, with breaks in between sets (this helps the earwitness believe that it is possible that the offender may not be included in a set). If the suspect is included in more than one set, a different recording of the suspect should be in each set, and, to avoid the suspect standing out, some foil speakers must also be included in more than one set (different recordings of each foil speaker in each set). Likewise, more than one recording of the suspect, and more than one recording of at least some of the foil speakers may be included in a set. If so, the earwitness should be told that there may be one, more than one, or no recordings of the offender in a set (this decreases the probability that the earwitness will not identify a speaker because they already identified a speaker in a recording that was presented earlier). The recordings (and sets of recordings) should be presented in random order, ideally presented by a computer program. To avoid the potential for the person administering the lineup to influence the earwitness, the administrator should have no knowledge of whether the suspect is included in a set or not, and should not be able to hear the recordings or see the earwitness's responses.

Descriptions of voice lineup protocols can be found in various works, including: Hollien et al. (1995); Hollien (1996); Nolan & Grabe (1996); Laubstein (1997); Broeders & van Amelsvoort (1999, 2001); Nolan (2003); Jessen (2008); Zetterholm et al. (2012); Hollien et al. (2014); Sarwar et al. (2014); de Jong-Lendle et al. (2015). Butcher (1996) emphasises the danger of a poor choice of foils such that the suspect stands out.

[99.970] Validity of speaker recognition by laypeople

There are multiple factors which have been reported in the research literature as related to listeners' abilities to correctly identify speakers:

[99.980] Variability between listeners

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. One of the most recent studies to find large between-listener differences is Sørensen (2012).

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 idiosyncratic between-listener variation.

In order to assess the validity of an individual listener's 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 (see section **[99.290]**). 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 be impractical.

[99.990] Listener certainty

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 studies have 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 target voice (e.g., the offender voice) was included in the lineup and not when the target voice was not included, even though the listeners were told that the target voice may or may not be in the lineup. This is to say that when the target 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 target voice. For three of the four target speakers in van Wallendael et al. (1994), when the target speaker was not included in the lineup listeners always incorrectly identified one of the speakers in the lineup as the target.

That listeners confidently pick one of the speakers in the lineup when the target is actually absent from the lineup is rather worrying if one considers the situation where the police investigators are mistaken and their suspect is not the offender. The listener may pick the suspect's voice, particularly if it happens to be the voice in the lineup which sounds most similar to that of the offender. An earwitness may with a high degree of certainty identify the suspect's voice as the voice of the offender even though the suspect is not the offender (Zetterholm et al., 2012).

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 only in cases in which the target 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 (high correlation does not imply good calibration).

In voice lineups in which the target speaker was present, Sarwar et al. (2014) found that listeners were overconfident in their identification, i.e., their reported levels of certainty in the correctness of their identifications was greater than their actual correct identification rate. They were overconfident by an average of 14 to 38 percentage points depending on the test conditions. In voice lineups in which the target speaker was not present and the listener said they were not present, Zetterholm et al. (2012) found that listeners were underconfident in their degree of certainty that the target speaker was not in the lineup.

Solan & Tiersma (2003) state:

> 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. (p. 412)

Laub (2010) concluded that:

> Average jurors are sensitive to certain variables, such as witness confidence, which has little reliable relation to accuracy. (p. 94)
>
> mock jurors mistakenly relied on the confidence rating of the victim in her identification of the defendant. (p. 96)

Yarmey (1995, pp 802–803), Bull & Clifford (1999a, p 218), Rose (2002, p 101), Solan & Tiersma (2003, p 412), Yarmey (2004, p 275), Öhman et al. (2010, p 176), McDougall et al. (2015, p 267) all conclude that the degree of certainty expressed by an earwitness should not be taken as an indicator of whether their identification is correct.

[99.1000] Listeners' familiarity with speakers' voices

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 long 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 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 the term "familiar" has been used in legal settings 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 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.

Edmond & San Roque (2009) state:

> 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. (p. 31).

In some studies (see summaries in Bull & Clifford, 1999b, p 198; Solan & Tiersma, 2003, pp 397–399) listeners have heard an unfamiliar target voice for a few seconds or a few minutes, and were

later asked to identify the speaker in lineups. Results were sometimes contrary to expectations in that sometimes longer exposure led to poorer performance by listeners.

We have not been able to find any studies which involved concerted efforts by the listeners to familiarise themselves with a speaker's voice. Some summaries seem to suggest that Clarke & Becker (1969) was such a study, but our reading of the original publication leads us to conclude that this is not the case.

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 whom 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) all conclude that the ability of listeners to identify familiar voices is generally overestimated, including by members of the legal profession – people think that they themselves and that other listeners are better at identifying familiar speakers than they really are.

[99.1010] Typicality of speakers' voices

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 (see, for example, Sørensen, 2012).

If two speakers are selected at random from the population, then it is more likely that they will both have relatively typical voices than that either of them or both of them will 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, however, exceptions to the general pattern of atypical voices being easier to correctly identify. If two different speakers have atypical voices which are both atypical in the same way then this may 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. The two speakers may actually be relatively dissimilar, but their shared atypicality causes the listener to think that they are more similar than they really are.

From the listener's perspective, two speakers may be atypical in the same way if they both speak with the same accent and that accent is unfamiliar to the listener. The listener's built-in relevant

population is the population of speakers whom they are used to hearing. Most Americans are most familiar with American accents of English. Most Americans cannot tell the difference between an Australian and a New Zealand accent. To people familiar with those accents (e.g., Australians and New Zealanders), the differences are obvious. Betancourt & Bahr (2010) reviewed studies involving cross-dialect speaker identification.

One study (Ladefoged, 1978) asked 10 members of the UCLA Phonetics Lab to identify the voices on 12 recordings. 11 of the voices were familiar to all the listeners in that they were other members of the lab, and 1 was a stranger to 7 of the listeners. Listeners were not told whether all the voices would be familiar or not. Only 2 members of the lab were African American, neither was included in the recordings. The stranger was African American. 5 of the 7 listeners for whom the African American speaker was a stranger misidentified the stranger's voice as one of the African American members of the Lab. The results would appear to be due to a prior expectation effect and an atypicality effect associated with a less familiar accent.

There has been at least one court case in which listeners misidentified one speaker for another even though the two speakers had different accents. Listeners from Southern California misidentified a questioned speaker with a Boston accent as a known speaker who had a New York City accent (Labov & Harris, 1994).

We can think of a listener's ability to identify a speaker as being based on their experience with the speakers they have heard over their lifetime. The parallel in a forensic voice comparison system is the database used to train the statistical model which calculates the typicality of an offender recording. If we put the wrong data into this model, for example if we use a database of American English voices when the known- and questioned-speaker recordings are both of Australian English speakers, then the system's estimate of the typicality of the questioned-speaker recording will be very low, very atypical. The resulting likelihood ratio will therefore be much larger than it would have been if a database of Australian English speakers had been used instead. The system will overestimate the strength of the evidence because it is using data from the wrong population. In the same way, a listener who is unfamiliar with a particular accent may tend to misidentify people who speak with that accent.

We have discussed different accents as an intuitively easy thing to understand, but in general, any time two or more speakers are relatively atypical in the same way, then the likelihood that a listener will misidentify them will increase. The problem will be compounded if the speakers are actually relatively similar to each other.

[99.1020] Duration and quality of speech material

Listeners have been found to be 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).

There is, however, a potential problem due to what is known as *confirmation bias* (a form of cognitive bias). A listener can make a speaker identification within a few seconds. Once they have made an identification, they may then focus on similarities in the voices which help confirm their identification, and downplay any differences which could point to their identification being incorrect. This can occur at a subconscious level, the listener does not have to intend to be biassed. Because of confirmation bias, if a listener's initial identification is incorrect, listening for longer may not result in them realising their mistake. There are a large number of research studies on confirmation bias. A review is presented in Kassin et al. (2013).

Edmond & San Roque (2009) state:

> 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. (p. 22)
>
> A witness may become more confident through repeated exposure without any corresponding improvement in accuracy. ... Also, 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. (p. 31)

Loss of acoustic information due to factors such as background noise, poor quality recording, poor quality playback, distortions due to transmission of the voice via a telephone system, and mismatches between the quality of the known- and questioned-speaker recordings, also tend to reduce correct-identification rates (Rathborn et al., 1981; Bull & Clifford, 1984, pp 114–116; Bull & Clifford, 1999b, p 197; Rose, 2002, p 102).

Nolan et al. (2013) found that when telephone-quality recordings of pairs of speakers were presented, they were perceived to be more similar than when good quality recordings of the same pairs of speakers were presented. When one member of the pair was good quality and the other member telephone quality, they were perceived to be more different.

In an identification task, however, it could be that since listeners know that people sound different on the telephone versus in-person, listeners could attribute some genuine between-speaker differences to the effect of the telephone, leading them to ignore differences and fail to distinguish speakers.

Rathborn et al. (1981) obtained a correct-identification rate of 63% when the original exposure to the target speaker was a good quality recording and the lineup recordings were good quality recordings, but when both were recorded via a telephone the correct-identification rate was 45%, and it was 42–45% when one was a good quality recording and the other a telephone quality recording. McDougall et al. (2015) found better performance for good quality recordings versus good quality recordings (75% correct), poorer for telephone recordings versus telephone recordings (64% correct) and for good quality recordings for the original exposure versus telephone quality recordings for the lineup (60% correct), and poorest for telephone quality recordings for the original exposure versus good quality recordings for the lineup (32% correct). The latter result is particularly noteworthy because it is likely to correspond to real-world conditions and the way one would naïvely set up a voice lineup.

Öhman et al. (2010) found no difference in performance for combinations of good quality and mobile telephone quality, but this was because listeners were performing at chance levels even for the good quality versus good quality condition.

Öhman et al. (2013) found that in an experiment in which the original exposure was to an angry voice and the speakers in a lineup conducted immediately afterwards spoke with normal voices, adult listeners never correctly identified the target speaker.

[99.1025] Time interval between exposure and lineup

In general, one would expect earwitnesses' voice identification accuracy to decline as the time interval between exposure to an unfamiliar voice and hearing the lineup increases. Research results are somewhat mixed, perhaps because of other differences between the studies. Sherrin (2015) reviews this topic.

Some studies suggest a relatively rapid decline in performance. For example, Öhman et al. (2013) found that both adult and 11–13 year old children had 19% and 25% correct identification rates immediately after exposure, but correct identification rates were at or below chance after a two week delay (chance was 12.5%). Other studies suggest little decline over longer periods of time. For example, Kerstholt et al. (2006) found correct identification rates of 24–32% with no significant difference between a three week and an eight week interval.

[99.1030] Prior expectation bias

If a listener expects to hear a particular voice, then they are more likely to identify the voice they hear as the voice 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 known person they expected to hear (Rose, 2002, p. 104).

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 will be the voice of a particular person, then whatever the properties of the voice the listener will still be 100% certain that the person they heard was the person they expected to hear – hearing the voice will have had no effect on their identification.

2. If a 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, if 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 substantial influence.

A practice which has been criticised for being inherently suggestive is a speaker lineup of one, called a *show-up* (Solan & Tiersma, 2003, pp 381–382, 390–393, 427–428). This is where an earwitness is played a recording of only one person, or hears that person live, and that person is the suspect. Show-ups cover any situation in which the earwitness is explicitly or implicitly asked whether one voice they heard is that of one particular speaker. The earwitness could be a member of the public, or could be a police officer. Eyewitness and earwitness show-ups are considered suggestive because it is reasonable for the witness to infer that the reason they are being shown only one person is that the police believe this person is guilty, and what the police require is for the witness to make a positive identification in order to help convict that person.

Solan & Tiersma (2003, pp. 381–382) summarise a number of cases in which, 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 then 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 (bias instilled in the earwitness by the police), and have ruled them inadmissible. US courts have, however, typically allowed such suggestively-tainted identifications when made by police officers (Solan & Tiersma, 2003, pp. 388–393).

[99.1040] Example of speaker recognition by a layperson

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 stated 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, but the prosecution did not have a forensic voice comparison conducted by a forensic practitioner. Instead, the prosecution put the police officer on the stand to give testimony related to her identification of the speaker. The defence called an expert witness to summarise research on the validity of speaker recognition by laypeople in general.

Below **[99.1050]**–**[99.1110]**, the factors listed above **[99.880]**–**[99.1030]** as being relevant to the validity of speaker recognition by laypeople are related to the police officer's written statements and oral testimony (only the oral testimony was presented to the jury).

[99.1050] Variability between listeners

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 therefore unknown.

[99.1060] Listener certainty

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).

Given 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 probability that her identification is correct.

[99.1070] Listeners' familiarity with speakers' voices

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).

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. 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 multiple 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.

[99.1080] Typicality of speakers' voices

The police officer stated that "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). 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.

[99.1090] Duration and quality of speech material

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 likely to be correct 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 voice [*sic*] 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 less likely to be correct than if she had spent more time comparing the properties of the known and questioned voices before coming to her definitive conclusion. Such an immediate identification could, however, 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.

[99.1100] Prior expectation bias

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 voice [*sic*] 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 speaker whom 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" (Statement of 17 September 2008, point 6). "Whilst listening to the calls I became familiar with the accuseds voice [*sic*]." (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).

[99.1110] Conclusion

The police officer was either correct or incorrect in her identification of Mr Mansell as the speaker she had heard on the telephone intercept recordings. The discussion above merely provides a guide to factors which should be considered in coming to an opinion as to the likelihood that the police officer's identification was correct, and highlights the multiple problems with this sort of testimony. This was not a true earwitness situation, and, in our opinion, a forensic voice comparison with a demonstrated degree of validity and reliability would have been a much better way to proceed.

[99.1120] Further reading

Reviews of speaker recognition by laypeople in a legal context are presented in Solan & Tiersma (2003), Edmond & San Roque (2009), Laub (2010), Edmond et al. (2011a,b), and Sherrin (2015).

DISPUTED UTTERANCE ANALYSIS

[99.1500] Introduction

Certain words of phrases sound similar to each other, for example:

- "thirteen" – "thirty"
- "fourteen" – "forty"
- "fifteen" – "fifty"
- "S" – "F"
- "C" – "Z" (in American English)
- "G" – "J"
- "N" – "M"
- "excuse me while I kiss this guy" – "excuse me while I kiss the sky"
- "holding back the ears" – "holding back the years"
- "I can do it" – "I can't do it"

Such words and phrases can be mistaken for each other, even under good listening conditions. The potential to mishear such words and phrases increases under poorer conditions such as if there is background noise or if the speech has been transmitted through a telephone system, also if the speaker is mumbling or speaking quickly, etc. Under poor conditions, listeners can even confuse words of phrases that under good conditions sound quite different.

Differences between the accent of the speaker and the listener can also lead to the listener hearing something different than what the speaker intended. For example:

- A Southern US speaker's pronunciation of "pen" may be heard as "pin" by speakers of other accents of English.
- A Brooklynite's pronunciation of "bird" may be heard as "Boyd" by speakers of other accents of English.
- An American's pronunciation of "writing" may be heard by Anglophone Canadians as "riding" (in Canadian English the vowel sounds in "riding" [ˈɹaɪːɾɪŋ] and "writing" [ˈɹʌɪɾɪŋ] are different).
- A Newfoundlander's pronunciation of "boy" may be heard as "buy" by speakers of other accents of English.
- A New Zealander's pronunciation of "late tackle" may be heard as "light tickle" by speakers of other accents of English.
- A speaker from the southeast of England may pronounce "book" in a way which is heard as "buck" by speakers from the north of England.

Listening to speech is not simply a *bottom-up* process of decoding the incoming acoustic signal into phonemes as building blocks for words. Listening to speech is also a *top-down* process in which the listener uses contextual information about what they are likely hear. If a waiter asks

what you would like to drink and you answer "ko-ra" instead of "cola", the waiter will probably not even notice your mispronunciation. One would probably have to emphasise the last word in "have a nice May" for a listener not to hear it as "day". Human communication has evolved to use a combination of top-down and bottom-up listening. Combining top-down information with bottom-up information makes understanding speech more efficient. Under poor listening conditions, in which bottom-up information is compromised, listeners rely more on top-down information. If listeners did not use top-down information, speakers would have to speak more clearly and communication would be more laborious. Top-down listening makes communication more efficient, but it can lead to mishearing when what the speaker says has a low probability with respect to what the listener expects to hear. What is heard can be highly dependent on what the listener expects to hear.

[99.1510] Approaches to disputed utterance analysis

Some practitioners conduct disputed utterance analyses by listening. They use high quality audio equipment to listen to the recording of the disputed utterance, and make a subjective judgement based on what they hear. This approach is inconsistent with the new paradigm for the evaluation of forensic evidence. It is susceptible to cognitive bias. If different practitioners hear different things and their abilities have not been empirically tested under conditions reflecting those of the case, there is no principled way to choose between them.

Relying on the testimony of what laypeople (such as police officers) hear, or playing the disputed utterance to the trier of fact and asking them to decide is also highly problematic. Laypeople are susceptible to cognitive bias, and are likely to be listening using relatively poor quality audio equipment in acoustically non ideal environments (e.g., laptop speakers in an office, or a loudspeaker system in a courtroom).

In addition, what anybody hears, whether they be layperson or expert, is irrelevant. The question of interest in a disputed utterance analysis is not what was heard, but what was said. The best way to evaluate what was said is not via listening, but via quantitative acoustic measurements made on relevant data, and statistical analysis of the resulting measurements. The question for the forensic practitioner to address is:

1. What is the probability of obtaining the acoustic properties of the disputed utterance if what was said is what the prosecution contends was said?

   versus

2. What is the probability of obtaining the acoustic properties of the disputed utterance if what was said is what the defence contends was said?

Also, the performance of the system used to calculate the strength of the evidence should be empirically assessed under conditions reflecting those of the disputed utterance in the case.

Disputed utterance analysis will not be effective as distinguishing words that are true homophones (they sound exactly alike). These could include, for example, "know" versus "no", "sun" versus "son", and "cite" versus "sight" versus "site".

Below, we provide two examples of disputed utterance analyses based on real cases. These are not analyses which were actually conducted and presented to the courts, they were conducted for

research purposes. The analyses calculate likelihood ratios on the basis of relevant data, quantitative acoustic measurements, and statistical models, and the results of empirical testing of performance are presented.

A topic somewhat related to disputed utterance analysis is the preparation of transcripts for use in legal proceedings, that topic is discussed in **Transcripts in the legal system [100]** (Fraser, 2010).

[99.1520] Example 1: VOT and formants

In a Swedish case [nº B1293-07 of Hovrätten för nedre Norrland, 2008-02-26], a word on an audio recording of a police interview with an eye witness (a female Swedish speaker) was disputed as being either the pronoun "dom" [dɔm] THEY or the proper name "Tim" [tʰɪm]. The quality of the recording was reasonable but not particularly good. It had originally been recorded on analogue equipment.

The analysis below is described in greater detail in Morrison et al. (2014). Thanks to Jonas Lindh, the forensic practitioner in the case, for providing the data and performing the acoustic analysis reported below.

[99.1530] Hypotheses

We calculated a likelihood ratio based on the following hypotheses:

1. The speaker said "dom".
2. The speaker said "Tim".

The words "dom" [dɔm] and "Tim" [tʰɪm] differ in both the first consonant and in the vowel (the last consonant is the same).

Both [d] and [tʰ] are alveolar plosives: a closure is made between the tip of the tongue and the alveolar ridge, the air pressure behind this closure is increased, then the closure is released. [d] is a prevoiced plosive: the vocal folds start vibrating before the closure is released. [tʰ] is a voiceless aspirated plosive: the vocal folds do not start vibrating until after the closure has been released and there is turbulent airflow before the vocal folds start vibrating. See section **[99.520]**.

The articulations of the [ɔ] and [ɪ] involve differences on the position of the jaw, tongue, and lips, resulting in the vowels having different formant values. See section **[99.460]**.

[99.1540] Relevant data

The relevant data, in addition to the recording of the disputed word itself, consisted of 29 undisputed tokens of "dom" and 16 undisputed tokens of "Tim" spoken by the same speaker in the same recording.

[99.1550] Acoustic analysis

Lindh measured the voice onset time (VOT, the time between the release of the closure and the beginning of voicing, see section **[99.520]**) of the disputed word, and of the undisputed tokens of "dom" and "Tim". A histogram of the resulting VOT values is shown in Figure 27 top panel.

Lindh also measured the first and second formant values (F1 and F2) in the middle of the vowel in the disputed word and in the middle of the vowel in each undisputed token of "dom" and "Tim". See section **[99.460]**. A scatterplot of the vowels' F1 and F2 values is shown in Figure 28.

FIGURE 27. Histogram of VOT measurements made on the plosives in the undisputed "dom" and "Tim" tokens. Top panel: Original VOT values. Bottom panel: Transformed VOT values. The vertical green line indicates the measured VOT of the plosive in the disputed word.

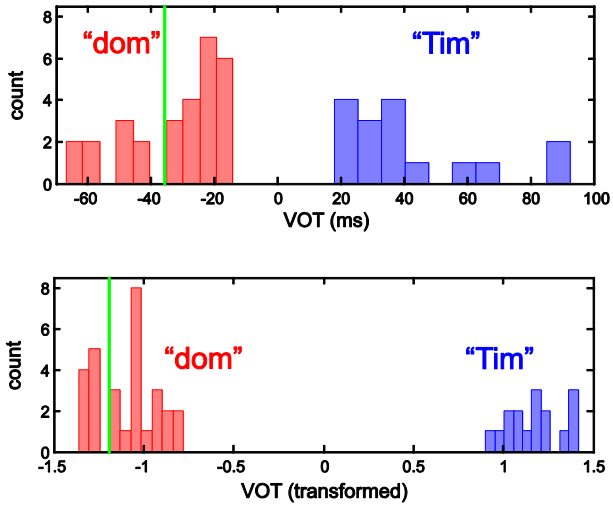
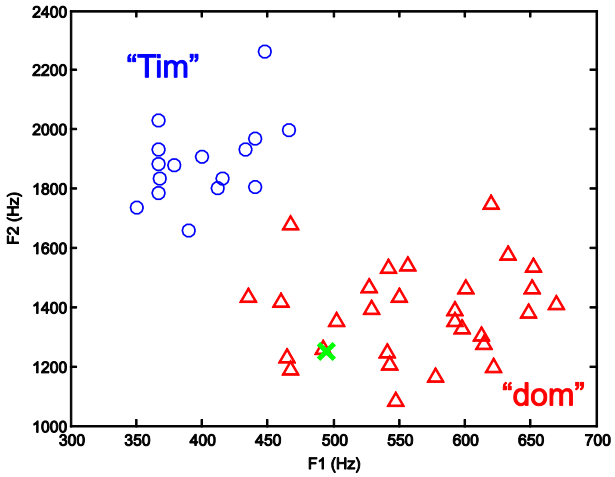


FIGURE 28. Scatterplot of F1 and F2 measurements made on the vowels in the undisputed "dom" and "Tim" tokens. The green cross indicates the measured F1 and F2 of the vowel in the disputed word.



[99.1560] Statistical analysis

The VOT values for the undisputed "dom" and "Tim" tokens clearly did not have Gaussian distributions. A transformation was applied so that the distribution of the transformed data had more Gaussian distributions. Such transformations are standard statistical techniques (in this case the transformation used was an arctangent function). The transformed data are shown in Figure 27 bottom panel.

Morrison et al. (2014) reported the details of four statistical analyses. Here we only report the results of a Bayesian analysis using uninformative priors. Such an analysis is intended to deal with the problem of having relatively little data to train the statistical models. The analysis was based on three-dimensional feature vectors consisting of transformed VOT, F1, and F2 (hereinafter [VOT, F1, F2] feature vectors). A "dom" model and a "Tim" model were trained on the feature vectors from the undisputed "dom" and "Tim" tokens. The likelihood of the [VOT, F1, F2] feature vector from the disputed word was then calculated for the "dom" model and for the "Tim" model, and the former likelihood divided by the latter to calculate a likelihood ratio. See section **[99.230]**. In a Bayesian analysis *Bayes factor* rather than *likelihood ratio* would be the technically correct term, but for simplicity we will continue to use the term *likelihood ratio*.

[99.1570] Results

An empirical test of system validity was conducted using a cross-validation procedure, i.e., by holding out one token from the training data, using the remaining data to train the models, using those models to calculate a likelihood ratio for the held out token, and cycling through holding out each of the tokens in the training data (see section **[99.820]**). C_{llr} was very close to zero and Tippett plots looked like vertical lines far apart from each other. Whether the empirically demonstrated level of validity was good enough was not our concern in this case. Our concern was whether,

given the small amount of training data, the very very good empirically demonstrated level of validity was misleading. That is why in Morrison et al. (2014) we explored several models for dealing with small amounts of training data, and here we only report the results of the Bayesian analysis using uninformative priors.

Based on the results of the Bayesian analysis with uninformative priors, the probability of obtaining the measured acoustic properties of the disputed word is approximately 3 billion times greater if the speaker had said "dom" than if the speaker had said "Tim".

[99.1580] Example 2: Fricative spectra

Before reading this example, we recommend that you listen to the three audio recordings posted at <http://expert-evidence.forensic-voice-comparison.net/#audio>. Write down what you think is said in each recording. This should help you appreciate that what is said in speech recordings is not always clear and why there can be disputes about what was said.

In 1995 in New Zealand, David Bain was convicted of murdering his family. The conviction was subsequently appealed and he was found not guilty at a retrial in 2009 (*Bain v R* [2009] NZSC 16; see also Innes, 2011). The prosecution contended that Bain killed his family, established an alibi, then returned home and made a call to emergency services. The defence contended that his family was killed while he was out, he discovered them dead upon his return, and then he made the call to emergency services.

The call to emergency services was recorded. The quality of the existing versions of the recording is poor. Bain was gasping for breath. The telephone system was an older analogue system, not a newer digital system. The telephone call was originally recorded on an analogue reel to reel system, and was then copied to an analogue microcassette. The microcassette went missing for a number of years, but resurfaced before the retrial. A digital copy of the microcassette was then made. There is a 50 Hz hum on the recording (such hums are due to interference from the mains electricity supply).

Before the retrial, a police officer listened to the recording and thought he heard Bain utter a phrase which amounted to a confession. This was disputed by the defence.

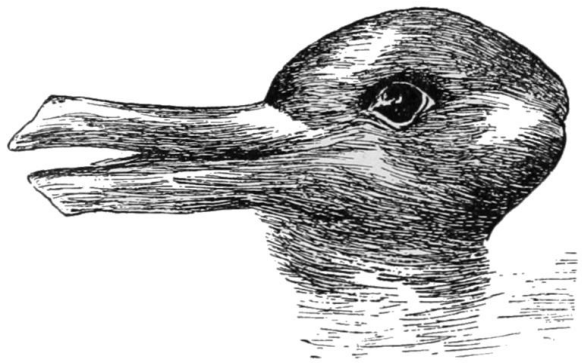
A number of forensic practitioners were asked to analyse the recording. Most gave an opinion based on listening to the recording. There was a lack of agreement between different practitioners, and the court ruled that the jury would not be allowed to hear the disputed-utterance portion of the recording, nor any reference to that part of the recording, nor what the police officer thought he heard. One forensic practitioner, Philip Rose, pointed out that what anybody heard was irrelevant, that what was relevant was what was said, and that the appropriate way to assess this was to calculate a likelihood ratio based on an acoustic and statistical analysis. Rose, however, did not conduct such an analysis at the time.

In a post-trial research activity, Fraser et al. (2011) conducted experiments in which mock jury members were asked to listen to the disputed utterance in the Bain recording and say what they heard. Some listeners were then exposed to testimony that what was said was what the prosecution contended was said, and others were exposed to testimony that what was said was something else. Almost no one heard what the prosecution contended was said on initial listening, but after being

exposed to the testimony, the group who heard the prosecution contention were much more likely to report hearing what the prosecution contended was said.

The study demonstrated the power of suggestion with respect to what listeners hear. This is an auditory version of ambiguous visual images that can be seen as one thing or another, e.g., a duck versus a rabbit, see Figure 29. One may perceive the image in one way and be unable to perceive it the other way even when the alternative is pointed out, then one may suddenly shift to perceiving it the other way and then have difficulty switching back. If before one sees the image, one is primed by a suggestion about what one is about to see, one will tend to perceive the image in the way suggested, e.g., if you hear a quack and then see the image you are more likely to perceive it as a duck. Confirmation bias can then set in, which is why it can be difficult to switch to seeing the image as a rabbit. Likewise, when one of primed to hear an ambiguous acoustic signal in a particular way, one is more likely to perceive it that way and have difficulty perceiving it in another way.

FIGURE 29. Ambiguous visual image (from *Fliegende Blätter*, 23 October 1892).



We have deliberately delayed telling you what the prosecution contended that Bain said in the disputed utterance. They contended that he said "I shot the prick." The defence never formally proposed an explicit alternative other than not "I shot the prick", but an alternative that was suggested by some forensic practitioners is "I can't breathe." Go back and listen to the three audio recording again and see if you can hear them as either "I shot the prick" or as "I can't breathe." You may be able to experience something akin to the experience of looking at an ambiguous visual images. Perhaps you did not hear either of "I shot the prick" or "I can't breathe" before they were suggested to you, but now you may be able to make yourself sometimes hear "I shot the prick" and sometimes hear "I can't breathe." In recording number 1 the speaker was actually saying "I can't breathe", and in recording number 2 the same speaker was actually saying "I shot the prick." We do not know what was said in recording number 3; it is the actual disputed utterance from the Bain case.

Below we describe a preliminary analysis of the disputed utterance from the Bain case. We calculate a likelihood ratio based on relevant data, quantitative acoustic measurements, and

statistical models. The validity of the system is also empirically tested. The analysis is described in greater detail in Morrison & Hoy (2012).

[99.1590] Hypotheses

Rather than try to analyse the whole of the disputed utterance, we focussed on the first consonant in the word which is contended to be either "shot" or "can't". Thus we calculated a likelihood ratio based on the following hypotheses:

1. The consonant spoken by Bain was /ʃ/.
2. The consonant spoken by Bain was /k/.

The consonant in the disputed utterance was clearly a fricative. /ʃ/ is realised as a voiceless postalveolar fricative [ʃ]. /k/ is normally realised in this context as a voiceless aspirated velar plosive [k^h], but if someone is gasping for breath, it can be realised as a voiceless palatal fricative [ç] – complete closure is not achieved and the constriction between the tongue and the roof of the mouth is further forward. See Figure 15 in section **[99.450]**, and the description of fricatives in section **[99.500]**. The postalveolar place of articulation is just behind the alveolar ridge, and palatal place is between the postalveolar and the velar places, so the articulations of [ʃ] and [ç] are actually quite similar (Magnetic Resonance Imaging videos and animations of these speech sounds are available at <www.seeingspeech.ac.uk/ipachart/>, still x-ray tracings can be found in Laver, 1994, p 246).

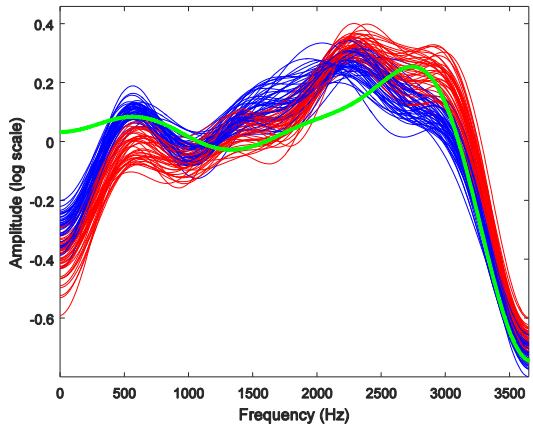
[99.1600] Relevant data

The recording of the call made to emergency services did not include undisputed tokens of both [ʃ] and [ç]. Recordings were made of another New Zealand English speaker who happened to be approximately the same height and age as Bain had been. The speaker mimicked Bain's out of breath speaking style and was recorded via a landline telephone. He said each phrase "I shot the prick" and "I can't breathe" multiple times. The recording of some utterances were excluded due to technical problems, leaving 41 usable tokens of [ʃ] and 44 usable tokens of [ç].

[99.1610] Acoustic analysis

We measured the spectrum of each of the undisputed tokens of [ʃ] and [ç], and of the consonant in the disputed utterance. The spectra were characterised using cepstral coefficients, similar but not exactly the same as MFCCs (see section **[99.720]**). A single spectral measurement made over the whole duration of the fricative, and a hertz rather than a mel frequency scale was used over the frequency range 0 to 3675 Hz. The coefficient values were amplitude normalised, and feature vectors consisting of the 1st through 8th cepstral coefficients were used for statistical analysis. Figure 30 shows the resulting smoothed spectra (the drops in amplitude at low and high frequencies are due to the telephone bandpass, the lack of low frequency drop for the disputed consonant is due to the 50 Hz hum).

FIGURE 30. Smoothed spectra of undisputed tokens of [ʃ] (red lines), undisputed tokens of [ç] (blue lines), and the consonant in the disputed utterance (thicker green line).



[99.1620] Statistical analysis

Two multivariate Gaussian distribution models were trained, one trained on the feature vectors from the undisputed tokens of [ʃ] and the other trained on the feature vectors from the undisputed tokens [ç]. The likelihood of each model were then evaluated at the value of the feature vector form the disputed consonant, and the likelihood from the first model divided by the likelihood from the second model to calculate a likelihood ratio (see section **[99.230]**).

The validity of the system was empirically tested using a cross-validation procedure. One undisputed token was held out, all the other undisputed tokens were used to train the models, then the likelihood ratio for the held out token was calculated. This was repeated for every undisputed [ʃ] token and every undisputed [ç] token. See section **[99.820]**.

[99.1630] Results

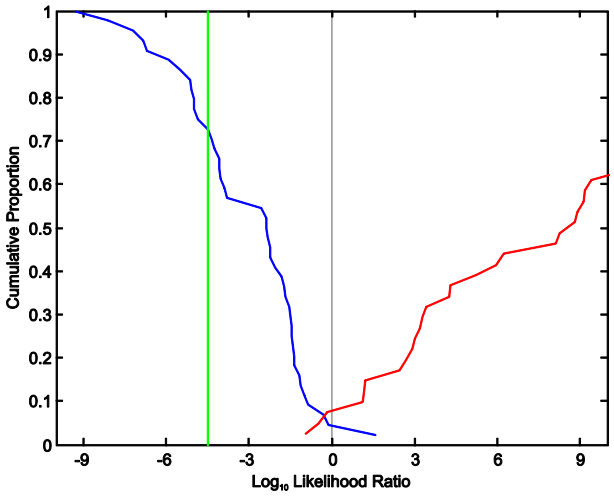
The results of the validation test are represented as the Tippett plot in Figure 31. The C_{llr} value was 0.172. Both the Tippett plot and the C_{llr} value indicate a good validity.

The likelihood ratio value calculated for the disputed consonant was approximately 1/31,000. The probability of obtaining the measured acoustic properties of the disputed consonant was 31,000 times greater if Bain had said [ç], as in “can’t” than if Bain had said [ʃ] as in “shot”.

This analysis was called a preliminary analysis, and there are caveats which should be noted. The training data came from a different speaker, not from Bain, and the recordings conditions were not exactly the same as those in the Bain case. A potential way to improve on the analysis described here would be to record multiple speakers. The statistical models would then be trained on data from multiple speakers, and cross-validated testing would be conducted by holding out all the recordings of a speaker from the training set. The recordings could be directly recorded or

processed so that they more closely reflect the recording conditions of the disputed utterance. Statistical mismatch compensation techniques could also be applied.

FIGURE 31. Tippett plot showing the results of the empirical validation of the system applied to the disputed utterance in the Bain case. The vertical green line indicated the likelihood ratio value calculated for the disputed consonant.



OTHER BRANCHES OF FORENSIC SPEECH SCIENCE

[99.1700] Introduction

Below we briefly describe other branches of forensic speech science:

- Lie detection **[99.1710]**
- Intoxication detection **[99.1720]**
- Voice disguise in forensic casework **[99.1730]**
- Speaker profiling **[99.1740]**
- Language analysis for determination of origin (LADO) **[99.1750]**

[99.1710] Lie detection

Listeners may be able to detect whether speakers are lying or not at levels slightly above chance. The situation may be somewhat akin to speaker recognition by laypeople. Some listeners may be better at detecting lies than others. Some speakers may be better at concealing lies than others. Familiarity of the listener with the speaker may be relevant.

Claims of ability to detect when a speaker is lying would have to be empirically validated under conditions reflecting those of the case under investigation. In forensically relevant situations the speaker may be under emotional stress, e.g., even an innocent person being questioned by police may feel stressed, and this is likely to affect a speaker's speech. To be relevant to such situations, empirical validation would have to test speakers under stressed conditions. See section **[99.290]ff**.

A number of products purport to automatically detect when a speaker is lying or purport to assist a human operator to detect when a speaker is lying. The products may overtly claim to detect stress rather than lying, they may be marketed as voice stress analysers, but it may be implied that they are lie detectors.

Several papers have reported on theoretical and empirical assessments of commercial systems whose explicit or implied function is lie detection via acoustic analysis of voice signals (see the following papers and reference cited therein: Damphousse et al., 2007; Eriksson & Lacerda, 2007; Hollien et al., 2008; Harnsberger et al., 2009; Horvath et al., 2013). Claims of scientific bases for how the systems worked were found to be spurious. Such systems may decrease the incidence of lying via a placebo effect whereby speakers who believe an effective lie-detection system is in use are less likely to lie. Beyond that, none of the published studies found any substantial evidence in support of the hypothesis that automatic systems performed at levels above chance or that human-operated systems performed at levels above those for unaided listening.

One paper (Eriksson & Lacerda, 2007) may have strayed from a sober scientific evaluation and implied that the manufacturer of one product was perpetrating a fraud. The manufacturer threatened to sue the publisher of the journal in which the paper was published, and the publisher removed the paper from its website. The paper was subsequently posted at multiple locations on the internet by multiple third parties.

Unless empirical demonstration of an adequate level of validity and reliability is obtained, we do not recommend the use of human-listener, acoustic, or automatic approaches which purport to detect lies on the basis of a speaker's speech.

[99.1720] Intoxication detection

Slurred or disfluent speech is often taken as an indication that the speaker is intoxicated. There could, however, be other reasons for slurred speech other than alcohol or illegal drugs, including the effects of fatigue, dental anaesthesia, prescription drugs, head trauma, stroke, and motor neuron disease.

Disfluencies include: Additions of phonemes or words, hesitations or pauses, repetitions, substitutions of phonemes or words, voicing/devoicing substitutions, omissions of phonemes or words, distortions, and lengthening (Hollien et al., 2001).

Hollien et al. (2014) reviewed literature on the effects of alcohol intoxication on speech. They pointed out that factors that would need to be controlled in research include the speaker's drinking habits and their level of intoxication. Also that it would be necessary to compare recordings of the same individual's speech when they are intoxicated versus when they are sober. In the literature review, Hollien et al. found that:

- Listeners tended to overestimate the level of intoxication of speakers who were mildly intoxicated, but underestimate the level of intoxication of speakers who were severely intoxicated.
- Sober speakers could deliberately speak in a way that listeners perceived as severely intoxicated, and even severely intoxicated speakers could deliberately speak in a way that made listeners perceive them as less intoxicated.
- Whether speakers were habitually light, moderate, or heavy drinkers had little effect on listeners' perception of their level of intoxication.
- As intoxication level increased, there were moderate increases in fundamental frequency (f0), moderate decreases in speaking rate, and large increases in disfluencies.

Baumeister et al. (2012), however, found that although the general trend was for f0 to increase with level of intoxication, some individuals' f0 actually decreased as intoxication level increased. No single f0-based model would therefore be effective for predicting an individual speaker's levels of intoxication. Schiel & Heinrich (2015) also found substantial between-speaker variability in the relationship between various types of disfluencies and level of intoxication.

[99.1730] Voice disguise in forensic casework

Clark & Foulkes (2007) reviewed the literature on voice disguise in forensic voice comparison casework.

Most cases did not involve voice disguise. Estimates of the percentage of cases involving voice disguise in different laboratories ranged from 2.5% to 25%. In some types of cases such as blackmail and ransom demands, however, the speaker has a high expectation of being recorded and probably does not want to be identified, hence voice disguise is the norm in such cases

(laboratories reporting higher rates of voice disguise probably had more cases of this type). Although voice disguise is more common in questioned-speaker recordings, in some cases a speaker may attempt to disguise their voice when they know that they are being recorded for a known-speaker recording, e.g., when they know that they are the speaker in the questioned-speaker recording and they spoke normally in the questioned-speaker recording.

Common forms of voice disguise in casework included: using creaky voice (very low fundamental frequency), falsetto (very high fundamental frequency), whisper, pinching one's nose, placing an object (e.g., a pencil) between one's teeth, and affecting an accent. Speakers often had difficulty maintaining a disguise. Speakers who affected an accent, for example, would revert back towards their usual accent. Such forms of disguise are usually readily identified as disguise, especially if they are not consistently maintained. Attempts at mimicking other particular speakers or particular foreign accents seldom convince listeners that the speaker really is the person they are impersonating or really is a first-language speaker of the foreign accent they are feigning, but may be reasonably effective means of disguise. Electronic voice disguise, i.e., using signal processing on a standard computer or dedicated hardware, was rare, but its prevalence was expected to increase as the technology became more readily accessible.

Voice disguise has been found to negatively affect the performance of automatic forensic voice comparison systems (e.g., Meuwly, 2001, §8.7; Künzel et al., 2004; Zhang & Tang, 2008; Perrot & Chollet, 2008). Research has also been conducted on automatic detection and classification of voice disguise (e.g., Perrot et al., 2009; Perrot & Chollet, 2012; Wu et al., 2014). For the purposes of conducting casework, voice disguise should be treated like speaking style or recording condition: The system should be trained and must be tested using recordings which include the disguise and any mismatches, e.g., a particular type of disguised speech in the questioned-speaker recording versus normal speech in the known-speaker recording.

Mohammadi & Kain (2017) present an overview of signal-processing systems for converting a speaker's voice so that it sounds like the voice of another particular speaker. At present, the ability of real-world systems to do this convincingly is substantially lower than that of their fictional counterparts.

For a relatively broad review of voice disguise in forensic contexts, including topics not covered here, see Perrot & Chollet (2012).

[99.1740] Speaker profiling

In contrast to forensic voice comparison in which a known-speaker recording and a questioned-speaker recordings are compared and the hypotheses to be addressed are same-speaker versus different-speaker, in speaker profiling there is only a questioned-speaker recording and the question concerns what one can we tell about the speaker based on this recording. Speaker profiling is usually performed for investigative purposes, e.g., to help investigators narrow the potential pool of suspects. A number of things can potentially be inferred from a recording of a speaker's voice, including the speaker's height, age, and sex, the social, regional, and/or foreign accent spoken, and the language spoken. The first three are physiological and the remainder learned.

On average, adult males have more massive vocal folds and longer vocal tracts than adult females leading to males having lower fundamental frequency (f0) and lower formant frequencies than

females. The perceptual differences are usually so obvious that the services of a forensic speech scientist are not needed. Sometimes, however, the sex of the speaker is not obvious and an analysis of properties such as f0 and the formant frequencies of different vowels may be informative. This, in theory, would be particularly amenable to acoustic and statistical analysis leading to a likelihood ratio. A speaker's sociological gender should be distinguished from a speaker's physiological sex. A forensic phonetician may be able to identify speech patterns often associated with a gay accent, but this would not necessarily be indicative of the speaker's actual sexuality.

There are large changes in speech properties through childhood and adolescence due to physical changes. These may allow a relatively accurate estimate of speaker age. A questioned-speaker recording of a child or adolescent is more likely to be a recording of a victim rather than an offender. On average, f0 for adult males gradually decreases to around age 40 then gradually increased from around age 50. For adult females the pattern is a gradual decrease in f0 which may accelerate at menopause but which may increase again in older age. Age estimate for adults based on speech is, however, of poor accuracy irrespective of whether an auditory or an acoustic and statistical analysis approach is used. The underlying problem is that the correlation between age and speech properties is slight.

Listeners tend to associate lower f0 and lower formants with taller speakers, but these acoustic features have not been found to be accurate predictors of speaker height. The accuracy of height estimation based on voice recordings is therefore poor irrespective of whether an auditory or an acoustic and statistical analysis approach is used.

Identification of social and regional accents usually depends on the knowledge and skill of a phonetician or dialectologist who has (via study or experience) some familiarity with the accent that happens to be being spoken. One individual may be an expert on English accents in the British Isles, another may be an expert on English accents in North America, and yet another an expert in Chinese languages and dialects. An auditory approach is almost always used. An acoustic and statistical analysis approach would be difficult to implement because of the difficulty of obtaining appropriate training data, especially since *a priori* one does not known which accents might be relevant. Also, in addition to the properties of the speaker's voice, what the speaker says, e.g., vocabulary and grammatical patterns, can provide useful information. A practitioner of speaker profiling will therefore analyse linguistic information as well as speech properties. Having identified an accent typical of a general region, the practitioner may be able to consult existing descriptions or data, or experts on accents of that region, or laypeople from that region and further refine the speaker's accent to a more local area.

Foreign accents can be challenging. It may be quickly determined that the speaker has a foreign accent, but identifying their first language (L1) may be harder. Even untrained English speakers can fairly accurately identify Spanish accented, French accented, and German accented English, because they are familiar, but have little to no idea about the vast majority of possible foreign accents. The task is complicated by the fact that foreign accents can be stronger or weaker, and can also be mixed with regional accents. A practitioner of speaker profiling will have to have some familiarity with the particular foreign accent, at least enough to narrow down the possibilities. A practitioner may, for example, be sufficiently familiar with Slavic accented speech to hypothesise that a speaker has a Slavic accent, but not be sufficiently familiar with Slavic accents to identify which Slavic language might be the speaker's first language. They would, however, be in a position to find other resource, experts, or laypersons who may be able to help with either identifying the particular Slavic language or disconfirming the Slavic-language hypothesis.

Which language is being spoken is something which is highly amenable to automation, at least at a screening level. There are thousands of different languages, so no one individual can be familiar with all of them. An automatic language recognition system can be trained on a large number of languages, and the probabilistic output of the system used to decide which human expert to refer the case to. The limitation is that the automatic system will not be trained on all languages (data from all languages will not be available for training), and when presented with a language it was not trained on it may output a fairly high probability for one of the languages it was trained on. Once a hypothesis as to the language spoken has been developed, the practitioner may be able to consult speakers of that language. L1 speakers of a given language (and even second-language, L2, speakers) are excellent at identifying that language, and are usually also familiar with and therefore good at identifying neighbouring languages.

Compared to forensic voice comparison and disputed utterance analysis, speaker profiling is much more difficult (if not impossible) to conduct using relevant data, quantitative measurements, and statistical models, and much more difficult to validate. Forensic phoneticians and linguists would, however, be expected to be able to and to actually provide detailed justifications for their conclusions. Speaker profiling may be a helpful investigative tool, but the results should be used with caution.

For more detailed reviews of speaker profiling, see: Jessen (2007, 2010), Schilling & Marsters (2015).

[99.1750] Language analysis for determination of origin (LADO)

Language analysis for determination of origin (LADO) can be considered a form of speaker profiling, but concentrating on the question of what a speaker's speech (accent) and their language use (dialect) can tell us about their place of origin. Rather than being only investigative, the purpose is usually to inform decision-makers in asylum or refugee hearings. Also, rather than having only a recording to work from, the practitioner may be able to directly interview the asylum seeker.

Guidelines published by a group of linguists (Language and National Origin Group, 2004) state:

> Language analysis can not be used reliably to determine national origin, nationality or citizenship. ... In some cases, language analysis CAN be used to draw reasonable conclusions about the country of socialization of the speaker. ... The way that people speak has a strong connection with how and where they were socialized: that is, the languages and dialects spoken in the communities in which people grow up and live have a great influence on how they speak.

Identification of languages and social and regional dialects has already been discussed in 0. In LADO, forensic practitioners usually work with informants who are ideally L1 speakers of the language the asylum seeker claims as their L1. In LADO, the situation is complicated by multiple factors: The speaker may genuinely have the background they claim, or may be deliberately attempting to deceive by speaking in a language and accent/dialect that does not reflect their true background. The speaker may be a speaker of a non-standard accent/dialect which may be unfamiliar to both the forensic practitioner and their informants. Speakers may accommodate the way they speak to the way the interviewer speaks, e.g., rather than speaking in their local accent/dialect the speaker may approximate a more prestigious standard form of the language used

by the interviewer. In some parts of the world it is common for speakers to have a local language as their L1, and have some degree of ability to use a more widely spoken regional language and/or a national language. These being mutually unintelligible languages rather than accents/dialects. The speaker may be interviewed in a more widely spoken language rather than in their first language, no L1-speaker interviewer/informant of the claimed L1 being available. Asylum seekers may have over a long period of time mixed with people from other language communities and consciously or unconsciously altered the way they speak. Some readers may be familiar with this phenomenon, e.g., somebody from Britain moves to Canada and lives there for many years, most people in Canada think they still have a British accent, but when they go back to Britain everyone thinks they have an American accent.

There are a number of controversies associated with LADO. Forensic practitioners and L1 informants may be exposed to biasing information. Forensic practitioners may express conclusions with absolute confidence. There is debate about who is qualified to perform LADO: L1 speakers without extensive linguistics training, linguists who have studied the language (but who may not be L1 speakers or have forensic expertise), forensic linguists, forensic linguists and L1 speakers working together. Cambier- Langeveld (2010) argues that:

> you can never just assume that a person is capable of making accurate judgments on regional origin on the basis of speech, no matter what his/her qualifications or background, and no matter how confident he/she is. You can only be sure that a person is capable of making accurate judgments within a particular language area after you have tested a person on this particular task.

For a more detailed review of LADO, see Schilling & Marsters (2015).

Abbreviations

ABRE	American Board of Recorded Evidence
AFSP	Association of Forensic Science Providers
bit	binary digit
cf	compare with (Latin “confer”)
C_{llr}	log-likelihood-ratio cost
DNA	deoxyribonucleic acid
E	evidence
EWCA	England and Wales Court of Appeal
f0	fundamental frequency
F1	first formant
F2	second formant
F3	third formant
F4	fourth formant
FBI	Federal Bureau of Investigation
FRE	Federal Rule of Evidence
ff	following
GIGO	garbage in garbage out
GMM	Gaussian mixture model
GMM-UBM	Gaussian mixture model - universal background model
H_d	different-origin hypothesis or different-speaker hypothesis
H_s	same-origin hypothesis or same-speaker hypothesis
Hz	hertz
IAFPA	International Association for Forensic Phonetics and Acoustics
IAI	International Association for Identification
IAVI	International Association of Voice Identification
IPA	International Phonetic Association or International Phonetic Alphabet
i-vector	identity vector
k	kilo-
L1	first language
L2	second language

log	logarithm
$\log_{10}$	log base ten
$\log_2$	log base two
LADO	Language analysis for determination of origin
LDA	linear discriminant analysis
LR	likelihood ratio
LR_d	likelihood ratios derived from different-speaker comparisons
LR_s	likelihood ratios derived from same-speaker comparisons
MFCC	mel-frequency cepstral coefficients
ms	millisecond(s)
N_d	number of different-speaker comparisons
NICA	Northern Ireland Court of Appeal
NRC	National Research Council
N_s	number of same-speaker comparisons
NSW	New South Wales
o	odds
p	probability
p.	page
PLDA	probabilistic linear discriminant analysis
pp.	pages
R	Regina
s	second(s)
UK	United Kingdom
US	United States
v	versus
VOT	voice onset time
WA	Western Australia

Glossary

This glossary is intended to indicate the specific meaning of particular words as used in the present document. The entries are not intended to provide exhaustive definitions, but usually reference the section in the present document where the relevant concepts are discussed.

accent – A regional or social *accent* refer to the way speakers from a particular region or a particular social group speak. Sometimes a contrast is made between pronunciation (accent) and words and grammar used (*dialect*).

accuracy – The extent to which a measurement or estimate approximates the true value (synonymous with *validity*) [99.290].

acoustic phonetics – The study of the acoustic properties of speech, i.e., the analysis of sound waves which are created by human vocal tracts, which propagate through the air, and which are detectable by human ears or by microphones [99.440].

acoustic-phonetic-statistical approach – An *approach* to forensic voice comparison based on making *acoustic-phonetic* measurements and analysing the measured values using statistical models [99.700].

alveolar ridge – Part of the *vocal tract* near the front of the roof of the mouth [99.450].

anti-resonance – frequency at which a sound is reduced in amplitude, particularly in *nasal* sounds [99.480]

approach – A method for extracting information from voice recordings for the purpose of conducting forensic voice comparison (cf, *framework*, *paradigm*) [99.650].

arytenoid cartilages – Part of the *larynx* [99.540].

aspiration – Turbulent airflow between vocal folds at the beginning of a plosive [99.520].

auditory-acoustic-phonetic approach – An *approach* to forensic voice comparison based on listening and *acoustic-phonetic* measurements [99.660].

auditory approach – An *approach* to forensic voice comparison based on listening [99.660].

automatic approach – An *approach* to forensic voice comparison based on acoustic measurements made using *signal-processing* techniques and analysing the measured values using statistical models [99.720].

bandpass – The range of frequencies which are transmitted by a transmission system, e.g., a telephone system [99.610].

Bayes' Theorem – A statement of the normative logical relationship between beliefs about competing hypotheses (e.g., same-speaker versus different-speaker hypotheses) before and after the presentation of the strength of evidence [99.160].

bilabial – A speech sound made with a closure or constriction of the lips [99.520].

bit rate – A measure of the amount of information encoded or transmitted per second of speech 0[99.600].

blade – Part of the front of the tongue [99.480].

breathy voicing – Voicing combined with turbulent airflow [99.540].

broad transcription – A written representation of speech sounds providing relatively little phonetic 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 – The portion of a vocal fold vibration during which the vocal folds are closed [99.540].

coarticulation – Overlap of the pronunciation of one speech sound with the pronunciation of the proceeding and/or following speech sound [99.480].

codec – A signal compression and decompression algorithm [99.610].

coefficient – The value calculated when fitting a model to data [99.720].

cognitive bias – The subconscious influence of task-irrelevant information on human decision making [99.160].

confirmation bias – A form of *cognitive bias* in which more weight is given to evidence that is consistent with an existing belief and less weight to evidence that is inconsistent with the existing belief [99.1020].

constriction – A narrowing made in the vocal tract in order to produce a speech sound [99.460].

creaky voicing – Irregular low-frequency vocal fold vibration [99.540].

cross validation – A procedure which allows a single set of data to be used for training and testing. It repeatedly divides the data onto different parts to avoid training and testing on the same data [99.820].

defence attorney's fallacy – A logical error in the interpretation of the meaning of a likelihood ratio [99.390].

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

dialect – A regional or social *dialect* refer to the way speakers from a particular region or a particular social group speak. Sometimes a contrast is made between pronunciation (*accent*) and words and grammar used (*dialect*).

diphthong – A vowel with substantial *formant* movement over its duration [99.460].

dorsum – The back of the tongue [99.450].

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 – The proportion of test results which are incorrect [99.300].

feature warping – A statistical procedure to compensate for mismatched in recording conditions. It is commonly used in the *automatic approach* to forensic voice comparison [99.870].

foil speakers – Speakers who sound broadly similar to the voice on the questioned speaker recording (or to the known speaker) [99.660], or to the suspect in a voice lineup [99.960].

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 – A system of reasoning in order to assess strength of evidence based on information (cf, *approach*, *paradigm*) [99.140].

fricative – A speech sound made with a constriction in the vocal tract that causes turbulent airflow 0[99.500].

fundamental frequency (f0) – The rate at which the vocal folds vibrate during voicing. The physical correlate of what is perceived as pitch [99.540].

fusion – A procedure for combining the parallel *score* or likelihood ratio output of several forensic comparison systems and converting them to likelihood ratios [99.240].

Gaussian distribution – (aka *normal distribution*) A simple *probability density function*, a widely used statistical model [99.220].

Gaussian mixture model – A complex *probability density function* [99.230].

Gaussian mixture model - universal background model (GMM-UBM) – A statistical model used in the *automatic approach* to forensic voice comparison [99.720].

harmonic – A multiple of the *fundamental frequency* [99.610].

hertz (Hz) – A measure of frequency, i.e., number of repetitions per second.

histogram – A graphical representation of the proportion of data falling within different ranges [99.220].

idiolect – The pronunciation peculiarities of an individual, finer grained than the peculiarities of a dialect [99.470].

intensity – The physical correlate of what is perceived as loudness [99.600].

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

i-vector – A statistical-model representation of the properties of the speech on a recording commonly used in the *automatic approach* to forensic voice comparison [99.720].

jitter – The degree to which the duration of individual voicing cycles vary [99.540].

laryngeal – Adjectival form of *larynx* [99.450].

larynx – The structure at the bottom of the vocal tract. It includes the vocal folds [99.450].

likelihood-ratio framework – A framework for the evaluation of forensic evidence. It is considered by most forensic statisticians to be the logically correct framework for evaluation of forensic evidence [99.140].

likelihood ratio (*LR*) – A numeric expression of the strength of evidence [99.150].

linear discriminant analysis (LDA) – A statistical model commonly used in *acoustic-phonetic-statistical* and *automatic approaches* to forensic voice comparison [99.810].

log likelihood ratio – A likelihood ratio expressed on a logarithmic scale [99.300].

log-likelihood-ratio cost (C_{llr}) – A measure of the accuracy of a forensic analysis system which quantifies strength of evidence as likelihood ratios [99.300].

logistic regression – A statistical model commonly used in *acoustic-phonetic-statistical* and *automatic approaches* to forensic voice comparison. A method for performing calibration and fusion [99.240].

match probability – The denominator of the likelihood ratio – only applicable when the data are discrete and the numerator of the likelihood ratio is 1 [99.190].

mel-frequency cepstral coefficients (MFCC) – Measurements of speech signals commonly made in the *automatic approach* to forensic voice comparison [99.720].

mel – A measure of frequency based on how humans perceive the frequency of sounds, which differs from the *hertz* scale [99.810].

monophthong – A vowel with negligible formant movement over its duration [99.460].

Monte Carlo simulation – A statistical procedure which can be used to explore the reliability of a statistical model [99.820].

narrow transcription – A relatively detailed written representation of speech sounds [99.460].

nasal cavities – The internal structures of the nose. Part of the vocal tract [99.450].

nasal – A speech sound produced with a closure in the oral cavity and an open velopharyngeal port. Voicing produced at the vocal folds excites the resonance frequencies of the supralaryngeal vocal tract including the oral and nasal cavities [99.480].

nasalised vowel – A vowel made with the velopharyngeal port open. Resonances and antiresonances are contributed by both the oral and nasal cavities [99.480].

nasopharyngeal tube – The pharyngeal cavity (throat) plus the nasal cavities [99.480].

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

normal distribution – (aka *Gaussian distribution*) A simple *probability density function*, a widely used statistical model [99.220].

open quotient – The portion of a vocal fold vibration during which the vocal folds are open [99.540].

oral cavity – The mouth. Part of the vocal tract [99.450].

oropharyngeal tube – The pharyngeal cavity (throat) plus the oral cavity (mouth) [99.450].

packet – The portion of a signal transmitted as a unit by a mobile telephone system or voice over internet system [99.610].

paradigm – "the entire constellation of beliefs, values, techniques, and so on shared by the members of a given [scientific] community" (Kuhn, 1970, p. 175). It includes the combination of *approach* and *framework* used [99.70].

paradigm shift – A change in the culture of science [99.70].

pharyngeal cavity – The 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 – A symbol used to transcribe a speech sound (e.g., a symbol of the International Phonetic Alphabet, IPA) [99.460].

phonetics – The study of the physical aspects of the production, transmission, and perception of human speech [99.440].

plosive – A speech sound made by creating complete oral cavity and velopharyngeal-port closure, 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 – The extent to which multiple measurements or estimates of the same value differ from each other (synonymous with *reliability*) [99.290].

probabilistic linear discriminant analysis (PLDA) – A statistical model commonly used in the *automatic approach* to forensic voice comparison [99.720].

probability density function – A statistical model of the distribution of one or more continuous variables [99.220].

prosecutor's fallacy – A logical error in the interpretation of the meaning of a likelihood ratio [99.380].

quantisation resolution – The number of values used to digitally encode the intensity of a signal [99.600].

relevant population – The population of speakers, any one of whom who could potentially be the questioned speaker if the questioned speaker were not the known speaker [99.180].

reliability – The extent to which multiple measurements or estimates of the same value differ from each other (synonymous with *precision*) [99.290].

resonance frequency – The frequency at which a sound is amplified by a resonator such as the *supralaryngeal vocal tract* [99.460].

rounded lips – Lips held in a configuration such as when saying the vowel sound of "who" [99.460].

sampling frequency – The number of times per second a measurement is taken when digitising a signal [99.600].

schwa [ə] – A non-stressed (neutral) vowel sound [99.460].

score – A value which quantifies the degree of similarity of samples of known and questioned origin and while also taking into consideration their typicality with respect to the relevant population. The value of a score, however, is not directly interpretable as a meaningful likelihood ratio [99.240].

shimmer – The degree to which the amplitude of voicing varies across cycles [99.540].

signal processing – A branch of engineering dealing with the analysis and manipulation audio and video signals, e.g., in telecommunications systems [99.720].

similarity – The extent to which the properties of the voice on the questioned-speaker recording resemble those of the voice on the known-speaker recording [99.150].

spectra – The plural of *spectrum*.

spectrogram – A graphical representation of a signal such as an acoustic speech signal. Time is represented on the *x* axis, frequency on the *y* axis, and intensity as the darkness of a monochrome scale (or the colour in a colour spectrum scale) [99.460].

spectrographic approach – An *approach* to forensic voice comparison based on looking at *spectrograms* [99.680].

spectrum – The frequency properties of a speech signal at a point in time (cf, *spectrogram*) [99.460].

speech – Sounds produced by a human vocal tract for the purpose of communication (used synonymously with *voice*). *Speech* contrasts with "language" which relates to words and grammar.

speech processing – A branch of signal processing dealing specifically with the analysis and manipulation speech signals [99.720].

spread lips – Lips held in a smiling-like configuration [99.460].

strength of evidence – The conclusion reached by the forensic practitioner with respect to the probability of obtaining the evidence if the prosecution hypothesis (e.g., same-speaker hypothesis) were true versus if the defence hypothesis (e.g., different-speaker hypothesis) were true [99.150].

stressed vowel – A relatively long vowel with well-defined formant values [99.460].

supralaryngeal vocal tract – The portion of the *vocal tract* above the *larynx* [99.450].

_t_ distribution – A *probability density function*, a statistical model sometimes used in *acoustic-phonetic-statistical* and *automatic approaches* to forensic voice comparison [99.810].

tip – Part of the front of the tongue [99.440].

train a statistical model – Perform the calculations necessary to build a statistical model. For example, in the case of a univariate *Gaussian distribution*, the mean and the standard deviation must be calculated [99.220].

training data – The data used to *train a statistical model*.

Tippett plot – A graphical representation of the results of a test of a forensic analysis system which quantifies strength of evidence using likelihood ratios [99.330].

tone – A fundamental-frequency pattern which distinguishes one phoneme from other phonemes [99.540].

transposition of conditionals – A logical error in the interpretation of the meaning of a likelihood ratio (aka *prosecutor's fallacy*) [99.380].

trier of fact – The decision maker in legal proceedings, who could be a judge, a panel of judges, or a lay jury, depending on the legal system.

trier of fact's fallacy – A logical error in the interpretation of the meaning of a likelihood ratio [99.394].

turbulent airflow – Irregular movement of air through a *constriction* in the vocal tract which causes a noise [99.500].

typicality – The extent to which the properties of the voice on the questioned-speaker recording resemble those of the voices on recordings of a sample of speakers from the relevant population [99.150].

validity – The extent to which a measurement or estimate approximates the true value (synonymous with *accuracy*) [99.290].

velopharyngeal port – The opening between the *pharyngeal* and *nasal cavities* [99.450].

velum – The soft palate. Part of the vocal tract [99.450].

vocal folds – The structure in the *larynx* which can be made to vibrate to produce *voicing* [99.450].

vocal tract – The *larynx*, throat (*pharyngeal cavity*), mouth (*oral cavity*), and nose (*nasal cavities*) when used to make speech sounds [99.450].

voice – Sounds produced by a human vocal tract for the purpose of communication (used synonymously with *speech*). *Voice* can also be used to refer more specifically to *laryngeal* and *vocal fold* activity.

voice lineup – The presentation of a series of voice recordings to an *earwitness* to see whether the *earwitness* recognises the voice of a speaker they heard earlier [99.960].

voice-onset time (VOT) – The time between the 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*. This particular term may be considered pejorative [99.680].

voicing – The vibration of the *vocal folds* used in the production of speech sounds [99.460].

vowel – A speech sound made with an open oral cavity. *Voicing* produced at the *vocal folds* excites the resonance frequencies of the *supralaryngeal vocal tract* [99.460].

References

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

Alexander A., Botti F., Dessimoz D., Drygajlo A. (2004). The effect of mismatched recording conditions on human and automatic speaker recognition in forensic applications. *Forensic Science International*, 146S, pp. S95–S99. http://dx.doi.org/10.1016/j.forsciint.2004.09.078

Alias D., Best V., Niall P.D., Semmler C., Woolford D.H. (2015). Hearing and the perception of sound. In: Freckelton I., Selby H. (Eds.), *Expert evidence*. Sydney: Thomson Reuters. Ch. 69.

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. http://dx.doi.org/10.1016/j.scijus.2009.07.004

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

Baumeister B., Heinrich C., Schiel F. (2012). The influence of alcoholic intoxication on the fundamental frequency of female and male speakers. Jo*urnal of the Acoustical Society of America*, 132, pp. 442–451. http://dx.doi.org/10.1121/1.4726017

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

Berger C.E.H., Robertson B., Vignaux G.A. (2016). Interpreting scientific evidence. In: Freckelton I., Selby H. (Eds.), *Expert evidence*. Sydney: Thomson Reuters. Ch. 28.

Betancourt K.S., Bahr R.H. (2010). The influence of signal complexity on speaker identification. *International Journal of Speech, Language and the Law*, 17, pp. 179–200. http://dx.doi.org/10.1558/ijsll.v17i2.179

Broeders A.P.A., van Amelsvoort A.G. (1999). Lineup construction for forensic earwitness identification: A practical approach. *Proceedings of the International Congress of Phonetic Sciences*. pp. 1373–1376.

Broeders A.P.A., van Amelsvoort A.G. (2001). A practical approach to forensic earwitness identification: Constructing a voice line-up. *Problems of Forensic Sciences*, 47, pp. 237–245.

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

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.

Butcher A. (1996. Getting the voice line-up right- Analysis of a multiple auditory confrontation. *Proceedings of the 6th Australasian International Conference on Speech Science & Technology*, pp. 97–102.

Cáo Hónglín 曹洪林, Lǐ Jìngyáng 李敬陽, Wáng Yīnglì 王英利, Kǒng Jiāngpíng 孔江平 (2013). Lùn shēngwén jiàndìng yìjiàn de biǎoshù xíngshì 論聲紋鑒定意見的表述形式 [On Expert Opinion of Forensic Speaker Identification], Zhèngjù Kēxué 證據科學 [Evidence Science], 21, 605–624

Cambier-Langeveld 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. http://dx.doi.org/10.1558/ijsll.2007.14.2.223

Cambier-Langeveld T. (2010). The role of linguists and native speakers in language analysis for the determination of speaker origin. *International Journal of Speech, Language and the Law*, 17, pp. 67–93. http://dx.doi.org/10.1558/ijsll.v17i1.67

Clark J., Foulkes P. (2007). Identification of voices in electronically disguised speech. *International Journal of Speech, Language and the Law*, 14, pp. 195–221. http://dx.doi.org/10.1558/ijsll.2007.14.2.195

Clarke F.R., Becker R.W. (1969). Comparison of techniques for discriminating among talkers. *Journal of Speech and Hearing Research*, 12, pp. 747–761. http://dx.doi.org/10.1044/jshr.1204.747

Curran J.M. (2016). Admitting to uncertainty in the LR, *Science & Justice*, 56, pp. 380–382. http://dx.doi.org/10.1016/j.scijus.2016.05.005

Damphousse K.R., Pointon L., Upchurch D., Moore R.K. (2007). Assessing the validity of voice stress analysis tools in a jail setting. Report submitted to the U.S. Department of Justice. http://www.ncjrs.gov/pdffiles1/nij/grants/219031.pdf (last visited March 2017)

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

Drygajlo A., Jessen M., Gfroerer S., Wagner I., Vermeulen J., Niemi T. (2015). Methodological guidelines for best practice in forensic semiautomatic and automatic speaker recognition, including guidance on the conduct of proficiency testing and collaborative exercises, European Network of Forensic Science Institutes. http://enfsi.eu/wp-content/uploads/2016/09/guidelines_fasr_and_fsasr_0.pdf (last visited October 2017).

Edmond G., Martire K., San Roque M. (2011a). 'Mere guesswork': Cross-lingual voice comparisons and the Jury. *Sydney Law Review*, 33, pp. 395–425.

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

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

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

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

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

Enzinger E. (2016). *Implementation of forensic voice comparison within the new paradigm for the evaluation of forensic evidence*. Doctoral dissertation, University of New South Wales. http://handle.unsw.edu.au/1959.4/55772 (last visited February 2017).

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

Enzinger E., Morrison G.S. (2015). Mismatched distances from speakers to telephone in a forensic-voice-comparison case. *Speech Communication*, 70, pp. 28–41. http://dx.doi.org/10.1016/j.specom.2015.03.001

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

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

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

Eriksson A., Lacerda F. (2007). Charlatantry in forensic speech science: A problem to be taken seriously. *International Journal of Speech, Language and the Law*, 14, pp. 169–193. http://dx.doi.org/10.1558/ijsll.2007.14.2.169

Faigman D.L., Blumenthal J.A., Cheng E.K., Mnookin J.L., Murphy E.E., Sanders, J. (2015). Talker identification: I. Legal Issues. In: Faigman D.L., Saks M.J., Sanders J., Cheng E.K., (Eds.), *Modern Scientific Evidence: The Law and Science of Expert Testimony*, vol. 5, §36.1–36.3.

Federle L., Pedersen N., Pollett A. (2017). Statistical evaluation in forensic DNA typing. In: Freckelton, I. Selby, H. eds. *Expert evidence*. Sydney: Thomson Reuters. Ch. 80A.

Fernández Gallardo L. (2014). *Human and automatic speaker recognition over telecommunication channels*. Doctoral dissertation, University of Canberra. https://www.canberra.edu.au/researchrepository/file/6e3d646b-ceb7-4f91-94bc-8944ca86c902/1/full_text.pdf (last visited February 2017).

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

Forensic Science Regulator (2016). *Codes of practice and conduct for forensic science providers and practitioners in the criminal justice system (version 3.0)*, Forensic Science Regulator, Birmingham, UK. https://www.gov.uk/government/publications/forensic-science-providers-codes-of-practice-and-conduct-2016 (last visited 27 January 2017).

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

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

Fraser H. (2010). Transcripts in the legal system. In: Freckelton I., Selby H. (Eds.), *Expert evidence*. Sydney: Thomson Reuters. Ch. 100.

Fraser H., Stevenson B., Marks T. (2011). Interpretation of a crisis call: Persistence of a primed perception of a disputed utterance. *International Journal of Speech, Language and the Law*, 18, pp. 261–292. http://dx.doi.org/10.1558/ijsll.v18i1.145

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. http://dx.doi.org/10.1558/ijsll.v17i1.143

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

González-Hautamäki R., Hautamäki V., Rajan P., Kinnunen T. (2013). Merging human and automatic system decisions to improve speaker recognition performance. *Proceedings of INTERSPEECH 2013*. International Speech Processing Association pp. 2519–2523. http://isca-speech.org/archive/interspeech_2013/i13_2519.html (last visited February 2017).

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. http://dx.doi.org/10.1109/TASL.2007.902747

Greenberg C., Martin A., Brandschain L., Campbell J., Cieri C., Doddington G., Godfrey J. (2010). Human assisted speaker recognition in NIST SRE10. *Proceedings of Odyssey 2010, The Speaker and Language Recognition Workshop*. International Speech Processing Association. pp. 180–185. http://isca-speech.org/archive_open/odyssey_2010/od10_032.html (last visited February 2017).

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. http://dx.doi.org/10.1558/ijsll.v15i2.193

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

Harnsberger J.D., Hollien H., Martin C.A., Hollien K.A. (2009). Stress and deception in speech: Evaluating layered voice analysis. *Journal of Forensic Sciences*, 54, pp. 642–650. http://dx.doi.org/10.1111/j.1556-4029.2009.01026.x

Hautamäki V., Kinnunen T., Nosratighods M., Lee K.A., Ma B., Li H. (2010). Approaching human listener accuracy with modern speaker verification. *Proceedings of Interspeech 2010*. International Speech Processing Association. http://isca-speech.org/archive/interspeech_2010/i10_1473.html (last visited February 2017).

Hicks T., Buckleton J.S., Bright J.A., Taylor D. (2016). A Framework for interpreting evidence. In: Buckleton J.S., Bright J.A., Taylor D (Eds.), *Forensic DNA Evidence Interpretation* (2nd Ed.). Boca Raton, FL: CRC. Ch. 2.

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

Hollien H. (1996). Consideration of guidelines for earwitness lineups. *Forensic Linguistics*, 3, 14–23. http://dx.doi.org/10.1558/ijsll.v3i1.14

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

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

Hollien H., de Jong G., Martin C.A., Schwartz R., Liljegren K. (2001). Effects of ethanol intoxication on speech suprasegmentals. *Journal of the Acoustical Society of America*, 110, pp. 3198–3206. http://dx.doi.org/10.1121/1.1413751

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

Hollien H., Harnsberger J.D., Martin C.A., Hollien K.A. (2008). Evaluation of the NITV CVSA. *Journal of Forensic Sciences*, 53, pp. 183–193. http://dx.doi.org/10.1111/j.1556-4029.2007.00596.x

Hollien H., Huntley R., Künzel H., Hollien P.A. (1995). Criteria for earwitness lineups. *Forensic Linguistics*, 2, 143–153. http://dx.doi.org/10.1558/ijsll.v2i2.143

Hollien H., Huntley Bahr R., Harnsberger J.D. (2014). Issues in forensic voice. *Journal of Voice*, 28, pp. 170–184. http://dx.doi.org/10.1016/j.jvoice.2013.06.011

Horvath F., McCloughan J., Weatherman D., Slowik S. (2013). The accuracy of auditors' and layered voice analysis (LVA) operators' judgments of truth and deception during police questioning. *Journal of Forensic Sciences*, 58, pp. 385–392. http://dx.doi.org/10.1016/10.1111/1556-4029.12066

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

Innes B. (2011). R v David Bain – A unique case in New Zealand legal and linguistic history. *International Journal of Speech, Language and the Law*, 18, pp. 145–155. http://dx.doi.org/1010.1558/ijsll.v18i1.145

Jessen M. (2007). Speaker classification in forensic phonetics and acoustics. In: Müller C. (Ed.), *Speaker Classification I*, Berlin: Springer, pp. 180–204. http://dx.doi.org/10.1007/978-3-540-74200-5_10

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

Jessen M. (2010). The forensic phonetician: Forensic speaker identification by experts. In: Coultard M., Johnson A. (Eds.), *The Routledge Handbook of Forensic Linguistics*. London: Routledge, pp. 378–394. http://dx.doi.org/10.4324/9780203855607.ch25

Jessen M. (2012). *Phonetische und Linguistische Prinzipien des Forensischen Stimmenvergleichs [Phonetic and linguistic principles of forensic voice comparison]*. Munich, Germany: Lincom.

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

Kassin S.M., Dror I.E., Kukucka J. (2013). The forensic confirmation bias: Problems, perspectives, and proposed solutions. *Journal of Applied Research in Memory and Cognition*, 2, pp. 42–51. http://dx.doi.org/10.1016/j.jarmac.2013.01.001

Kaye D.H., Bernstein D.A., Mnookin J.L. (2011). Quantitative testimony on forensic science identification. In: *The New Wigmore: A Treatise on Evidence: Expert Evidence*, 2nd ed., New York: Wolters Kluwer, Ch. 14.

Kerstholt J.H., Jansen N.J.M., Van Amelsvoort A.G., Broeders A.P.A. (2006). Earwitnesses: Effects of accent, retention and telephone. *Applied Cognitive Psychology*, 20, pp. 187–197. http://dx.doi.org/10.1002/acp.1175

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

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. (2014). Forensic fallacies and a famous judge. *Jurimetrics*, 54, pp. 211–219.

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.

Kuhn T.S. (1970). *The structure of scientific revolutions*, 2nd Ed. Chicago: University of Chicago Press.

Künzel H., González-Rodríguez J., Ortega-García J. (2004). Effect of voice disguise on the performance of a forensic automatic speaker recognition system. In: *Proceedings of Odyssey*

2004 The Speaker and Language Recognition Workshop. http://isca-speech.org/archive_open/odyssey_04/ody4_153.html (last visited 22 May 2017).

Labov W., Harris W.A. (1994). Addressing social issues through linguistic evidence. In: Gibbons J. (Ed.), *Language and the Law*. Harlow, UK: Longman.

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

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

Ladefoged P., Disner, S.F. (2012). *Vowels and consonants: An introduction to the sounds of language* (3rd Ed.). Chichester, UK: Wiley.

Ladefoged P., Johnson, K. (2014). *A course in phonetics* (7th Ed.). Stamford, CT: Cengage.

Language and National Origin Group (2004). Guidelines for the use of language analysis in relation to questions of national origin in refugee cases. *Speech, Language and the Law*, 11, pp. 261–266. http://dx.doi.org/10.1558/ijsll.v11i2.261

Laub C.E. (2010). Can earwitness limitations be overcome by the court system? Strategies to help mock jurors appreciate the limitations of earwitness testimony. PhD dissertation. University of Nebraska.

Laubstein A.S. (1997). Problems of voice lineups. *Forensic Linguistics*, 4, pp. 262–279. http://dx.doi.org/10.1558/ijsll.v4i2.262

Laver J. (1994). *Principles of phonetics*. Cambridge, UK: Cambridge University Press.

Lindh J. (2017). Forensic comparison of voices, speech and speakers tools and methods in forensic phonetics. PhD dissertation. University of Gothenburg.

Marks D.B. (2017. A framework for performing forensic and investigatory speaker comparisons using automatic methods. MSc thesis. University of Colorado Denver.

Matějka P., Glembek O., Plchot O., Schwarz M., Cipr T., Cumani S., Kudla R., Szöke I., Svobodová M., Malý K., Černocký J. (2012). BUT HASR'12 experience: Are developers of SRE systems naïve listeners? Technical Report, Brno University of Technology. http://www.fit.vutbr.cz/research/view_pub.php?id=10777 (last visited February 2017).

McDougall K., Nolan F., Hudson T. (2015). Telephone transmission and earwitnesses-performance on voice parades controlled for voice similarity. *Phonetica*, 72, pp. 257–272. http://dx.doi.org/10.1159/000439385

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

Meuwly D. (2003a). Le mythe de l'empreinte vocale I. *Revue Internationale de Criminologie et Police Technique*, 56, pp. 219–236.

Meuwly D. (2003b). Le mythe de l'empreinte vocale II. *Revue Internationale de Criminologie et Police Technique*, 56, pp. 361–374.

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.

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

Mohammadi S.H., Kain A. (2017). An overview of voice conversion systems. *Speech Communication*, 88, pp. 65–82. http://dx.doi.org/10.1016/j.specom.2017.01.008

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

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

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

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

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

Morrison G.S., Enzinger E. (2016). Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (forensic_eval_01) – Introduction, *Speech Communication*, 85, pp. 119–126. http://dx.doi.org/10.1016/j.specom.2016.07.006

Morrison G.S., Enzinger E. (2018). An introduction to forensic voice comparison. In W.F. Katz, P.F. Assmann (Eds.), *Routledge Handbook of Phonetics*, Oxon, UK: Routledge.

Morrison G.S., Enzinger E., Zhang C. (2016). Refining the relevant population in forensic voice comparison – A response to Hicks et alii (2015) The importance of distinguishing information from evidence/observations when formulating propositions, *Science & Justice*, 56, pp. 492–497. http://dx.doi.org/10.1016/j.scijus.2016.07.002 [see also Morrison, G.S., Enzinger, E., Zhang, C. (2017). Reply to Hicks et alii (2017) http://arxiv.org/abs/1704.07639]

Morrison G.S., Hoy M. (2012). What did Bain really say? A preliminary forensic analysis of the disputed utterance based on data, acoustic analysis, statistical models, calculation of likelihood ratios, and testing of validity. In *Proceedings of the 46th Audio Engineering Society (AES) Conference on Audio Forensics: Recording, Recovery, Analysis, and Interpretation*, pp. 203–207. Stable URL: http://www.aes.org/e-lib/browse.cfm?elib=16331

Morrison G.S., Kaye D.H., Balding D.J., Taylor D., Dawid P., Aitken C.G.G., Gittelson S., Zadora G., Robertson B., Willis S.M., Pope S., Neil M., Martire K.A., Hepler A., Gill R.D., Jamieson A., de Zoete J., Ostrum R.B., Caliebe A. (2017). A comment on the PCAST report: Skip the "match"/"non-match" stage. *Forensic Science International*, 272, pp. e7–e9. http://dx.doi.org/10.1016/j.forsciint.2016.10.018

Morrison G.S., Lindh J., Curran J.M. (2014). Likelihood ratio calculation for a disputed-utterance analysis with limited available data. *Speech Communication*, 58, pp. 81–90. http://dx.doi.org/10.1016/j.specom.2013.11.004

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

Morrison G.S., Poh N. (2017). Avoiding overstating the strength of forensic evidence: Shrunk likelihood ratios / Bayes factors. Submitted manuscript (copy available on request).

Morrison G.S., Thompson W.C. (2017). Assessing the admissibility of a new generation of forensic voice comparison testimony, *Columbia Science and Technology Law Review*, 18, 326–434. http://www.stlr.org/cite.cgi?volume=18&article=morrisonThompson

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.

Nolan F. (2003). A recent voice parade. *Forensic Linguistics*, 10, 277–291. http://dx.doi.org/10.1558/sll.2003.10.2.277

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

Nolan F., Grabe E. (1996). Preparing a voice lineup. *Forensic Linguistics*, 3, 74–95. http://dx.doi.org/10.1558/ijsll.v3i1.74

Nolan F., McDougall K., Hudson T. (2013). Effects of the telephone on perceived voice similarity: Implications for voice line-ups. *International Journal of Speech, Language and the Law*, 20, 229–246. http://dx.doi.org/10.1558/ijsll.v20i2.229

Öhman L., Eriksson A., Granhag P.A. (2010). Mobile phone quality vs direct quality: How the presentation format affects earwitness identification accuracy. *European Journal of Psychology Applied to Legal Context*, 2, pp. 161–182.

Öhman L., Eriksson A., Granhag P.A. (2013). Angry voices from the past and present: Effects on adults' and children's earwitness testimony. *Journal of Investigative Psychology and Offender Profiling*, 10, 57–70. http://dx.doi.org/10.1002/jip.1381

Perrot P., Chollet G. (2008). The question of disguised voice. *Proceedings of Acoustics 08*, pp. 9681–9685. http://www.conforg.fr/acoustics2008/cdrom/data/articles/002691.pdf (last visited 22 May 2017).

Perrot P., Chollet G. (2012). Helping the forensic research institute of the French Gendarmerie to identify a suspect in the presence of voice disguise or forgery. In: Neustein A., Patil H.A. (Eds.), *Forensic Speaker Recognition: Law Enforcement and Counter-Terrorism*, pp. 469–503, New York: Springer. http://dx.doi.org/10.1007/978-1-4614-0263-3_16

Perrot P., Razik J., Chollet G. (2009). Vocal forgery in forensic sciences. In: M. Sorell (Ed.), *Forensics in Telecommunications, Information and Multimedia: Second International Conference on e-Forensics*, pp. 179–185, Berlin: Springer. http://dx.doi.org/10.1007/978-3-642-02312-5

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. http://dx.doi.org/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.

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

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

Ramos D., Franco-Pedroso J., González-Rodríguez J. (2011). Calibration and weight of the evidence by human listeners. The ATVS-UAM submission to NIST human-aided speaker recognition 2010. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing* (ICASSP 2011), pp. 5908–5911. http://dx.doi.org/10.1109/ICASSP.2011.5947706

Rathborn H., Bull R., Clifford B.R. (1981). Voice recognition over the telephone. *Journal of Police Science and Administration*, 9, pp. 280–284.

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

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

Rogers H. (2000). *The sounds of language: An introduction to phonetics*. Harlow, UK: Pearson Education. (Republished: 2013. London, UK: Routledge.)

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. http://dx.doi.org/10.1016/j.csl.2005.07.003

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

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

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

Saeidi R., van Leeuwen D.A. (2012). The Radboud University Nijmegen submission to NIST SRE-2012. Technical Report. https://users.aalto.fi/~saeidir1/file_library/SRE12.pdf (last visited February 2017).

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

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

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

Sarwar F., Allwood C.M., Zetterholm E. (2014). Earwitnesses: The type of voice lineup affects the proportion of correct identifications and the realism in confidence judgments. *International Journal of Speech, Language and the Law*, 21, pp. 139–155. http://dx.doi.org/10.1558/ijsll.v21i1.139

Schiel F., Heinrich C. (2015). Disfluencies in the speech of intoxicated speakers. *International Journal of Speech, Language and the Law*, 22, pp. 19–34. http://dx.doi.org/10.1558/ijsll.v22i1.24767

Schilling N., Marsters A. (2015). Unmasking identity: Speaker profiling for forensic linguistic purposes. *Annual Review of Applied Linguistics*, 35, pp. 195–214. http://dx.doi.org/10.1017/S0267190514000282

Schwartz R., Campbell J.P., Shen W., Sturim D.E., Campbell W.M., Richardson F.S., Dunn R.B., Granville R. (2011). USSS-MITLL 2010 Human assisted speaker recognition. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP 2011)*. Institute of Electrical and Electronics Engineers (IEEE). pp. 5904–5907. https://dx.doi.org/10.1109/ICASSP.2011.5947705

Shaw F., Crocker V. (2015). Creaky voice as a stylistic feature of young American female speech: an intraspeaker variation study of Scarlett Johansson. *Lifespans & Styles*, 1, article 3. http://dx.doi.org/10.2218/ls.v1i0.2015.1180

Sherrin C. (2015). Earwitness evidence: The reliability of voice identifications. *Osgoode Legal Studies Research Paper Series*. Paper 101. http://digitalcommons.osgoode.yorku.ca/olsrps/101 (last visited February 2017).

Silva G.D. da, Medina C.A. (2017). Evaluation of MSR Identity Toolbox under conditions reflecting those of a real forensic case (forensic_eval_01). *Speech Communication*, 94, 42–49. http://dx.doi.org/10.1016/j.specom.2017.09.001

Solewicz Y.A., Becker T., Jardine G., Gfroerer S. (2012). Comparison of speaker recognition systems on a real forensic benchmark. In: *Proceedings of Odyssey 2012, The Speaker and Language Recognition Workshop*. International Speech Communication Association. pp. 85–91. http://isca-speech.org/archive/odyssey_2012/od12_086.html (last visited February 2017).

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

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

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

Thompson W.C., Schumann E.L. (1987). Interpretation of statistical evidence in criminal trials: The prosecutor's fallacy and the defense attorney's fallacy. *Law and Human Behavior*, 11, pp. 167–187. http://dx.doi.org/ 10.1007/BF01044641

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

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. http://dx.doi.org/10.1121/1.1913064

van der Vloed, D. (2016). Evaluation of Batvox 4.1 under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01). *Speech Communication*, 85, pp. 127–130. http://dx.doi.org/10.1016/j.specom.2016.10.001 [errata published in (2017) 92, p. 23. http://dx.doi.org/10.1016/j.specom.2017.04.005]

van der Vloed D., Bouten J., van Leeuwen D. (2014). NFI-FRITS: A forensic speaker recognition database and some first experiments. In *Proceedings of Odyssey 2014, The Speaker and Language Recognition Workshop*. International Speech Communication Association. pp. 6–13. http://cs.uef.fi/odyssey2014/program/pdfs/21.pdf (last visited February 2017).

van Wallendael L.R., Surace A., Hall-Parsons D., Brown M. (1994). 'Earwitness' voice recognition: Factors affecting accuracy and impact on jurors. *Applied Cognitive Psychology*, 8, 661–677.

Willis S.M., McKenna L., McDermott S., O'Donell G., Barrett A., Rasmusson A., Nordgaard A., Berger C.E.H., Sjerps M.J., Lucena-Molina J.J., Zadora G., Aitken C.G.G., Lunt L., Champod C., Biedermann A., Hicks T.N., Taroni F. (2015). *ENFSI guideline for evaluative reporting in forensic science*, European Network of Forensic Science Institutes. http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf (last visited October 2017).

Wu H., Wang Y., Huang J. (2014). Identification of electronic disguised voices. *IEEE Transactions on Information Forensics and Security*, 9, pp. 489–500. http://dx.doi.org/10.1109/TIFS.2014.2301912

Yarmey A.D. (1995). Earwitness speaker identification. *Psychology, Public Policy and Law*, 1, pp. 792–816. http://dx.doi.org/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. http://dx.doi.org/10.1002/acp.702

Yarmey A.D. (2004). Common-sense beliefs, recognition and the identification of familiar and unfamiliar speakers from verbal and non-linguistic vocalizations. *Speech, Language and the Law*, 11, 1350–1771.

Zetterholm E., Sarwar F., Thorvaldsson V., Allwood C.M. (2012). Earwitnesses: The effect of type of vocal differences on correct identification and confidence accuracy. *International Journal of Speech, Language and the Law*, 19, pp. 219–237. http://dx.doi.org/10.1558/ijsll.v19i2.219

Zhang C. (2009). *法庭语音技术研究 [Forensic Speech Technology Research]*. 中国社会出版社 [China Social Press].

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

Zhang C., Morrison G.S. (2017). Forensic voice comparison. In: Sybesma, R., Behr, W., Gu, Y., Handel, Z., Huang, C.-T. J., Myers, J. (Eds.), *Encyclopedia of Chinese Language and Linguistics*. Leiden: Brill. pp. 256–260. http://dx.doi.org/10.1163/2210-7363_ecll_COM_000205

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

Zhang C., Morrison G.S., Enzinger E. (2016). Use of relevant data, quantitative measurements, and statistical models to calculate a likelihood ratio for a Chinese forensic voice comparison case involving two sisters. *Forensic Science International*, 267, pp. 115–124. http://dx.doi.org/10.1016/j.forsciint.2016.08.017

Zhang C., Tan T. (2008). Voice disguise and automatic speaker recognition. *Forensic Science International*, 175, pp. 118–122. http://dx.doi.org/10.1016/j.forsciint.2007.05.019