Assessing the Admissibility of a New Generation of Forensic Voice Comparison Testimony

Geoffrey Stewart Morrison [a,b,c] & *William C Thompson* [c,d]

[a] Independent Forensic Consultant, Vancouver, British Columbia, Canada
[b] Department of Linguistics, University of Alberta, Edmonton, Alberta, Canada
[c] Isaac Newton Institute for Mathematical Sciences, Cambridge, England, United Kingdom
[d] Department of Criminology, Law & Society, and School of Law, University of California, Irvine, California, USA

Abstract

This article provides a primer on forensic voice comparison (aka forensic speaker recognition), a branch of forensic science in which the forensic practitioner analyzes a voice recording in order to provide an expert opinion that will help the trier-of-fact determine the identity of the speaker. The article begins with an explanation of ways in which human speech varies within and between speakers. It then discusses different technical approaches that forensic practitioners have used to compare voice recordings, and frameworks of reasoning that practitioners have used for evaluating the evidence and reporting its strength. It then discusses procedures for empirical validation of the performance of forensic voice comparison systems. It also discusses the potential influence of contextual bias and ways to reduce this. Building on this scientific foundation, the article then offers analysis, commentary, and recommendations on how courts evaluate the admissibility of forensic voice comparison testimony under the *Daubert* and *Frye* standards. It reviews past rulings such as *U.S. v. Angleton*, 269 F.Supp 2nd 892 (S.D. Tex. 2003) that found expert testimony based on the spectrographic approach inadmissible under *Daubert*. The article also offers a detailed analysis of the evidence presented in the recent *Daubert* hearing in *U.S. v. Ahmed, et al*. 2015 EDNY 12-CR-661, which included testimony based on the newer automatic approach. The scientific testimony proffered in *Ahmed* is used to illustrate the issues courts are likely to face when considering the admissibility of forensic voice comparison testimony in the future. The article concludes with a discussion of how proponents of forensic voice comparison testimony might meet a reasonably rigorous application of the *Daubert* standard and thereby ensure that such testimony is sufficiently trustworthy to be used in court.

Keywords: forensic voice comparison, voiceprint, automatic speaker recognition, admissibility, Daubert, Frye

This manuscript is a preprint version (version 2016-12-11a) of: Geoffrey Stewart Morrison & William C. Thompson, *Assessing the admissibility of a new generation of forensic voice comparison testimony*, 18 Columbia Science and Technology Law Review (2017) http://stlr.org/. There may be formatting and content difference between the current version and the final published version. © Geoffrey Stewart Morrison & William C. Thompson 2016

1 Introduction

In criminal and civil cases, disputes sometime arise about the identity of a speaker on an audio recording. In such cases a forensic practitioner may be asked to perform a forensic voice comparison.[1] This involves comparing recordings of one or more known speakers with a recording of a speaker of questioned identity. The goal is to provide an expert opinion that will help the trier-of-fact determine the identity of that speaker.

Forensic voice comparison testimony has a long and troubled history in the United States. In the 1970s, 80s, and 90s, courts frequently admitted forensic voice comparison testimony.[2] The testimony of that era was typically based on the *spectrographic* approach or *auditory-spectrographic* approach.[3,4] Almost

[1] *Forensic voice comparison* is our preferred term; see Geoffrey Stewart Morrison, *Forensic voice comparison*, *in* Expert Evidence ch. 99 (Ian Freckelton & Hugh Selby 2010) (hereinafter Morrison 2010) at §99.170 for reasons. Other terms which have been used include: *forensic speaker comparison*, *forensic speaker recognition*, *forensic speaker identification*, *forensic voice identification*, *forensic talker identification*, *voiceprint identification*, and *voicegram identification*. The differences between the different terms may reflect subtle philosophical differences, but in general the courts can simply interpret all these terms as equivalent (see, however, note 26 *infra*).

[2] Based on published rulings, the rate of admission appears to have been somewhat greater than the rate of exclusion. In the rulings listed in Table 1 of David L. Faigman, Jeremy A. Blumenthal, Edward K. Cheng, Jennifer L. Mnookin, Erin E. Murphy, & Joseph Sanders (2015) *Talker identification: I. Legal Issues*, *in* Modern Scientific Evidence: The Law and Science of Expert Testimony vol. 5 ch. 36 §37.1–37.3 (David L. Faigman, Michael J. Saks, Joseph Sanders, & C. Edward K. Cheng 2015) (hereinafter Faigman *et al*. 2015), between 1967 and 1999 the counts were 22 versus 15 for admission versus exclusion.

[3] When we use the term *approach* in the singular with the definite article, for example, *the* acoustic-phonetic *approach*, this is a cover term for all *methods* which could be classed as acoustic-phonetic. For example, one method could be based on format measurements and another method could be based on fundamental frequency measurements, but they would both be classed as acoustic-phonetic approaches (see Section 2.3.3). When we use the term *approach* in the plural or in the singular without the definite article, *e.g.*, acoustic-phonetic *approaches* or an acoustic-phonetic *approach*, its meaning is interchangeable with *method*. We use the term *system* to mean a concrete implementation of a method. Different approaches to forensic voice comparison can be conceptualized as broadly different ways of extracting information from speech recordings. We use the term *framework* to refer to different ways of making use of that information in order to derive a strength of evidence statement (see Section 3). Although in practice there may be correlation between the use of particular approaches and particular frameworks, approaches and frameworks are in principle orthogonal to one another. We use the term *paradigm* to subsume a particular combination of approach and framework, and "the entire constellation of beliefs, values, techniques, and so on shared by the members of a given community" Thomas S. Kuhn, The Structure of Scientific Revolutions (2nd ed. 1970) (hereinafter Kuhn 1970) at p. 175. The terms *system*, *approach*, *framework* and *paradigm* are used consistently with these meanings in the writings of Morrison.

[4] See descriptions of auditory and spectrographic approaches in Sections 2.3.1 and 2.3.2. Spectrograms are graphical representations of the acoustic properties of short sections of recordings of speech. The *auditory-spectrographic* approach (also called the *aural-spectrographic* approach) involves both listening to the audio recordings and looking at spectrograms. In the early 1970s, there was debate about whether it was better to use a visual only or a visual plus auditory approach. The latter won out. We are not concerned with this debate in the present paper, and since it is not always clear from published rulings which of the two was actually used, we will often use either *spectrographic* or *auditory-spectrographic* as a cover for both approaches. Based on published rulings, auditory-only approaches appear to have seldom been presented to U.S. courts, and they do not appear to have ever been admitted under the standards set forth in *William Daubert et al. v Merrell Dow Pharmaceuticals Inc*., 509 US 579 (1993) (hereinafter *Daubert*). Auditory-only approaches were proffered but excluded in: *U.S. v. Jones*, 24 F.3d 1177 (9th Cir. 1994); and in *U.S. v. Salimonu*, 182 F.3d 63 (1st Cir. 1999). Harry Hollien, *An approach to speaker identification*, 61 Journal of Forensic Sciences 334–344 (2016) http://dx.doi.org/10.1111/1556-4029.13034 (hereinafter Hollien 2016) includes a statement that auditory approaches "have satisfied Daubert ... in well over 150 cases and 40 trials", but provides no references to substantiate this claim. We requested references from the author, but they were not

from its inception, however, this testimony was soundly criticized by members of the scientific community.[5] Following a 1979 National Research Council (hereinafter NRC) report,[6] the FBI stopped using the spectrographic approach in court,[7] and the number of reported cases in which it was used by others gradually fell to a trickle.[8]

After an extensive *Daubert* hearing[9] in *U.S. v. Angleton* (2003),[10] where the defense attempted to introduce conclusions reached using the auditory-spectrographic approach, a federal judge ruled the testimony inadmissible, finding specifically that:

> The testimony and evidence show that voice identification techniques using the aural spectrographic method are not widely accepted by the scientific community. ... there is great dispute among researchers and the few practitioners in the field over the accuracy and reliability of voice spectrographic analysis ... The evidence also shows that error rates for voice spectrographic techniques are unknown and vary widely depending on the conditions under which the analysis is made.[11]

Since *Angleton*, there are no reported cases in which testimony based on the spectrographic approach has overcome a *Daubert* challenge.[12]

---

provided.

[5] Richard A. Bolt, Franklin S. Cooper, Edward E. David Jr., Peter B. Denes, James M. Pickett, Kenneth N. Stevens, *Speaker identification by speech spectrograms: a scientists' view of its reliability for legal purposes*, 47 Journal of the Acoustical Society of America 597–612 (1970) http://dx.doi.org/10.1121/1.1911935 (hereinafter Bolt *et al*. 1970); Richard A. Bolt, Franklin S. Cooper, Edward E. David Jr., Peter B. Denes, James M. Pickett, Kenneth N. Stevens, *Speaker identification by speech spectrograms: some further observations*, 54 Journal of the Acoustical Society of America 531–534 (1973) http://dx.doi.org/10.1121/1.1913613 (hereinafter Bolt *et al*. 1973)

[6] National Research Council, On the Theory and Practice of Voice Identification (1979) (hereinafter NRC 1979)

[7] According to Dr Hirotaka Nakasone, Senior Scientist, Digital Evidence Section, FBI laboratory, the FBI continued using the spectrographic approach for investigative purposes until 2011. The laboratory then abandoned this approach in favor automatic approaches, but still only for investigative purposes (personal communication 30 November 2011).

[8] See Jordan S. Gruber, Fausto T. Poza, *Voicegram identification evidence*, 54 American Jurisprudence Trials 1 (1995) (hereinafter Gruber & Poza 1995), Geoffrey Stewart Morrison, *Distinguishing between forensic science and forensic pseudoscience: Testing of validity and reliability, and approaches to forensic voice comparison*, 54 Science & Justice 245–256 (2014) http://dx.doi.org/10.1016/j.scijus.2013.07.004 (hereinafter Morrison 2014), and references cited therein for the history of the scientific debate. Gruber & Poza 1995 is the most comprehensive published review of the controversy over the use of the spectrographic approach.

[9] When a litigant challenges the admissibility of expert evidence under the *Daubert* standard, *supra* note 3, the judge may hold a hearing (called a *Daubert* hearing) at which the parties may present evidence and argument outside the presence of the jury on whether the expert's reasoning and methodology are sufficiently valid to meet the *Daubert* standard.

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

[11] *Ibid*. at 905.

[12] Prior to *Angleton*, there were at least three published rulings on the admissibility of the spectrographic approach under *Daubert*. The spectrographic approach was ruled admissible in *U.S. v. Salimonu*, 182 F.3d 63 (1st Cir. 1999), and in *State v. Coon*, 974 P.2d 386 (Alaska 1999), and inadmissible in *U.S. v. Bahena et al.*, 223 F.3d 797 (8th Cir. 2000), but the latter two

Judicial rejection of the spectrographic approach does not, however, mean the end of forensic voice comparison testimony. Over the last 15–20 years there have been substantial advances in automatic speaker recognition technology and in the application of this technology to forensic voice comparison.[13] In April 2015, in a terrorism prosecution in federal district court in New York (*Ahmed*),[14] the prosecution attempted to introduce testimony by a forensic practitioner who had, in part, used an approach based on automatic speaker recognition (he also used auditory and acoustic-phonetic approaches). Because the automatic approach is fundamentally different from the auditory-spectrographic approach considered in cases like *Angleton*, this testimony could not be dismissed out of hand and required an extensive *Daubert* hearing. No ruling was issued, however, because soon after the hearing the case was resolved through a negotiated plea. Nevertheless, the evidence offered in the *Daubert* hearing is worth careful consideration because testimony based on automatic approaches will surely be offered in other cases in the not too distant future. Due to its complexity, deciding whether to admit such testimony and what weight it deserves will be challenging for the courts.

The present article is designed to guide lawyers and judges in their evaluation of the new generation of forensic voice comparison testimony. We begin with a primer on forensic voice comparison. We describe

---

rulings explicitly stated that they were specific to those particular cases and not generalizable to other cases in which the circumstances could be different.

U.S. v. Drones, 218 F.3d 496 (5th Cir. 2000) denied a petition for federal habeas corpus relief for ineffective assistance of counsel where defense counsel had failed to call a voice comparison expert; the court concluded that failure to call such an expert was not unreasonable "given the uncertainty of the current state of the law regarding the reliability and admissibility of expert voice identification evidence, and the vulnerability of such expert testimony at trial." 218 F.3d at 504.

Post *Angleton*, use of the spectrographic approach was ruled inadmissible in: *State of Louisiana v. Gary Morrison,* 2003 KW 1554; *People v. Hubbard,* 738 N.W.2d 769 (Mich. 2007); and *State of Vermont v. Gregory S Forty,* 2009 VT 118.

Forensic voice comparison testimony was also ruled inadmissible in *U.S. v. Ramos,* 71 Fed. App'x. 334 (5th Cir. 2003) and in *U.S. v. Arce-Lopez,* 979 F. Supp. 2d 228 (D. Puerto Rico, 2013), although neither appellate ruling stated what approach to voice comparison the practitioner had used. In *Arce-Lopez* the court (citing *Salimonu*) found that "the jury is 'perfectly well-equipped' to listen to the witnesses testify in court, compare their voices to the voice in the audio recordings, and draw conclusions about whose voice is in the audio recordings. ... Accordingly, this is 'not an area in which expert testimony would be helpful to the jury.'" Since the court found that the expert testimony would not be of assistance to the trier of fact, it did not rule on whether it satisfied the other Rule 702 criteria. We think the court's confidence in the ability of jurors to draw conclusions about the identity of speakers from audio recordings was misplaced. Speaker identification by laypeople is highly problematic. It varies widely from listener to listener and depending on speaking, recording, and/or listening conditions. Also, people think that they and other listeners are better at speaker identification than they really are, see reviews of legal and/or research literature in: Lawrence M. Solan & Peter M. Tiersma, *Hearing voices: speaker identification in court*, 54 Hastings Law Journal 373–435 (2003) (hereinafter Solan & Tiersma 2003); Morrison 2010 note 1 *supra*; Gary Edmond, Kristy A. Martire, Mehera San Roque, *Unsound law: Issues with ('expert') voice comparison evidence*, 35 Melbourne Law Review 52–112 (2011) (hereinafter Edmond *et al*. 2011); Christopher Sherrin, *Earwitness evidence: The reliability of voice identifications*, paper 101 Osgoode Legal Studies Research Paper Series (2015) http://digitalcommons.osgoode.yorku.ca/olsrps/101 (hereinafter Sherrin 2015). Hence, this is a topic on which expert testimony may well assist the trier-of-fact, assuming the expert's methods and their implementation are valid.

[13] See Section 2.3.4 for a description of the automatic approach.

[14] *United States v. Ali Ahmed, Madhi Hashi, & Muhamed Yusuf*, No. 12-661 (E.D.N.Y.) (hereinafter *Ahmed*). The first-named author of the present paper is a forensic scientist who advised the Yusuf defense.

different approaches to forensic voice comparison, and frameworks for reasoning in assessing the strength of forensic evidence. We offer guidance on how to evaluate the scientific validity and reliability of forensic analysis systems. We also discuss the dangers of contextual bias and ways of shielding forensic practitioners from its potential effects. We then discuss the admissibility of forensic voice comparison under *Daubert* and *Frye*.[15] To provide concrete examples and argumentation regarding the underlying issues, we take a close look at the forensic voice comparison testimony from the *Daubert* hearing in *Ahmed* We examine this hearing in some depth because we believe that the same issues will recur in future cases. Finally, we describe the showing that we believe proponents of voice comparison testimony should be required to make in order to meet the standards for admissibility under Rule 702[16] and *Daubert*.

Earlier legal commentaries on and guides to forensic voice comparison evidence have focused primarily or exclusively on the auditory-spectrographic approach.[17] As far as we know, the present article is the first law review article to provide a detailed discussion of the newer automatic approach.[18]

Just as we were completing the final draft of the present article, on September 20, 2016, The President's Council of Advisors on Science and Technology (hereinafter PCAST) issued its report on *Forensic science in criminal courts: Ensuring scientific validity of feature-comparison methods*.[19] We believe that the views we express in the present article are in broad agreement with the thrust of the PCAST report.[20]

2 Primer on forensic voice comparison[21]

In a forensic voice comparison case there are at least two voice recordings, one a recording of a speaker of *known* identity and the other a recording of a speaker of *questioned* identity (there could be multiple known-speaker and multiple questioned-speaker recordings, but for simplicity the following description assumes one of each). One party in the trial contends that the speaker of questioned identity is the same as the speaker of known identity, and the other party contends that it is not the same speaker. The task of

[15] *Frye v. United States*, 293 F. 1013 (D.C.Cir.1923)

[16] Federal Rule of Evidence 702 as amended Apr. 17, 2000, eff. Dec. 1, 2000; Apr. 26, 2011, eff. Dec. 1, 2011.

[17] For example: Michele Meyer McCarthy, *Admissibility and weight of voice spectrographic analysis evidence*, 95 American Law Reports 5th 471 (2002); Solan & Tiersma 2003 note 12 *supra*; Faigman *et al*. 2015 note 2 *supra*.

[18] We restrict the present paper to forensic voice comparison performed by experts; we do not here review speaker identification by laypeople. Reviews of legal and/or research literature on the latter are included in: Solan & Tiersma 2003 note 12 *supra*, Morrison 2010 note 1 *supra*, Edmond *et al*. 2011 note 12 *supra*, Sherrin 2015 note 12 *supra*.

[19] President's Council of Advisors on Science and Technology, Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods (2016) https://www.whitehouse.gov/administration/eop/ostp/pcast/docsreports (hereinafter PCAST 2016)

[20] Although we believe that the PCAST report has several shortcomings, and we will be critical of some of those shortcomings.

[21] A more detailed introduction to the technical aspects of forensic voice comparison, still intended to be accessible to a legal audience, appears in Morrison 2010 note 1 *supra*.

the forensic scientist is to analyze the two voice recordings and to report a conclusion that will aid the trier of fact in deciding whether the recorded voices are those of the *same speaker* or of *different speakers*.

Often the speaker of known identity will be a suspect or defendant, and the speaker of questioned identity will be an offender. The recording of the speaker of known identity could be a recording of an interview at a police station, and the recording of the speaker of questioned identity could be a recording of an intercepted telephone call during which incriminating statements are made or during which the crime is actually committed (*e.g.*, a fraud is perpetrated). In such cases the prosecution will contend that the two recordings are of the same speaker and the defense will contend that they are of different speakers. There are other possible scenarios, for example, the speaker of questioned identity could be hypothesized to be a kidnap victim.

To understand and evaluate forensic voice comparison evidence it is necessary to understand a number of topics which we will introduce and discuss in the following sections. First, it is necessary to have a basic understanding of the nature of human speech and how it may vary within individuals and between individuals. Second, it is important to understand that additional variability between speech recordings may be introduced by differences in recording conditions. Third, one needs to understand the different approaches that practitioners use when analyzing speech recordings. We will explain four major methodological approaches that voice comparison practitioners have used. Fourth, it is crucial to understand the framework that the practitioner uses when evaluating and reporting the strength of the evidence. We use the term *framework* to refer to the reasoning process by which the practitioner goes from observations about the properties of the recorded speech to the conclusions that the expert states in written reports and oral testimony. Fifth, it is important to understand how the validity and reliability of practitioners' analytical procedures can be empirically tested. And sixth, it is important to understand what contextual bias is, and its potential to influence the conclusions reached by forensic practitioners.

2.1 The nature of speech

The nature of speech is quite different from that of DNA and from that of fingerprints. With some minor exceptions a person's DNA and the pattern of friction ridges on a person's finger pads do not change over time. In contrast, even if a speaker attempts to say the same thing exactly the same way twice, there will almost inevitably be measurable differences in the acoustic properties of the speech they produce. These are intrinsic differences in the speech produced, not just differences due to measurement error.

There are physical differences between speakers which cause differences in the properties of their speech. Men generally have more massive vocal folds and longer vocal tracts than women so men generally have deeper voices and lower resonance frequencies than women. *Ceteris paribus*, physical differences between different men or between different women will also result in differences in the properties of their speech.

An audio recording of a person speaking is, however, not just a representation of physical attributes in the same way as a DNA profile or a fingerprint would be. The properties of speech are also influenced by the speaker's behavior. An obvious behavioral difference is that some speakers speak one language while others speak another. A person who is bilingual can speak one language on one occasion and another language on another occasion – they have not changed the anatomy of their vocal tracts, they have changed their behavior. Along the same lines, different people speaking the same language may speak with different regional or social accents and dialects, and smaller groups of people such as family or friendship groups may share behavioral speech patterns which differ from those of other groups. An individual may even have some behavioral speech patterns which are peculiar to them. Note, however, that a person's behavior may change depending on context. For example, the way a person speaks when giving a formal presentation will probably differ from the way they speak when socializing with friends. When a speaker is calm they will speak differently from when they are excited. A speaker may whisper on one occasion and shout on another. The way a person speaks could also vary due to changes in physical conditions, for example, a person's voice may be creaky when they have not spoken for a long time or harsh if they have stressed their vocal folds by speaking loudly for a prolonged period. Medical conditions such as laryngitis or nasal congestion will also affect the properties of a person's speech. Also, behaviorally, the words a speaker says on one occasion is unlikely to be exactly the same as the words they say on another occasion, and they are even less likely to attempt to say the same words exactly the same way twice.

Although in general there are differences between different speakers (*between-speaker variability*), there are also differences in the properties of a person's speech from occasion to occasion (*within-speaker variability*). Just looking at how similar or different two voice recordings are is therefore not sufficient to be able to tell whether they were produced by the same speaker or not. One has to ask whether the properties of the speech in the questioned-speaker recording are more likely to occur if they were produced by the known speaker (any difference being due to within-speaker variability) or by some other speaker from the relevant population[22] (any similarity between the known- and questioned-speaker recording being due to chance). The same logic applies to DNA analysis, latent print analysis, and other types of forensic comparison, but the degree of intrinsic within-person variability is much larger for speech than for DNA or fingerprints.

2.2 The nature of speech recordings

As well as speech being intrinsically variable, there are also differences between recordings of speech due to variability in the conditions under which the recordings are made. A common scenario in forensic voice comparison is that the known-speaker recording is a recording of a police interview and the

[22] The concept of relevant population will be discussed below in Section 3.1.

questioned-speaker recording is a recording of a telephone call. A speaker may use different speaking styles when talking on the telephone and when being interviewed by the police, but the acoustic environment and technical aspects of the recording will also differ. The police interview may be made in a small room with hard walls. Such a room would have a substantial amount of reverberation (echoes). The room may also have an audible ventilation system. The questioned-speaker recording could be made in the street on a mobile telephone. There could be traffic noise in the background. The known-speaker recording may be made with a relatively good microphone directly in front of the speaker. The questioned-speaker recording may have been transmitted through one or more communications systems such as a landline telephone system, a mobile telephone system, or using a Voice over Internet Protocol (VoIP) system such as Skype. Such communication systems distort and remove acoustic information. Some file formats which make file sizes smaller, such as MP3, also do so by distorting and removing acoustic information. Another potential source of difference between voice recordings is the distance from the speaker to the microphone, for example, a covert recording device may be far from the speaker but an interview microphone close. Even if the sound coming out of the speaker's mouth were the same, different distances to the microphone would affect the acoustics of the recorded signal. Not all microphones have the same characteristics, and changing microphones can also affect the properties of the recorded signal. Another factor which can affect the performance of forensic voice comparison analysis is the duration of the recordings. Performance may be very poor for recordings which are only a few seconds long.

In forensic casework, there is usually a mismatch in recording conditions between known-speaker and questioned-speaker recordings. Recording-condition mismatch can make two recordings more different than they would otherwise be. Poor recording conditions can also mask intrinsic between-speaker differences. Genuine between-speaker differences could be absent, obscured, or distorted in the recorded signals. On the other hand, the cause of genuine between-speaker differences which persist in the recorded signal could be incorrectly attributed to differences due to recording conditions.

All of these variables must be taken into account when performing a forensic comparison of known- and questioned-speaker recordings.

2.3 Approaches to forensic voice comparison

Historically, and still in current practice, there are four basic *approaches* to forensic voice comparison, which we denominate *auditory*, *spectrographic*, *acoustic-phonetic*, and *automatic*. We will further divide acoustic-phonetic into *acoustic-phonetic non-statistical* and *acoustic-phonetic statistical*. Practitioners frequently use a mixture of different approaches (*e.g.*, auditory-spectrographic, auditory-acoustic-phonetic), but for clarity we will describe each one separately.

2.3.1 Auditory approach[23]

In an auditory approach (aka aural approach) the practitioner listens to the known-speaker and questioned-speaker recordings. They listen in search of similarities which they would expect to hear if the two recordings consisted of speech from the same speaker, but which they would not expect to be likely to hear if the recordings consisted of speech from different speakers. They also listen in search of differences which they would expect to hear if the two recordings consisted of speech from different speakers, but which they would not expect to be likely to hear if the recordings consisted of speech from the same speaker. They may listen to the pronunciation of particular vowel sounds or of particular consonant sounds, the pronunciation of particular words or phrases, and other more global properties such as intonation patterns and the auditory effects of physical properties and configurations of vocal folds. Practitioners will typically have training in auditory phonetics, including training in transcribing the speech sounds they hear using a phonetic alphabet. Thus the practitioner will have a means of documenting what they hear and highlighting the similarities and differences that they consider to be pertinent. Practitioners may have tools which allow them to listen to short sections of speech from each recording, one immediately after the other. They may also listen to sections of speech from other speakers who act as foils, *i.e.*, speakers who sound broadly similar to the questioned speaker. The practitioner may be presented with multiple recordings of each of a number of speakers, without being told which are of the known speaker, the questioned speaker, and the foils, and be asked to group the recordings by speaker.

The conclusion emerging from an auditory approach is the practitioner's subjective[24] judgment based on listening to the speech recordings.

2.3.2 Spectrographic approach[25]

[23] For other introductions to the auditory approach, see: Francis Nolan, *Speaker recognition and forensic phonetics*, *in* The Handbook of Phonetic Sciences 744–767 (William J. Hardcastle & John Laver 1997) (hereinafter Nolan 1997); Philip J. Rose, Forensic Speaker Identification (2002) (hereinafter Rose 2002); Philip J. Rose, *Technical forensic speaker recognition*, 20 Computer Speech and Language 159–191 (2006) http://dx.doi.org/10.1016/j.csl.2005.07.003 (hereinafter Rose 2006); Michael Jessen, *Forensic phonetics*, 2 Language and Linguistics Compass 671–711 (2008) http://dx.doi.org/10.1111/j.1749-818x.2008.00066.x (hereinafter Jessen 2008); Morrison 2010 note 1 *supra*; Hollien 2016 note 3 *supra*; Harry Hollien, Grace Didla, James D. Harnsberger, Keith A. Hollien, *The case for aural perceptual speaker identification*, 269 Forensic Science International 5–20 (2016) http://dx.doi.org/10.1016/j.forsciint.2016.08.007

[24] Some consider the term *subjective* to be pejorative. In scientific writing, this is generally not the case. Throughout the present paper we use the term *subjective* in accordance with the following definition from Merriam-Webster (http://www.merriam-webster.com/dictionary/subjective): "3a: characteristic of or belonging to reality as perceived rather than as independent of mind ... b: relating to or being experience or knowledge as conditioned by personal mental characteristics or states".

[25] For other introductions to the spectrographic approach, see: Oscar Tosi, Voice Identification: Theory and Legal Applications (1979); NRC 1979 note 6 *supra*; Gruber & Poza 1995 note 8 *supra*; Harry Hollien, Forensic Voice Identification (2002) (hereinafter Hollien 2002); Rose 2002 note 23 *supra*; Didier Meuwly, *Le mythe de l'empreinte vocale I*, 56 Revue Internationale de Criminologie et Police Technique 219–236 (2003); Didier Meuwly, *Le mythe de l'empreinte vocale II*, 56

In a spectrographic approach the practitioner takes parts of the audio recordings (typically words or phrases) and converts them into pictures. These pictures are called spectrograms. In the context of forensic voice comparison, spectrograms have also been called *voiceprints*[26] and *voicegrams*. An example of a spectrogram is shown in Fig. 1.[27] The practitioner looks at spectrograms derived from the known-speaker recording and spectrograms derived from the questioned-speaker recording, and may also look at spectrograms derived from recordings of foil speakers. Usually the practitioner will look at multiple words or phrases that occur in both the known-speaker and questioned-speaker recordings. They may look at particular details in the pictures in search of similarities which they would expect to see if the two recordings were of the same speaker but not expect to be likely if they were of different speakers, and also in search of differences they would expect to see if the two recordings were of different speakers but not expect to be likely if they were of the same speaker. In contrast to other approaches, there has been a tradition for practitioners of the spectrographic approach to make new recordings of the known speaker in which the known speaker is required to say the same words as on the questioned-speaker recording and in the same manner as they were said on the questioned-speaker recording. This practice has been criticized by others, but has been enshrined as a requirement in published standards.[28]

The conclusion emerging from a spectrographic approach is the practitioner's subjective judgment based on looking at pictures of parts of the speech recordings.

Revue Internationale de Criminologie et Police Technique 361–374 (2003); Morrison 2010 note 1 *supra*

[26] The term *voiceprint* in a forensic context dates back to at least the 1960s; Lawrence G. Kersta, *Voiceprint identification*, 196 Nature 1253 (1962) http://dx.doi.org/10.1038/1961253a0. *Voiceprinting* referred to a particular approach, and *voiceprint* was even a registered trademark. The term quickly fell into disrepute among forensic practitioners, even among practitioners of the spectrographic approach. One objection was that the term implied a false analogy with *fingerprint*. Unfortunately, the term is still widely used by the general public and in legal circles, where it is often incorrectly used to refer to forensic voice comparison in general. Lawyers and judges should be aware that many forensic voice comparison practitioners will consider it an insult if they are called a *voiceprint expert*. We recommend that the term not be used (except in relation to its proper historical referent).

[27] Spectrograms were initially produced using specialized hardware which was first developed in the 1940s. Measurements of acoustic properties of speech could be made from the spectrogram, *i.e.*, by lining up a ruler with graphical features and reading off values on the time or frequency axis. Since at least the early 1990s, it has been possible to produce spectrograms using ordinary computers running signal processing software. Such software calculates numbers which describe the acoustic properties of the speech on the recording, then converts those numbers into pictures. Continued reliance on spectrograms as a basis for subjective judgments could be criticized as anachronistic given that measurements of acoustic properties can be directly extracted using software and those numbers can be immediately entered into statistical models.

[28] American Board of Recorded Evidence, Voice comparison standards (1999) http://www.tapeexpert.com/pdf/abrevoiceid.pdf (last visited Oct 22, 2016) (hereinafter ABRE 1999); International Association for Identification 1991 voice comparison standard, as quoted in Gruber & Poza 1995 note 8 *supra* at §57–60 (hereinafter IAI 1991); International Association of Voice Identification recommended procedures, as reproduced in Appendix A of NRC 1979 note 6 *supra* (hereinafter IAVI 1979)

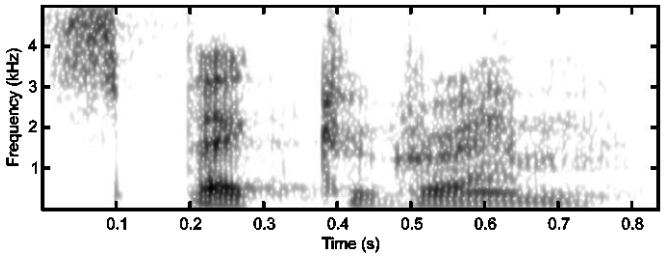

Figure 1. Example of a spectrogram.[29]

2.3.3 Acoustic-phonetic approach[30]

In an acoustic-phonetic approach the practitioner usually uses computer software to make quantitative measurements of acoustic properties of parts of the voice recordings. Measurements may be made on particular speech sounds that occur in both the known-speaker and questioned-speaker recordings. The types of measurements made are generally the same as the types of measurements which are made in acoustic phonetics, an area of research which studies the transmission of human speech between the speaker's vocal tract and the listener's ear. An example of properties commonly measured are *formants*. Formants are the *resonances* of the vocal tract. In the same way that longer tubes of wind instruments have lower resonances than shorter tubes (*e.g.*, bassoon versus oboe or tuba versus trumpet), longer human vocal tracts have lower resonances than shorter human vocal tracts. The length of the vocal tract can vary from person to person, but when speaking a person changes the length and shape of their vocal tract to produce a range of different resonance frequencies. The differences between vowel sounds such as "ee", "oo", and "ah" are the result of different resonances resulting from the speaker moving their tongue, jaw, lips, etc. to make different vocal tract shapes. Another commonly made measurement is *fundamental frequency*, which is the acoustic correlate of what listeners perceive as the pitch of someone's voice, *e.g.*, a deep voice or a high-pitched voice. Whereas formants are related to the length and shape of the vocal tract, fundamental frequency is related to the size of the speaker's vocal folds and the configuration in which they hold and put tension on their vocal folds. To return to the analogy of a wind instrument, the vocal folds are like the vibrating reed/reeds of a woodwind instrument or the vibrating lips of a musician playing a brass instrument. In the same way that the musician can alter the frequency of vibration of the reed/reeds or their lips, a speaker can alter the frequency of vibration of their vocal folds. The same vowel sound can be sung using different musical notes, the different musical

[29] This figure was first published in Geoffrey Stewart Morrison, *Forensic voice comparison and the paradigm shift*, 49 Science & Justice 298–308 (2009) http://dx.doi.org/10.1016/j.scijus.2009.09.002 (hereinafter Morrison 2009).

[30] For other introductions to the acoustic-phonetic approach, see: Nolan 1997 note 23 *supra*; Hollien 2002 note 25 *supra*; Rose 2002 note 23 *supra*; Rose 2006 note 23 *supra*; Jessen 2008 note 23 *supra*; Cuiling Zhang, 法庭语音技术研究 [Forensic Speech Technology Research] (2009); Morrison 2010 note 1 *supra*; Michael Jessen, Phonetische und Linguistische Prinzipien des Forensischen Stimmenvergleichs (2012); Philip J. Rose, *Where the science ends and the law begins- likelihood ratio-based forensic voice comparison in a $150 million telephone fraud*, 20 International Journal of Speech, Language and the Law 227–324 (2013) http://dx.doi.org/10.1558/ijsll.v20i2.277

notes are due to the singer changing the frequency of vibration of their vocal folds. Many types of acoustic measurements are the quantitative acoustic parallels of the subjective auditory properties that practitioners of the auditory approach listen for, and many are quantitative parallels of properties which are represented graphically in spectrograms.

A practitioner will usually manually search for all occurrences of a particular speech sound, or word, or phrase which occurs in both the known-speaker and questioned-speaker recordings. They will then make measurements of the acoustic properties of those units. The numbers resulting from the measurements can then be compared. The practitioner may also make the same types of measurements on the same units in voice recordings from other speakers. The latter could be foil speakers, or could be intended to be a sample of speakers representative of the relevant population in the case. The practitioner will usually make measurements on several different speech sounds, words, and/or phrases, not just one.

There are both statistical and non-statistical versions of the acoustic-phonetic approach. In the *non-statistical* version the conclusion is the practitioner's subjective judgment based on considering the raw numbers from the measurements or based on looking at graphical plots of the numbers. In the *statistical* version the conclusion is based on a statistical model applied to the numbers. Statistical models can take such numbers as input and calculate numeric expressions of strength of evidence in a more objective manner.[31] Some practitioners directly report the numeric output of the statistical model as their conclusion, other practitioners report a subjective judgment based on consideration of the output of the statistical model.

2.3.4 Automatic approach[32]

[31] An example of a simple statistical model is a normal distribution. This has two *parameters*: a mean, and a standard deviation. Data are used to calculate estimates of these parameter values (these estimates are called *statistics*). The process of using data to estimate parameter values is called model *training*. If we asked you "What is the probability that an adult American male would be between 5 feet 6 inches tall and 6 feet tall?" you could make a subjective estimate. A statistical model would give you a more objective estimate. Imagine that we obtain data which consist of measurements of the heights of 5,232 adult male Americans, we assume that our sample is representative of the population and that heights in the population are normally distributed, and we calculate that the mean height is 69.2 inches and the standard deviation is 6.0 inches, values from Cheryl D. Fryar, Qiuping Gu, Cynthia L. Ogden, & Katherine M. Flegal, Anthropometric reference data for children and adults: United States, 2011–2014, 3(39) Vital Health Statistics (2016) http://www.cdc.gov/nchs/data/series/sr_03/sr03_039.pdf (last accessed Oct 22, 2016); we can then use a normal distribution with this mean and standard deviation to calculate an estimate of the probability that an adult American male is between 66 and 72 inches tall. The answer is 38%. Note that use of a statistical model is not entirely objective since choices have to be made about what particular statistical model to use and what data to use to train the model. Poor choices may lead to poor results, but once these choices have been made, the remainder of the process is objective. In fact a normal distribution is a poor choice for modelling human height, and a more complex model taking account of population substructure would give better results. In §3.1we will discuss how statistical models can be used to evaluate strength of forensic evidence.

[32] For other introductions to the automatic approach, see: Didier Meuwly, *Reconnaissance de locuteurs en sciences forensiques: L'apport d'une approche automatique* (2001) http://www.unil.ch/webdav/site/esc/shared/These.Meuwly.pdf (last accessed Oct 22, 2016); Anil Alexander, *Forensic automatic speaker recognition using Bayesian interpretation and*

In an automatic approach the practitioner uses computer software to make measurements of the acoustic properties of the known-speaker and questioned-speaker recordings, and of voice recordings from other speakers who are intended to represent the relevant population for the case. Generally the acoustic measurements are made on the whole of the speakers' speech in the recordings, and there is no focus on individual speech sounds, words, or phrases. The types of measurements made are usually the same as those used in speech processing (a branch of signal processing, in turn a branch of electrical engineering). These types of measurements are also applied to other tasks such as automatic speech recognition. An example of a common type of measurement is *mel frequency cepstral coefficients* (MFCCs). MFCCs are a set of numbers, *e.g.*, 14 numbers, which describe the frequency components (the *spectrum*) of the speech during a short interval of time, *e.g.*, 20 milliseconds. MFCC measurements are usually made once every 10 milliseconds, *i.e.*, 100 times per second (with a 50% overlap of adjacent 20 millisecond long intervals). A set of 14 MFCCs provides a more detailed measurement of the speech spectrum than do traditional acoustic-phonetic measurements, such as fundamental frequency plus two or three formants, but the value of an individual cepstral coefficient is not usually directly interpretable in terms of acoustic-phonetic theory.[33]

In an automatic system, the numbers from the measurements are always used as input to statistical models. The practitioner may be involved in selecting what they consider to be appropriate statistical models, appropriate types of measurements, appropriate data for training the statistical models, and in selecting which portions of the audio recordings correspond to the speaker of interest, but the measurements and statistical models themselves run automatically without additional human intervention. In automatic speaker recognition, a number of statistical techniques have been developed to deal with differences in recording conditions between known-speaker and questioned-speaker recordings. Many of these techniques can also be applied in automatic approaches to forensic voice comparison.

The conclusion emerging from an automatic approach will be based on the output of the statistical model. Some practitioners directly report the numeric output of the statistical model as their conclusion, other practitioners report a subjective judgment based on consideration of the output of the statistical model.

---

statistical compensation for mismatched conditions (2005); Daniel Ramos Castro, *Forensic evaluation of the evidence using automatic speaker recognition systems* (2007) http://atvs.ii.uam.es/files/2007_11_28_thesis_daniel_ramos_searchable_v1.pdf; Tomi Kinnunen & Haizhou Li, *An overview of text-independent speaker recognition: From features to supervectors*, 52 Speech Communication 12–40 (2010) http://dx.doi.org/10.1016/j.specom.2009.08.009; Morrison 2010 note 1 *supra*; Timo Becker, Automatischer Forensischer Stimmenvergleich (2012); John H.L. Hansen & Taufiq Hasan, *Speaker recognition by machines and humans: A tutorial review*, IEEE Signal Processing Magazine 74–99 (2015, November) http://dx.doi.org/10.1109/MSP.2015.2462851; Ewald Enzinger, *Implementation of forensic voice comparison within the new paradigm for the evaluation of forensic evidence* (2016) http://handle.unsw.edu.au/1959.4/55772 (hereinafter Enzinger 2016)

[33] The boundary between acoustic-phonetic-statistical and automatic approaches is fuzzy. A recent trend in automatic speaker recognition is to incorporate acoustic-phonetic information (or acoustic-phonetic like information) by, for example, using automatic speech recognition to divide the speech recording into different phonetic units and then make use of those units in subsequent analysis.

3 Frameworks for the evaluation and the reporting of the strength of forensic evidence

We next consider frameworks that practitioners use to reason and draw conclusions from their analyses of voice recordings. Practitioners who use the same approach may apply different frameworks to evaluate and present the strength of the evidence. To understand a practitioner's conclusions, one must understand both the approach used for technical analysis of the voice recordings, and the framework applied to reason and draw conclusions from that analysis. A practitioner who uses sophisticated methods of analysis but draws illogical or otherwise unjustifiable conclusions from that analysis will not produce trustworthy evidence.

Below we discuss the likelihood-ratio framework, similarity-only framework, posterior-probability framework, two-stage framework, and the UK framework.

The reader interested in more comprehensive introductions to the likelihood ratio framework within the context of forensic voice comparison may wish to consult Rose (2002)[34] and Morrison (2010).[35] General introductions to the likelihood ratio framework which should be accessible to a legal audience include Kaye *et al*. (2011) chapter 14,[36] and Robertson & Vignaux (1995)[37] (or the recently released second edition, Robertson *et al*. 2016).[38]

For additional discussion of similarity-only and posterior-probability frameworks and other non-likelihood-ratio means of expressing strength of evidence, the following are highly recommended: Jackson (2009),[39] Kaye (2015).[40]

---

[34] note 23 *supra*

[35] note 1 *supra*

[36] David H. Kaye, David A. Bernstein, & Jennifer L. Mnookin, The New Wigmore: A Treatise on Evidence: Expert Evidence (2nd ed. 2011)

[37] Bernard Robertson & G.A. Vignaux, Interpreting Evidence: Evaluating Forensic Science in the Courtroom (1995) (hereinafter Robertson & Vignaux 1995)

[38] Bernard Robertson, G.A. Vignaux, Charles E.H. Berger, Interpreting Evidence: Evaluating Forensic Science in the Courtroom (2nd ed. 2016) (hereinafter Robertson *et al*. 2016). Some readers may find chapters 7 and 8 of Robertson *et al*. 2016 somewhat technical, and may wish to skip these, at least on a first reading. Other recommended, but more technical introductions include: David J. Balding & Christopher D. Steele, Weight-of-evidence for Forensic DNA Profiles (2nd ed. 2015) chapters 1–3 and 11; Colin G.G. Aitken, Paul Roberts, Graham Jackson G, Fundamentals of Probability and Statistical Evidence in Criminal Proceedings: Guidance for Judges, Lawyers, Forensic Scientists and Expert Witnesses (2010) http://bit.ly/1WnoXRx

[39] Graham Jackson, *Understanding forensic science opinions*, *in* Handbook of Forensic Science 419–445 (Jim Fraser & Robert Williams 2009)

[40] David H. Kaye, Presenting forensic identification findings: The current situation, in Communicating the Results of Forensic Science Examinations, Penn State Law Research Paper No. 23-2015 (Cedric Neumann, Anjali Randive, & David H. Kaye 2015) http://ssrn.com/abstract=2690891. Also recommended, but more technical: John S. Buckleton, *A framework for interpreting evidence*, *in* Forensic DNA Evidence Interpretation 27–63 (John S. Buckleton, Christopher M. Triggs, Simon J.

3.1 The likelihood-ratio framework

In the opinion of many leading scholars in the field of forensic inference and statistics,[41] the logically correct framework for the evaluation of forensic evidence is the *likelihood ratio framework* (which has also been called the *Bayesian framework*[42] and the *logical framework*). In the context of forensic voice comparison, this framework requires the practitioner to consider two questions:

> 1. What is the probability of obtaining the observed properties of the voice on the questioned-speaker recording if it were produced by the known speaker?
>
> 2. What is the probability of obtaining the observed properties of the voice on the questioned-speaker recording if it were produced not by the known speaker, but by some other speaker from the relevant population?[43]

The answer to the first question quantifies the *similarity* of the recording of the voice of the questioned speaker with respect to the known speaker, and the answer to the second question quantifies the *typicality* of the recording of the voice of the questioned speaker with respect to the relevant population.

The need to consider both similarity and typicality is more intuitively understood if we use an example based on a simpler (and simplified) evidence type. Imagine that all the eyewitnesses to a crime agree that the offender had blond hair. Further imagine that the eyewitnesses are not mistaken, that blond is clearly different from every other hair color, and that no one ever dyes their hair or wears a wig, *etc*. – we work in a simplified world to make the example easier. Also imagine that a suspect has been arrested, and the suspect also has blond hair. What is the probability that the offender would have blond hair if he were the suspect? Please think about this for a moment before reading the next sentence. Given all the simplifications, the probability of the offended having blond hair if he were the suspect should be 100%.[44] Now, another question: What is the probability that the offender would have blond hair if he were not the suspect but instead someone selected at random from the population? Please think about this for a

Walsh 2005) / Tasha Hicks, John S. Buckleton, Jo-Anne Bright, Duncan Taylor (2015). *A framework for interpreting evidence*, *in* Forensic DNA Evidence Interpretation 37–86 (John S. Buckleton, Jo-Anne Bright, Duncan Taylor 2nd ed. 2015)

[41] for example: Colin G.G Aitken, Charles E.H. Berger, John S. Buckleton, Christophe Champod, James M. Curran, A. Philip Dawid, Ian W. Evett, Peter Gill. Joaquín González-Rodríguez, Graham Jackson, Ate Kloosterman, Tina Lovelock, David Lucy, Pierre Margot, Louise McKenna, Didier Meuwly, Cedric Neumann, Niamh Nic Daéid, Anders Nordgaard, Roberto Puch-Solis, Birgitta Rasmusson, Michael Redmayne, Paul Roberts, Bernard Robertson, Claude Roux, Marjan J. Sjerps, Franco Taroni, Tjark Tjin-A-Tsoi, G.A. Vignaux, Shiela M. Willis, Grzegorz Zadora, *Expressing evaluative opinions: A position statement*, 51 Science & Justice 1–2 (2011) http://dx.doi.org/10.1016/j.scijus.2011.01.002

[42] Although it has been called the *Bayesian framework*, it is not necessarily the case that the forensic practitioner applies Bayes theorem or uses a Bayesian concept of probability.

[43] The concept of relevant population will be discussed below.

[44] Statisticians use numbers between 0 and 1 for probabilities, whereas laypeople usually use percentages. Divide or multiply by 100 to convert from one system to the other.

moment before reading the next sentence. Maybe you have decided that you actually can't answer the question as posed because the question didn't specify which population was relevant. Which is the relevant population in this case? Let's assume that the offender must have been someone from the geographical area in which the crime was committed and that people in that geographical area therefore form the relevant population. What if the crime were committed in Stockholm? What is the probability that someone would have blond hair if they were selected at random from the population of Stockholm? We don't know the exact number for that probability, but we imagine that it is pretty high, maybe as high as 80%. It should be intuitively obvious that the fact that both the suspect and offender have blond hair does not constitute strong evidence in favor of the hypothesis that they are the same person if the alternative hypothesis is that the offender is someone selected at random from the population of Stockholm. If we were to use the values of 100% and 80% we previously mentioned, then the probability that the offender would have blond hair if they were the suspect is 100/80 = 1.25 times higher than if the offender were someone selected at random from the population of Stockholm.

The number we just calculated has a name, it is called a *likelihood ratio*. In the present context, a forensic likelihood ratio is the probability of the evidence if the same-origin hypothesis were true divided by the probability of the evidence if the different-origin hypothesis were true.[45] A likelihood ratio is a quantitative statement of the strength of the evidence. In this case the *evidence* is that the offender has blond hair, the *same-origin hypothesis* is that the offender is the suspect, and the *different-origin hypothesis* is that the offender is not the suspect but someone else selected at random from the population of Stockholm. If the likelihood ratio has a value greater than one, then the evidence is more likely under the same-origin hypothesis than under the different-origin hypothesis, and the larger the likelihood ratio value the greater the relative support for the same-origin hypothesis over the different-origin hypothesis. *Mutatis mutandis*, if the likelihood ratio has a value less than one, then the evidence is more likely under the different-origin hypothesis than under the same-origin hypothesis, and the smaller the likelihood ratio value the greater the relative support for the different-origin hypothesis over the same-origin hypothesis. If the likelihood ratio value is close to one, then the evidence is about equally likely under each hypothesis.

What if instead of Stockholm the crime had been committed in Beijing and it is the population of Beijing that is the relevant population? We will leave it up to you to think about that. You can even take a guess at the probability that someone selected at random from the population of Beijing will have blond hair and calculate a likelihood ratio using the number you guessed.

Rather than guessing the probability that someone selected at random from the population would have blond hair, we can estimate this probability based on relevant data. If we go back to Stockholm, there are

[45] In scientific literature on forensic inference and statistics, the generic terms for the competing hypotheses are often the *prosecution hypothesis* and the *defense hypothesis*. It is not necessarily the case, however, that they are overtly advanced by the prosecution and defense respectively. They must be two alternative hypotheses which are of interest to the trier of fact in that together they pose a question the answer to which will assist the trier of fact to determine a fact at issue in the trial. The two hypotheses must be mutually exclusive, and, within the circumstances of the case, exhaustive.

more than 1 million people in the greater Stockholm area and it isn't practical to look at the hair color of all of them (the population of the greater Beijing area is 25 million, so that would be even less practical). It is practical, however, to look at the hair color of a few hundred people and base our estimate on that. The data from a few hundred people are a *sample* of the population. We want the sample to be representative of the population as a whole, so we need to take some care in selecting who to include in the sample. Maybe there is a neighborhood in Stockholm where lots of Chinese immigrants live and in that neighborhood blond hair is relatively uncommon compared to the rest of the city. If we want our sample to be representative of the city as whole, we should not take our whole sample from that neighborhood. Maybe we decide that people walking in the city center will give us a representative sample of the population as a whole. For every person who passes by we note their hair color, is it blond or some other color? Once we have noted the hair color of a few hundred people we can calculate the percentage (or the proportion) who have blond hair. We can then use that percentage as our estimate of the probability that someone selected at random from the population of Stockholm will have blond hair. Note that, in addition to sampling in the right parts of the city so that we expect the sample to be representative of the population of the city as a whole, the size of our sample also matters to some extent. What if we only sampled two people? If we repeatedly sampled groups of two people we would probably find that our estimate of the probability of blond hair in the population varied wildly from sample group to sample group, sometimes 0%, sometimes 50%, sometimes 100%. What if our sample size were 10? The situation would be better, but we could still have considerable variability in our estimate. We could make our sample size 100 and get a more stable estimate, or even make the sample size 1000. At some point, however, we should find that the estimate is quite stable, *i.e.*, adding substantially more individuals to the sample would not result in a substantial difference in the estimate, and that the costs associated with collecting a larger sample would not be warranted. At some point we decide that our sample is large enough to get a sufficiently accurate and precise[46] estimate of the probability of blond hair in the population, and we do not collect a larger sample. In some (or many) instances the cost of collecting data may dictate how much we can afford to collect and then we will have to *post hoc* assess accuracy and precision given this amount of data.

Speech data will be less intuitive for most readers, but the same principles apply. The forensic practitioner must estimate both the degree of similarity of the properties of the voice on the questioned-speaker recording with respect to the known speaker, and the degree of typicality of the properties of the voice on the questioned-speaker recording with respect to the relevant population. One must estimate the probability of the properties of the speech on the questioned-speaker recording had it been produced by the suspect, and one must also estimate the probability of the properties of the speech on the questioned-speaker recording had it been produced by someone else from the relevant population. Because of within-speaker variability, even if it does come from the known speaker the probability of obtaining the measured properties of the voice on the questioned-speaker recording will always be much less than

[46] Accuracy and precision (aka validity and reliability) are discussed below in Section 4.

100%.

Deciding what constitutes the relevant population may not be trivial. The defense position is usually that the questioned speaker is not the known speaker, so a different-speaker hypothesis including a specific relevant population could be explicitly stated by the defense, but the forensic scientist often has to work without being provided with an explicit hypothesis from the defense. In this case the forensic scientist must adopt a hypothesis that they expect will be deemed appropriate by the trier of fact. The relevant population is the population of people who could plausibly have produced the voice on the questioned-speaker recording if it were not produced by the known speaker. Information in the questioned-speaker recording itself will usually (but not always) indicate whether the speaker is a male or female, what language is being spoken, and broadly what accent is being spoken. For example, if it is clear that the questioned speaker is an adult male who speaks English with a Boston accent, and this is not likely to be disputed by either party, then it would be reasonable to adopt as the relevant population men who speak English with a Boston accent.[47]

It is vital that the forensic practitioner clearly communicate to the trier-of-fact the choices they make regarding the particular hypotheses and the particular relevant population that they, the forensic practitioner, adopt. In order to know whether the forensic practitioner's testimony addresses a relevant question, and in order to understand the forensic practitioner's answer to that question, the trier of fact must know what hypotheses the forensic scientist is evaluating. In order to understand the meaning of the likelihood ratio value provided by the forensic practitioner, the trier-of-fact must know what the forensic practitioner adopted as the relevant population.

One cannot understand the answer if one does not understand the question. Imagine a forensic scientist working on a case in Beijing who uses hair color data from a sample of the population of Stockholm, but just presents their conclusion without explaining the different-origin hypothesis and relevant population which they had adopted, or similarly a forensic scientist working on a case in which the questioned-speaker recording is clearly of a female speaking Chinese but the forensic scientist uses a database of male speakers of Swedish to estimate the probabilities of speech properties in the population. The questions being asked and the relevant populations implied by the sample data would be nonsensical in the context of these cases, and the answers would be meaningless. Consider a less obvious example: Both the questioned speaker and the known speaker are speaking English and pronounce the word "car"

[47] Note that information about the known speaker cannot be used (see Robertson & Vignaux 1995 note 37 *supra* at pp. 43–44 / Robertson *et al*. 2016 note 38 *supra* at pp. 39–40). For example, we may know for a fact that the known speaker is 30 years old, but this information cannot be used to refine the relevant population. We don't know who the questioned speaker is, that is the question which the court proceedings will answer, so we do not know exactly how old they are. The defense contends that the questioned speaker is not the known speaker, so for the relevant population which is part of the different-origin hypothesis we are not justified in assuming that because the known speaker is 30 years old the questioned speaker is also 30 years old. Just by listening to a person's voice we cannot tell exactly how old they are. On the basis of listening (or acoustic measurements and statistical models) we might believe that the questioned speaker is not a child or a teenager and not elderly but it would be unlikely that we would be correct in estimating their exact age in years.

without pronouncing an "r" sound. If the relevant population is that of the United States in general, then the degree of typicality (the value of the denominator of the likelihood ratio) will be very different than if the relevant population is restricted to that of Boston.[48] The forensic scientist must clearly state the hypotheses they are addressing so that the judge at an admissibility hearing and/or the trier of fact at trial can decide if the question the forensic scientist asked is appropriate and so that the trier of fact can understand the forensic scientist's answer to the question.

3.2 Similarity-only framework

Some forensic practitioners *only* consider *similarities* between the voices on the known-speaker and questioned-speaker recordings. Similarity may be expressed verbally using terms such as "match" or "consistent with", *e.g.*, "the fundamental frequency of the voice on the questioned-speaker recording matches that of the voice on the known-speaker recording", or "the spectral properties of the voice on the recording of the bomb threat are consistent with it having been made by the defendant". Alternatively, the practitioner may simply point out properties that are similar in the known- and questioned-speaker recordings, *e.g.*, "on both recordings the speaker pronounces the word 'ask' like 'axe'".

As already explained in Section 3.1, degree of similarity alone is not meaningful; degree of typicality with respect to a relevant population also needs to be considered. Saying that the suspect and the offender both have blond hair, could be highly misleading without also providing information about the probability of finding blond hair in the relevant population. Likewise, only considering similarities between voice recordings could be highly misleading.

3.3 Posterior-probability framework

Some practitioners present *posterior probabilities*, *e.g.*, there is a 95% probability that the voice on the questioned-speaker recording was produced by the known speaker. Expressions of posterior probabilities need not be numerically exact. Expressions such as "identification", "probable identification", "possible identification", "inconclusive", "possible elimination", "probable elimination", and "elimination"[49] are verbal expressions of posterior probabilities.

---

[48] If it is obvious to the trier of fact that the questioned speaker has a Boston accent, then they will already have taken this into account, and (assuming the known speaker also has a Boston accent) they will be interested in the strength of evidence associated with whether the questioned speaker is the known speaker versus someone else who speaks with a Boston accent. See discussion of this issue in Geoffrey Stewart Morrison, Ewald Enzinger, & Cuiling Zhang, *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 (2016) http://dx.doi.org/10.1016/j.scijus.2016.07.002.

[49] ABRE 1999 note 28 *supra*, IAI 1991 note 28 *supra*

Logically, posterior probability cannot be derived solely via comparison of the properties of the known- and questioned-speaker recordings (irrespective of whether comparison is also made with a sample form a relevant population). Logically, in order to arrive at a *posterior probability* one has to combine two things: a *prior probability* and a *likelihood ratio*. That this is logically true can be formally proven and is not disputed among logicians or statisticians. This piece of logic is called *Bayes' Theorem*, and formal descriptions of this logic date back to the mid 1700s. Popular literature on this topic includes McGrayne (2011)[50] and Lindley (2006).[51]

In Appendix A, we explain how a prior probability and a likelihood ratio are combined to arrive at a posterior probability.

In the context of a court case in which forensic voice comparison testimony is presented, the prior probability is properly in the mind of the trier of fact, it is the belief that the trier of fact has as to the probability that the questioned-speaker is the known-speaker before the forensic voice comparison testimony is presented. The trier of fact's prior probability will depend on other information and evidence which have already been presented to them during the trial. If the trier of fact were to use the normative logic of Bayes' Theorem, they would combine the likelihood ratio with their prior probability to arrive at a posterior probability – the belief that the trier of fact has as to the probability that the questioned-speaker is the known-speaker after the forensic voice comparison testimony has been presented. The posterior probability would be either higher or lower than the prior probability depending on whether the likelihood ratio was greater than or less than 1, and the extent to which the posterior probability was higher or lower than the prior probability would depend on how high or low the likelihood ratio was.

Even if they are not aware of it, a forensic practitioner presents a posterior probability must have at least implicitly used a prior probability. Unless the trier of fact tells the forensic practitioner what prior probability to use, however, the forensic scientist cannot calculate the appropriate posterior probability. The posterior probability the practitioner presents will instead reflect the practitioner's own views or assumptions about the prior probability, which may differ from those of the trier of fact. The trier of fact may be unaware that the putatively scientific conclusion offered by the forensic practitioner depended partly on the practitioner's views or assumptions about the prior probability. Even if they are aware, it may not be clear to the trier of fact the extent to which the practitioner's conclusions were influenced by matters other than consideration of the voice evidence.

[50] Sharon B. McGrayne, The Theory that Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy (2011)

[51] Denis V. Lindley, Understanding Uncertainty (2006)

3.4 Identification / inconclusive / exclusion

Many forensic practitioners report definitive posterior-probability conclusions. They report conclusions such as "identification, "inconclusive", or "exclusion".[52] "Identification" or "same speaker" and "exclusion" or "different speaker" are extreme cases of verbal probabilities corresponding to 100% and 0% respectively. If the probability is 100% then no other evidence such as an alibi could outweigh it. If the probability is 0% then no other evidence such as an eyewitness statement could outweigh it. If no other evidence could outweigh the evidence presented by the forensic practitioner, then logically the forensic practitioner would have made a definitive decision on the ultimate issue of identity. That is a decision which should be made by the trier of fact after weighing all the relevant evidence presented to them, it should not be made by a forensic practitioner. The forensic practitioner should only analyze and express the strength associated with the one piece of evidence that they were asked to analyze.

The PCAST report opines that:

> the expert should not make claims or implications that go beyond the empirical evidence and the applications of valid statistical principles to that evidence.[53]

And recommends that:

> courts should never permit scientifically indefensible claims such as: "zero," "vanishingly small," "essentially zero," "negligible," "minimal," or "microscopic" error rates; "100 percent certainty" or proof "to a reasonable degree of scientific certainty;" identification "to the exclusion of all other sources;" or a chance of error so remote as to be a "practical impossibility."[54]

3.5 A reasonable degree of scientific certainty

Some forensic practitioners may state conclusions such as the following: "To *a reasonable degree of scientific certainty* the voice on the questioned-speaker recording was produced by the known speaker." The expression *a reasonable degree of scientific certainty* is not used in science in general, and it has no clearly defined meaning. It appears to have been invented to increase the likelihood that medical and forensic practitioners' conclusions would be accepted by a court without further enquiry. The phrase *a*

[52] See Geoffrey Stewart Morrison, Farhan Hyder Sahito, Gaëlle Jardine, Djordje Djokic, Sophie Clavet, Sabine Berghs, & Caroline Goemans Dorny, *INTERPOL survey of the use of speaker identification by law enforcement agencies*, 263 Forensic Science International 92–100 (2016) http://dx.doi.org/10.1016/j.forsciint.2016.03.044 (hereinafter INTERPOL survey)

[53] PCAST 2016 note 19 *supra* at p. 6

[54] *Ibid*. at p. 19

reasonable degree of medical certainty has been called the "magic words".[55] The NCFS and the PCAST report have recommended that expressions of this sort not be used.[56]

3.6 Two-stage framework[57]

The framework for evaluation of evidence recommended in the PCAST report[58] is essentially a likelihood-ratio framework. It is not, however, appropriate for data resulting from acoustic measurements made on voice recordings.

The PCAST report recommends a procedure in which if a forensic practitioner declares a "match", they also report the results of an empirical assessment of the probability of declaring a "match" if the questioned-source specimen came from the known-source (this is the numerator for a likelihood ratio) and the probability of declaring a "match" if the questioned-source specimen came from some other source (this is the denominator for a likelihood ratio).

> The forensic examiner should report the overall false positive rate [denominator of the likelihood ratio] and sensitivity [numerator of the likelihood ratio] for the method established in the [empirical] studies of foundational validity and should demonstrate that the samples used in the foundational studies are relevant to the facts of the case.[59]

Acoustic measurements made on voice recordings result in continuously-valued data with within-speaker variability.[60] For such data it is not appropriate to include a "match"/"non-match" stage, *i.e.*, a stage which assesses "whether the features in an evidentiary sample and the features in a sample from a suspected source lie within a pre-specified measurement tolerance".[61] Such a procedure suffers from a

---

[55] Jonas R. Rappeport, *Reasonable medical certainty*, 13 Bulletin of the American Academy of Psychiatry Law 5–15 (1985)

[56] National Commission on Forensic Science, *Recommendations on use of the term "reasonable scientific certainty"* (2016) https://www.justice.gov/ncfs/file/839726/download; PCAST 2016 note 19 *supra* at p. 19

[57] An expanded version of this section is published as Geoffrey Stewart Morrison, David H. Kaye, David J. Balding, Duncan Taylor, Philip Dawid, Colin G.G. Aitken, Simone Gittelson, Grzegorz Zadora, Bernard Robertson, Sheila Willis, Susan Pope, Martin Neil, Kristy A. Martire, Amanda Hepler, Richard D. Gill, Allan Jamieson, Jacob de Zoete, J., R. Brent Ostrum, & Amke Caliebe, *A comment on the PCAST report: Skip the "match"/"non-match" stage*. Forensic Science International (2016) http://dx.doi.org/10.1016/j.forsciint.2016.10.018 (hereinafter Morrison, Kaye, *et al.* 2016)

[58] PCAST 2016 note 19 *supra*

[59] *Ibid*. at p. 56. We note that the PCAST report actually confuses assessment of strength of evidence with empirical validation of system performance, see §4 *infra*. If the results of empirical tests form the basis for the calculation of the strength of the evidence (via correct acceptance and false acceptance rates), then a second set of empirical tests using a separate set of test data (or using cross validation) should be conducted to assess the performance of the system which calculates the strength of the evidence.

[60] Discrete (non-continuous) data can have values such a 1, 2, or 3, but do not allow for intermediate values such as 1.5, 1.9, 1.99999, or 2.00001. Continuously-values data allow any value to occur. Measurements made on objects of interest in many branches of forensic science will naturally result in continuously-valued data with within-source variability.

[61] PCAST 2016 note 19 *supra* at p. 48

cliff-edge effect: A questioned-source specimen which falls just above the threshold for "match" with the known-source sample and a questioned-source specimen which falls just below the threshold will result in very different conclusions as to the strength of the evidence, even though the difference between the two is negligible (the two specimens could in fact be from the same source, with the difference between them due to within-source variability). Also, a procedure that includes a "match"/"non-match" stage limits the strength-of-evidence conclusion to one of two possible values: A questioned-source specimen which vastly exceeds the threshold will be assessed as having exactly the same strength of evidence as a questioned-source specimen which just exceeds the threshold, even if the former should in theory constitute much stronger evidence than the latter. *Mutatis mutandis* for a specimen which falls just short of the threshold and one which falls far below the threshold.

A more appropriate procedure for continuously-valued data with within-source variability would calculate a likelihood ratio using statistical models which work directly with the continuously-valued measurements. The history of forensic science includes multiple examples in which two-stage procedures were proposed and used, but subsequently replaced by procedures which use statistical models that directly calculate likelihood ratios from continuously-valued measurements. See Aitken & Taroni (2004)[62] pp. 10–11, and Foreman *et al*. (2003)[63] pp. 474–476 for examples from glass and DNA respectively.

3.7 UK framework

Another framework that the courts could potentially encounter is the so-called UK framework. This comes from a position statement produced in 2007 by a number of forensic voice comparison practitioners and researchers in the United Kingdom.[64] The framework is similar to the two-stage framework in that it first has a "match"/"non-match" stage. The practitioner first makes a subjective judgment as to "whether the known and questioned samples are compatible, or consistent, with having been produced by the same speaker".[65] The choices are "consistent", "not consistent", or "no-decision". If "consistent", the practitioner then makes a subjective judgment as to whether the known- and

[62] Colin G.G. Aitken & Franco Taroni, Statistics and the Evaluation of Forensic Evidence for Forensic Scientist (2nd ed. 2004)

[63] Lindsey A. Foreman, Christophe Champod, Ian W. Evett, J.A. Lambert, Susan Pope, *Interpreting DNA evidence: A review*, 71 International Statistical Review 473–495 (2003) http://dx.doi.org/10.1111/j.1751-5823.2003.tb00207.x

[64] J. Peter French & Philip Harrison, *Position Statement concerning use of impressionistic likelihood terms in forensic speaker comparison cases*, 14 International Journal of Speech, Language and the Law 137–144 (2007) http://dx.doi.org/10.1558/ijsll.v14i1.137 (hereinafter French & Harrison 2007); J. Peter French, Francis Nolan, Paul Foulkes, Philip Harrison, & Kirsty McDougall, *The UK position statement on forensic speaker comparison: a rejoinder to Rose and Morrison*, 17 International Journal of Speech, Language and the Law 143–152 (2010) http://dx.doi.org/10.1558/ijsll.v17i1.143 (hereinafter French *et al*. 2010)

[65] French & Harrison 2007 note 64 *supra* at p. 141

questioned-speaker recordings fall into one of five levels of distinctiveness with respect to the population: "exceptionally-distinctive", "highly-distinctive", "distinctive", "moderately-distinctive", or "not-distinctive". The framework also allows for "categorical statements of identification",[66] and "making the statement that the samples are spoken by different speakers".[67]

The UK position statement has been criticized as not logically tenable, for suffering from cliff-edge effects, and for failing to consider testing of validity and reliability.[68] Unlike the PCAST procedure, the UK procedure is not based on empirical validation. Instead, it is based on "research literature and general experience", with "education, training and experience" as pre-requisites.[69]

> forensic phoneticians ... need to judge the distinctiveness of the features found in the criminal and suspect samples ... informally via the analyst's experience and general linguistic knowledge rather than formally and quantitatively.[70]

As of 2015, the lead authors of the UK position statement have abandoned the framework proposed in that document in favor of the likelihood ratio framework. In a presentation on 7 September 2015 at the Interspeech conference in Dresden, Germany, Dr Philip Harrison of JP French Associates stated that they had adopted the Association of Forensic Science Providers' standards,[71] which require the reporting of either a numeric or a verbal likelihood ratio.

4 Testing validity and reliability

The 2009 NRC Report to Congress on *Strengthening Forensic Science in the United States*,[72] was highly critical of existing practice in many branches of forensic science. Its recommendations for improvements included "The development and establishment of quantifiable measures of the reliability and accuracy of forensic analyses" (p. 23). The Forensic Science Regulator of England & Wales (hereinafter FSR) has mandated that, in all branches of forensic science, the methods applied be validated prior to being used

---

[66] *Ibid*. at p. 142

[67] *Ibid*. at p. 141

[68] Philip J. Rose, Geoffrey Stewart Morrison, *A response to the UK position statement on forensic speaker comparison*, 16 International Journal of Speech, Language and the Law 139–163 (2009) http://dx.doi.org/10.1558/ijsll.v16i1.139; Morrison 2009 note 29 *supra*; Morrison 2010 note 1 *supra*; Morrison 2014 note 8 *supra*

[69] French & Harrison 2007 note 64 *supra* at p. 138

[70] French *et al*. 2010 note 64 *supra* at p. 144

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

[72] National Research Council, Strengthening Forensic Science in the United States: A Path Forward (2009) http://www.nap.edu/catalog.php?record_id=12589 (hereinafter NRC 2009)

to perform analyses for presentation to the courts.[73] The NRC and the US National Commission on Forensic Science (hereinafter NCFS) have also both recommended that all forensic science providers be accredited, which includes a requirement to conduct method validation.[74] Morrison (2014)[75] presents a review of calls from the 1960s onwards for the validity and reliability of forensic voice comparison to be empirically tested under casework conditions.

The PCAST report opines:

> Without appropriate estimates of accuracy, an examiner's statement that two samples are similar—or even indistinguishable—is scientifically meaningless: it has no probative value, and considerable potential for prejudicial impact.[76]

In science, validity and reliability are distinct concepts. The terms *validity* and *reliability*, can be used with a broader range of meanings, but we use *validity* as a synonym of *accuracy*, and *reliability* as a synonym of *precision*. Fig. 2 illustrates the difference between these concepts. Imagine that we have four archers who each fire arrows at a target. One of the archers has a tight grouping of arrows, this archer's results are reliable/precise, but on average the arrows are far from the center of the target, this archer's results are not valid/accurate. For another archer, averaging over the location of all the arrows, that average is close to the center of the target, this archer's results are valid/accurate, but the spread of the arrows is wide, this archer's results are not reliable/precise. A third archer has results which are neither valid/accurate nor reliable/precise, they have a wide spread and on average are not close to the center of the target. A fourth archer is both valid/accurate and reliable/precise, this archer has a tight grouping of arrows and on average they are close to the center of the target. We have described these results in terms of valid versus not valid and reliable versus not reliable (accurate versus not accurate and precise versus not precise), but it is important to understand that these are not binary concepts, there are degrees of greater or lesser validity and degrees of greater or lesser reliability (degrees of greater or lesser accuracy and degrees of greater or lesser precision).

[73] Forensic Science Regulator, Codes of practice and conduct for forensic science providers and practitioners in the criminal justice system (v. 2.0, 2014) https://www.gov.uk/government/publications/forensic-science-providers-codes-of-practice-and-conduct-2014 (last visited Oct 27, 2016) (hereinafter FSR 2014)

[74] NRC 2009 note 72 *supra*; National Commission on Forensic Science, Universal Accreditation (2015) http://www.justice.gov/ncfs/file/477851/download (last visited Oct 27, 2016) (hereinafter NCFS 2015 *universal*)

[75] note 8 *supra*

[76] PCAST 2016 note 19 *supra* at p. 6

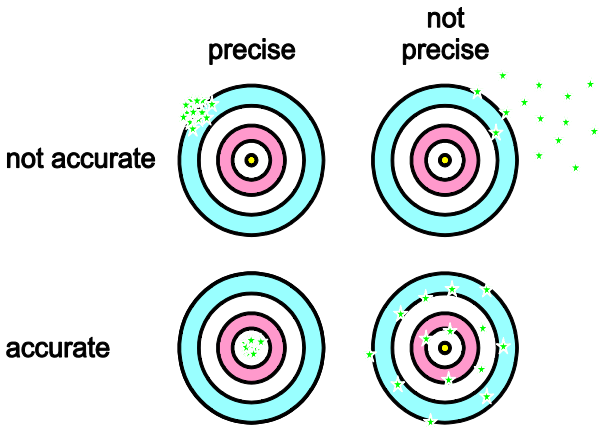

Figure 2. Results of four different archers firing arrows at a target. Each archer has a different pattern of validity and reliability (accuracy and precision).

In legal literature, terms such as validity and reliability are seldom explicitly and unambiguously defined. From context, it is clear that when legal texts use the term *reliability* they are often primarily concerned with what scientists would call *validity*. In the present paper we use the terms validity and reliability with their scientific meanings.

Empirical testing of validity and reliability is the only way to demonstrate how well a forensic analysis system actually works. We use the term *system* to designate the whole of the data and the processes used to evaluate the strength of evidence after the forensic scientist has stated what competing hypotheses they intend to evaluate. A forensic voice comparison system can include the sampling and selection of relevant voice recordings, procedures used to measure properties of the voice recordings, and statistical models used to calculate values which will be reported. A system also includes any actions taken by a human. The forensic practitioner is part of the system. A system could be entirely a human who listens and makes subjective judgments. Empirical testing of validity and reliability should be blind to the internal workings of the system. The system could be auditory, spectrographic, acoustic-phonetic non-statistical, acoustic-phonetic statistical, automatic, or a room full of monkeys with keyboards. The testing procedure would treat each system as a black box. The only condition would be that any system to be tested conform to the input and output requirements of the test protocol.

In order to empirically test the validity of a forensic voice comparison system,[77] one must have a set of test data. The data must include pairs of voice recordings for which the tester knows that the two members of each pair were produced by the same speaker. The test data must also include pairs of voice recordings for which the tester knows that the two member of each pair were produced by different speakers. What constitutes relevant test data will vary from case to case. There can be major differences between cases

[77] With appropriate changes in vocabulary, *etc*., the following also applies to testing the validity and reliability of systems which compare other objects of forensic interest.

as to what constitutes the relevant population, speaking styles, and recording conditions. The variability between cases is generally such that the results of a test of the performance of a forensic voice comparison system under the conditions of one case may be very different to the performance of that system under the conditions of another case. A forensic voice comparison system which works well under good recording conditions may work poorly under conditions which include background noise, reverberation, and transmission through communication systems, and it may work especially poorly when there is a mismatch between known-speaker and questioned-speaker recording conditions. The test data must therefore be representative of the relevant population for the case, and the conditions of one member of each test pair must reflect the speaking style and recording conditions of the known-speaker recording, and the conditions of the other member of each test pair must reflect the speaking style and recording conditions of the questioned-speaker recording.[78]

A key question is who decides whether the test data are sufficiently representative of the relevant population and sufficiently reflective of the conditions of the known-speaker and questioned-speaker recordings? In the first instance the tester must be satisfied. Who must be satisfied in the context of an admissibility hearing is a question we address in Section 6.2 below.

A general protocol for testing the validity of a forensic voice comparison system is as follows: A pair of voice recordings is presented to the system, one recording with conditions reflecting those of the known-speaker recording and the other with conditions reflecting those of the questioned-speaker recording. The tester knows whether this pair of recordings is a same-speaker pair or a different-speaker pair, but the system being tested must not be told which of these is true. The tester compares the output of the system with their knowledge about whether the input was a same-speaker pair or a different-speaker pair, and assesses the goodness of the output accordingly. A large number of same-speaker and different-speaker pairs are presented to the system, and the average goodness assessed. Additional details of this protocol are presented in Appendix B.

For any system in which the conclusion is based primarily or directly on subjective judgment, the tester must not be the same person as the practitioner who performs the forensic voice comparison. The tester must know the truth as to whether each pair is a same-speaker pair or a different-speaker pair, but the person being tested must not. For systems in which the conclusion is directly the output of a statistical model, and subjective judgment is confined to early parts of the process (selecting relevant training data

[78] The principle also applies across other branches of forensic science. The PCAST report notes that "for DNA analysis, the frequency of genetic variants is known to vary among ethnic groups; it is thus important that the sample collection reflect relevant ethnic groups to the case at hand. For latent fingerprints, the risk of falsely declaring an identification may be higher when latent fingerprints are of lower quality; so, to be relevant, the sample collections used to estimate accuracy should be based on latent fingerprints comparable in quality and completeness to the case at hand." (PCAST 2016 note 19 *supra* at p. 56). In the context of forensic analysis of physicochemical data, e.g., measurements of glass fragments, flammable liquid residue, car paint, fibers, and ink, Grzegorz Zadora, Agnieszka Martyna, Daniel Ramos, & Colin Aitken, Statistical Analysis in Forensic Science: Evidential Value of Multivariate Physicochemical Data (2014) §6.2 also stress the need for test data to reflect the conditions of the forensic cases to which forensic analysis systems will be applied.

and appropriate statistical models to use *etc*.), the tester can be the same person as the practitioner who performs the forensic voice comparison. In fact in the latter case the test procedure will be automated – the tester will select appropriate test data and then have a computer program test the forensic voice comparison system using those data.

Measured and calculated numbers in science are not absolutely precise, they have a degree of imprecision. It is good practice in science to quantify the degree of imprecision. Several factors can affect the precision (reliability) of a forensic voice comparison system, including intrinsic variability at the source, sampling variability, and measurement variability. For example, using one recording of the known speaker rather than another, or using one sample of the relevant population rather than another, or re-measuring the same recordings again can result in a different value for a calculated likelihood ratio. There are several solutions proposed for dealing with imprecision in forensic likelihood ratios. We discuss testing the precision (reliability) of a forensic voice comparison system in Appendix C.

Whatever the approach or combination of approaches used, what needs to be tested is the entire system. Only knowing the performance of components of a system would not be sufficient.[79] For example, in an acoustic-phonetic approach, it would not be enough to test the performance of the tools and procedures used for measuring fundamental frequency. The performance of such tools could be excellent, but if fundamental frequency measurements contain little useful information, or if subsequent components of the system are not able to effectively utilize the information they may contain, then the validity and reliability of the output of the system will be poor. Similarly, if the output of multiple systems are combined (*e.g.*, an automatic system and an acoustic-phonetic statistical system), it is the combined system which must be tested. If a practitioner uses the output of an automatic system or an acoustic-phonetic statistical system as input to a subjective judgment process, then it is the output of the final subjective judgment process which must be tested. The system which needs to be tested is the system which is actually used to evaluate the strength of evidence in the actual case.

Whereas automatic systems (and to a lesser extent acoustic-phonetic statistical systems) can quickly and cheaply run hundreds or thousands of test comparisons, for systems which are based primarily on subjective judgment (and systems in which the final stage is a subjective judgment) each test comparison may take considerable investment of a human practitioner's time. The higher time and financial costs, however, should not excuse subjective judgment systems from the requirement that they be tested. If the time and financial costs are such that a subjective judgment system cannot be adequately tested, then the system should not be used. Experience is not a substitute for empirical testing.

Experience in applying spectrographic voice identification in law enforcement has led

[79] Forensic Science Regulator, Draft Guidance: Digital Forensics Method Validation (FSR-G-218 Second consultation, 2015) https://www.gov.uk/government/consultations/digital-forensics-method-validation-draft-guidance-second-consultation (last visited Oct 27, 2016) at §13.3

> proponents of the method to express confidence its reliability. The basis for this confidence is not, however, accessible to objective assessment.[80]
>
> For an expert to say "I think this is true because I have been doing this job for *x* years" is, in my view, unscientific. On the other hand, for an expert to say "I think this is true and my judgement has been tested in controlled experiments" is fundamentally scientific.[81]
>
> Validation of this approach to voice identification becomes a matter of replicable experiments on the expert himself, considered as a voice identifying machine. ... validation requires experimental assessment of performance on relevant tasks. ... It may be objected that this minimal set of tests is unreasonably arduous. We do not believe that it is. As scientists we could accept no less in checking the reliability of a "black box" supposed to perform speaker identification.[82]

The PCAST report opines (emphasis in original):

> 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.[83]

And recommends that:

> Where there are not adequate empirical studies and/or statistical models to provide meaningful information about the accuracy of a forensic feature-comparison method, DOJ attorneys and examiners should not offer testimony based on the method.[84]

Readers interested in more comprehensive introductions to measuring the validity and reliability of

---

[80] Bolt *et al*. 1970 note 5 *supra* at p. 603

[81] Ian W. Evett, *Interpretation: a personal odyssey*, *in* The Use of Statistics in Forensic Science 9–22 (Colin G.G. Aitken & David A. Stoney, 1991) at p. 21

[82] Bolt *et al*. 1970 note 5 *supra* at p. 602

[83] PCAST 2016 note 19 *supra* at p. 6

[84] *Ibid*. at p. 19

forensic analysis systems may wish to consult Appendices A and B, Morrison (2011),[85] and Meuwly *et al*. (2016).[86]

A multi-laboratory evaluation of multiple forensic voice comparison systems under conditions reflecting those of one real forensic case is currently under way, and the results are being published in a virtual special issue of the journal *Speech Communication*.[87]

5 Contextual bias

The 2009 NRC report found that:

> forensic science experts are vulnerable to cognitive and contextual bias, ... Contextual information renders experts vulnerable to making erroneous identifications.[88]

The PCAST report advised that:

> Subjective methods require particularly careful scrutiny because their heavy reliance on human judgment means they are especially vulnerable to human error, inconsistency across examiners, and cognitive bias. In the forensic feature-comparison disciplines, cognitive bias includes the phenomena that, in certain settings, humans may tend naturally to focus on similarities between samples and discount differences and may also be influenced by extraneous information and external pressures about a case.[89]

The NCFS recommended that:

> Forensic laboratories should take appropriate steps to avoid exposing analysts to task-irrelevant information through the use of context management procedures detailed in written policies and protocols.[90]

Concern about contextual bias in forensic science arose in part from empirical studies showing that forensic practitioners are sometimes influenced by information that is irrelevant to their assessment of

---

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

[86] Didier Meuwly, Daniel Ramos, Rudolf Haraksim, *A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation*, Forensic Science International (2016) http://dx.doi.org/10.1016/j.forsciint.2016.03.048

[87] http://www.sciencedirect.com/science/journal/01676393/vsi/10KTJHC7HNM

[88] NRC 2009 note 72 *supra* at p. 4 note 8

[89] PCAST 2016 note 19 *supra* at p. 5

[90] National Commission on Forensic Science, Ensuring that forensic analysis is based upon task-relevant information (2015) https://www.justice.gov/ncfs/file/818196/download (last visited Oct 27, 2016) (hereinafter NCFS 2015 *task*)

the evidence.[91] For example, latent print examiners were less likely to report a "match" between a latent print from a crime scene and a suspect's print when they were told the suspect had a solid alibi.[92] Contextual bias is not, however, limited to forensic scientists. It is a universal phenomenon that affects decision making by people from all walks of life and in all professional settings, including science.[93] People are particularly vulnerable to contextual bias when making judgments based on data that may be somewhat ambiguous and subject to differing interpretations. Contextual bias occurs without conscious awareness; it does not require misconduct or bad intent.[94] Rather, exposure to contextual information can bias the conclusions of forensic practitioners who perform their jobs with utmost honesty and professional commitment.

Forensic practitioners who rely on subjective judgment to reach conclusions may need to evaluate data that are somewhat ambiguous and subject to differing interpretations. Under these circumstances, there is clearly a potential for practitioners to be influenced by contextual bias.[95] Such circumstances will inevitably arise in approaches to forensic voice comparison in which practitioners rely heavily on

[91] The term *context effect* originated in psychology and has been used to describe circumstances in which the perception of a stimulus is affected by the surrounding context, "as where a gray object looks lighter against a dark background than against a light background", William C. Thompson, *Interpretation: observer effects*, *in* Wiley Encyclopedia of Forensic Science 1575–1579 (Allan Jamieson & Andre A. Moenssens, 2009) (hereinafter Thompson 2009). In forensic science, the term *context effect* has been used more broadly to describe situations in which the results of a forensic analysis are affected by the information available to the analyst, "as when an analyst becomes more likely to identify a latent print as that of a suspect when told that another analyst has already made the identification or when told that other evidence indicates the suspect made the print" *ibid*. The "other evidence" might be said to provide a "context" that changes the analyst's interpretation of the scientific data. When the "other evidence" includes information that should have no bearing on the analyst's judgment, the phenomenon is called a *contextual bias.* See: D. Michael Risinger, Michael J. Saks, William C. Thompson, & Robert Rosenthal, The Daubert/Kumho implications of observer effects in forensic science: Hidden problems of expectation and suggestion, 90(1) California Law Review 1–56 (2002) http://www.jstor.org/stable/3481305 (hereinafter Risinger *et al*. 2002); William C. Thompson, *What role should investigative facts play in the evaluation of scientific evidence?* 43 Australian Journal of Forensic Sciences 123–134 (2011) (hereinafter Thompson 2011); NCFS 2015 *task* note 90 *supra*

[92] Itiel E. Dror & David Charlton, *Why experts make errors*, 56 Journal of Forensic Identification 600–616 (2006); Itiel E. Dror, David Charlton, & Ailsa E. Peron, *Contextual information renders experts vulnerable to making erroneous identifications*, 156 Forensic Science International 174–178 (2006); Itiel E. Dror & Robert Rosenthal, *Meta-analytically quantifying the reliability and biasability of forensic experts*, 53 Journal of Forensic Sciences 900–903 (2008); Expert Working Group on Human Factors in Latent Print Analysis, Latent Print Examination and Human Factors: Improving the Practice through a Systems Approach (2012) http://www.nist.gov/manuscript-publication-search.cfm?pub_id=910745 (hereinafter NIST/NIJ 2012)

[93] Saul M. Kassin, I.E. Dror & J.Kukucka, *The forensic confirmation bias: problems, perspectives, and proposed solutions*. 2 J. Applied Res. Mem. Cognition 42–52 (2013); Risinger *et al*. 2002 note 91 *supra*

[94] Thompson 2011 note 91 *supra*

[95] see reviews in Risinger *et al*. 2002 note 91 *supra*; Michael J. Saks, D. Michael Risinger, Robert Rosenthal, & William C. Thompson, *Context effects in forensic science: a review and application of the science of science to crime laboratory practice in the United States*, 43 Science & Justice 77–90 (2003) http://dx.doi.org/10.1016/S1355-0306(03)71747-X; Bryan Found, *Deciphering the human condition: The rise of cognitive forensics*, 47 Australian Journal of Forensic Sciences 386–401 (2015) http://dx.doi.org/10.1080/00450618.2014.965204; Reinoud D. Stoel, Charles E.H. Berger, Wim Kerkhoff, Erwin J.A.T. Mattijssen & Itiel E. Dror, *Minimizing contextual bias in forensic casework*, *in* Forensic Science and the Administration of Justice: Critical Issues and Directions 67–86 (Kevin J. Strom & Matthew J. Hickman, 2015) http://dx.doi.org/10.4135/9781483368740.n5

subjective judgment.

One way to minimize contextual bias is to avoid exposing practitioners to "task-irrelevant" information, *i.e.*, information that is not necessary for assessing the strength of the forensic evidence.[96] *Context management procedures* (often called blinding procedures) are used to prevent bias in many areas of science.[97] Although context management is relatively new in forensic science, procedures for implementing context management have been discussed extensively in forensic science literature,[98] and many laboratories have implemented such procedures, including for forensic DNA analysis and latent print analysis.[99]

A key issue in implementing a context management procedure is determining which information is relevant and irrelevant to a particular task. In forensic voice comparison, task-relevant information would clearly include information about and the recording conditions, and information pertinent to understanding what would constitute the relevant population. Task-irrelevant information would include the crime that the defendant is charged with, the results of other forensic analyses such as DNA and fingerprint analyses, and whether the forensic voice comparison analysis has been requested by the prosecution or the defense.

There are several ways to prevent practitioners from being exposed to task-irrelevant information. In large laboratories it may be practical to use a case manager to interact with the client and decide what constitutes task-relevant and task-irrelevant information for a practitioner. The case manager then passes on to the practitioner only the task-relevant information. Sometimes, information that is task-irrelevant and potentially biasing at one stage of an analysis becomes necessary and task-relevant at a later stage. For example, information about the DNA profile of a suspect is unnecessary and biasing when determining what potential profiles are present in a DNA mixture, but necessary when determining the

[96] see NCFS 2015 *task* note 90 *supra*; Thompson, W.C. (2016). Determining the Proper Evidentiary Basis for an Expert Opinion: What Do Experts Need to Know and When Do They Know Too Much? in *Blinding as a Solution to Bias in Biomedical Science and the Courts: A Multidisciplinary Approach*. C Robertson and A. Kesselheim, Eds. Elsevier; Itiel E. Dror, William C. Thompson, Christian A. Meissner, Irv Kornfield, Dan Krane, Michael Saks & D. Michael Risinger, *Context management toolbox: A linear sequential unmasking (LSU) approach for minimizing cognitive bias in forensic decision making*, 60 Journal of Forensic Sciences 1111–1112 (2015) (hereinafter Dror *et al*. 2015)

[97] Risinger *et al*. 2002 note 91 *supra*

[98] Simon A. Cole, *Implementing counter-measures against confirmation bias in forensic science*, 2 Journal of Applied Research in Memory and Cognition 61–62 (2013); Itiel E. Dror & Simon A. Cole, *The vision in 'blind' justice: Expert perception, judgment, and visual cognition in forensic pattern recognition*, 17 Psychonomic Bulletin & Review 161-167 (2010); Bryan Found & John Ganas, *The management of domain irrelevant context information in forensic handwriting examination casework*, 53 Science & Justice 154–158 (2013); Reinoud D. Stoel, Charles E.H. Berger, Wim Kerkhoff, Erwin J.A.T. Mattijssen & Itiel E. Dror, *Minimizing contextual bias in forensic casework*, *in* Forensic Science and the Administration of Justice: Critical Issues and Directions 67–86 (Kevin J. Strom & Matthew J. Hickman, 2015) http://dx.doi.org/10.4135/9781483368740.n5 (hereinafter Stoel et al 2015); Thompson 2011 note 91 *supra*; Dror *et al*. 2015 note 96 *supra*.

[99] NIST/NIJ 2012 note 92 *supra*; Stoel *et al*. 2015 note 98 *supra*; NCFS 2015 *task* note 90 *supra*

probability of the DNA mixture evidence if the suspect were a contributor versus if they were not a contributor.[100] The potential for bias can be reduced by withholding the potentially-biasing information from the analyst until it is needed, a procedure known as *sequential unmasking*.[101] Another example is found in latent print analysis, where some laboratories require examiners to evaluate poor quality latent prints from crime scenes, and to identify all the points (minutiae) that they consider relevant, before they see the high-quality known-origin print image.[102] Withholding information about the known-origin print prevents the examiner from being biased towards seeing indistinct parts of the poor quality questioned-origin image as having the same pattern as in the high-quality known-origin image.[103]

Another way to reduce the potential for cognitive bias is to avoid using approaches in which the strength of evidence conclusion is primarily or directly based on subjective judgment. As previously mentioned, an approach based on relevant data, quantitative measurements, and statistical models distances subjective elements from the final output of the system (subjective elements include decisions as to what constitute relevant data for training and testing the system). As long as the likelihood ratio output by the statistical model is directly reported as the strength of evidence statement, such a system is therefore much less susceptible to the potential influence of contextual bias.[104]

6 Admissibility

In this section, we will review the legal standards for admissibility of expert testimony established by Federal Rule of Evidence 702, *Daubert v. Merrell Dow Pharmaceuticals* (1993), and *Frye v. U.S.* (1923). We also consider how these standards have been and should be applied when evaluating the admissibility of forensic voice comparison testimony.

The admissibility of scientific evidence in Federal courts is governed by Rule 702, which states:

> Testimony by Expert Witnesses
>
> A witness who is qualified as an expert by knowledge, skill, experience, training, or education

[100] Thompson 2009 note 91 *supra*; Itiel E. Dror & Greg Hampikian, *Subjectivity and bias in forensic DNA mixture interpretation*, 51 Science & Justice 204–208 (2011)

[101] Dan E. Krane, Simon Ford, Jason R. Gilder, Keith Inman, Allan Jamieson, Roger Koppl, Irving L. Kornfield, D. Michael Risinger, Norah Rudin, Marc Scott Taylor,William C. Thompson, Sequential unmasking: A means of minimizing observer effects in forensic DNA interpretation, 53 Journal of Forensic Sciences 1006–1007 (2008)

[102] Dror *et al*. 2015 note 96 *supra*

[103] NIST/NIJ 2012 note 92 *supra*

[104] For additional arguments as to why the output of the statistical model should be directly reported and not used as input to a subjective judgment process, see Geoffrey Stewart Morrison & Reinoud D. Stoel, *Forensic strength of evidence statements should preferably be likelihood ratios calculated using relevant data, quantitative measurements, and statistical models – a response to Lennard (2013) Fingerprint identification: How far have we come?* 46 Australian Journal of Forensic Sciences 282–292 (2014) http://dx.doi.org/10.1080/00450618.2013.833648

> may testify in the form of an opinion or otherwise if:
>
> (a) the expert's scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue;[105]
>
> (b) the testimony is based on sufficient facts or data;
>
> (c) the testimony is the product of reliable principles and methods; and
>
> (d) the expert has reliably applied the principles and methods to the facts of the case.[106]

The United States Supreme Court addressed the admissibility of expert evidence in a series of cases that began with *Daubert v. Merrell Dow Pharmaceuticals* (1993) and included *General Electric v. Joiner* (1997)[107] and *Kumho Tire v. Carmichael* (1999).[108] The Court in *Daubert* explained that Federal Rule of Evidence 702 requires the trial judge to act as gatekeeper to "ensure that any and all scientific testimony or evidence admitted is not only relevant, but reliable."[109] The Court used the term "reliable" to refer to "*evidentiary* reliability–that is, trustworthiness."[110] It explained that: "In a case involving scientific evidence, *evidentiary reliability* will be based upon *scientific validity*."[111]

Before the *Daubert* ruling, most courts applied an admissibility test articulated in *Frye v. United States* (1923) which required courts to determine whether a method had "general acceptance in the particular field to which it belongs." *Daubert* rejected the idea that "general acceptance" should be the sole criterion for admissibility, but retained it as one of several factors for federal judges to consider when deciding whether expert testimony meets the requirements of Rule 702. In state courts, judges follow state evidence codes that sometimes differ from the Federal Rules of Evidence. Although many states have adopted the *Daubert* standard, some states, including California[112] and New York, continue to use

[105] The *Daubert* ruling states that: "The adjective 'scientific' implies a grounding in the methods and procedures of science. Similarly, the word 'knowledge' connotes more than subjective belief or unsupported speculation." 509 U.S. at 590

[106] Rule 702 as amended Apr. 17, 2000, eff. Dec. 1, 2011. When *Daubert*, *Joiner*, and *Kumho Tire* were decided, Rule 702 read as follows: "If scientific, technical or other specialized knowledge will assist the trier of fact to understand the evidence or to determine a fact in issue, a witness qualified as an expert by knowledge, skill, experience, training, or education, may testify thereto in the form of an opinion or otherwise." For a history of the amendment process, and the subsequent failure of some courts to abide by the amended version of the Rule, see David E. Bernstein & Eric G. Lasker, *Defending Daubert: It's time to amend Federal Rule of Evidence 702*, 57(1) William and Mary Law Review 1–48 (2015).

[107] *General Electric Co. v. Joiner*, 522 U.S. 136 (1997)

[108] *Kumho Tire Co. v. Carmichael*, 526 U.S. 137 (1999)

[109] *Daubert*, 509 U.S. at 589.

[110] *Daubert*, 509 U.S. at n. 9.

[111] *Ibid*. (emphasis in original).

[112] In California the applicable standard is known as *Kelly/Frye* because the state supreme court adopted the *Frye* standard in *People v. Kelly* [(1976) 17 Cal.3d 24], a case concerning the admissibility of testimony based on the spectrographic approach to forensic voice comparison. The court ruled that the proponent in that case had failed to demonstrate that the spectrographic approach was generally accepted by the scientific community.

versions of the *Frye* general acceptance test. Outside of the U.S., Section 33 of the 2014 Criminal Practice Directions in England & Wales[113] is clearly based (at least in part) on FRE 702 / *Daubert*, and the *Daubert* criteria have also influenced Canadian decisions on admissibility of expert evidence.[114]

In *Daubert*, the Supreme Court provided a non-exclusive list of factors for courts to consider when evaluating whether scientific testimony meets the requirements of Rule 702:

- Whether the reasoning or methodology underlying the testimony is scientifically valid and ... whether that reasoning or methodology properly can be applied to the facts in issue.
- Whether [a theory or technique] can be (and has been) [empirically] tested. ... In the case of a particular scientific technique, the court ordinarily should consider the known or potential rate of error.
- Whether the theory or technique has been subjected to peer review and publication.
- In the case of a particular scientific technique, ... the existence and maintenance of standards controlling the technique's operation.
- General acceptance within a relevant scientific community.

Below, we consider each of these requirements in turn. First, however, we discuss what the Supreme Court called the consideration of "fit", *i.e.*:

- whether expert testimony proffered in the case is sufficiently tied to the facts of the case that it will aid the jury in resolving a factual dispute[115]

6.1 Whether expert testimony is sufficiently tied to the facts of the case

Rule 702 requires that expert evidence "help the trier of fact to understand the evidence or to determine a fact in issue." *Daubert* explained that this helpfulness requirement of Rule 702 is essentially a matter of "fit" between the expert evidence and the issue upon which it is offered. There must be "a valid scientific connection to the pertinent inquiry as a precondition to admissibility."[116] In other words,

[113] Criminal Practice Directions Amendment No. 2 [2014] EWCA Crim 1569; Law Commission, Expert Evidence in Criminal Proceedings in England and Wales (2011) http://www.lawcom.gov.uk/expert_evidence.htm

[114] Graham D. Glancy & John M.W. Bradford, *The admissibility of expert evidence in Canada*, 35 Journal of the American Academy of Psychiatry and the Law 350–356 (2007); Gary Edmond & Kent Roach, *A contextual approach to the admissibility of the state's forensic science and medical evidence*, 61 University Of Toronto Law Journal 343–409 (2011) http://dx.doi.org/10.3138/utlj.61.3.343

[115] *Daubert* quoting *United States v. Downing*, 753 F. 2d 1224, 1242 (CA3 1985)

[116] *Daubert* 509 U.S. at 590.

admissibility depends not only on whether the expert's evidence is trustworthy *per se* but also whether it addresses in a scientifically valid manner an issue properly before the trier of fact.[117]

In *General Electric v. Joiner* the Supreme Court offered further clarification of the requirements of Rule 702, giving particular emphasis to the need for a reasonably close connection between any data on which the expert relies and the conclusions that the expert draws from it with respect to the case under consideration. The court upheld a trial judge's decision to exclude expert testimony linking PCB exposure to respondent Joiner's cancer where the expert's conclusion was supported only by animal studies:

> Trained experts commonly extrapolate from existing data. But nothing in either *Daubert* or the Federal Rules of Evidence requires a district court to admit opinion evidence that is connected to existing data only by the *ipse dixit* of the expert. A court may conclude that there is simply too great an analytical gap between the data and the opinion proffered.[118]

A subsequent case, *Kumho Tire,* also emphasized the need for a close connection between the expert's methodology and the conclusion drawn. The issue the judge must consider when evaluating admissibility is not the general validity of the underlying theory or method but whether it is valid for drawing the specific kinds of conclusions that the expert drew in the case at hand:

> the specific issue before the court was not the reasonableness *in general* of a tire expert's use of a visual and tactile inspection …. Rather, it was the reasonableness of using such an approach, along with [the expert's] particular method of analyzing the data thereby obtained, to draw a conclusion regarding *the particular matter to which the expert testimony was directly relevant*.[119]

Following *Kumho Tire*, Rule 702 was revised in a manner that further emphasized the need for a case-

---

[117] To illustrate how scientific knowledge may be valid but unhelpful, the court gave the following example: "The study of the phases of the moon, for example, may provide valid scientific 'knowledge' about whether a certain night was dark, and if darkness is a fact in issue, the knowledge will assist the trier-of-fact. However (absent creditable grounds supporting such a link), evidence that the moon was full on a certain night will not assist the trier of fact in determining whether an individual was unusually likely to have behaved irrationally on that night." *Id*. at 482; see also, *In re* Paoli R.R. Yard PCB Litig., 35 F.3d 717, 743 (3d Cir. 1994) reasoning that a valid connection between chemical exposure and animal cancer was insufficient to make the animal studies of carcinogenicity admissible because "there must be good grounds to extrapolate from animals to humans, just as the methodology of the studies must constitute good grounds to reach conclusions about the animals themselves." For further discussion of the concept of "fit," see D. Michael Risinger, *Defining the "task at hand": Non-science forensic science after Kumho Tire v. Carmichael*, 57 Washington and Lee Law Review 767 (2000).

[118] *Joiner*, 522 U.S. at 146. The analytic gap in *Joiner* arose from uncertainty about whether the exposure conditions modeled in the animal studies were sufficiently comparable to the conditions under which *Joiner* was exposed to render the research relevant. "Of course, whether animal studies can ever be a proper foundation for an expert's opinion was not the issue. The issue was whether *these* experts' opinions were sufficiently supported by the animal studies on which they purported to rely. The studies were so dissimilar to the facts presented in this litigation that it was not an abuse of discretion for the District Court to have rejected the experts' reliance on them." 522 U.S. at 144.

[119] *Kumho Tire*, 526 U.S. at 154 (emphasis in original)

specific inquiry into the validity and trustworthiness of expert evidence.[120] The new language requires, as a condition of admissibility, that "the expert has reliably applied the principles and methods to the facts of the case."

Kumho Tire, and the revised Rule 702, clarify which issues related to expert evidence must be considered by the judge as part of the *Daubert* inquiry, and which are matters of weight to be left to the trier of fact. Issues left to the trier of fact include assessing the veracity and sincerity of the expert, and determining how to weigh an expert's conclusions against other evidence in a case. A judge should not exclude expert testimony as unreliable simply because other evidence convinces the judge that the expert is wrong.[121] Issues for the judge at an admissibility hearing include any factors related to the scientific validity of the methods used by the expert for drawing conclusions of the type drawn in the case at hand. For forensic voice comparison, this would include questions about whether the expert chose an appropriate relevant population for the case at hand, whether they obtained a sample of data sufficiently representative of that relevant population, and whether the data or analytical techniques adequately accounted for the speaking styles and recording conditions in the known- and questioned-speaker recordings. We would argue that the judge should also consider whether the expert took adequate steps to avoid cognitive and contextual bias, as this also affects the validity of the expert's methods as applied and hence the trustworthiness of the expert's conclusions in the case at hand.[122] All of these matters relate to the relevance of the expert evidence to the "task at hand" and hence all go to admissibility under *Daubert*.

The need to evaluate such case-specific factors as part of the *Daubert* inquiry suggests that there can be no definitive precedent-setting ruling on whether a particular approach (auditory, spectrographic, acoustic-phonetic, or automatic) is admissible. Instead, courts will need to consider in each case whether there is adequate evidence that the system employed by the practitioner is sufficiently valid and reliable for the conditions in that case.[123]

---

[120] See note 106 *supra*.

[121] Nor should a judge admit expert evidence simply because other evidence suggests the expert's conclusion is correct. The judge must ask "whether the reasoning or methodology underlying the testimony is scientifically valid, and ... whether that reasoning or methodology properly can be applied to the facts in issue" (*Daubert*, 509 U.S. at 592–3), not whether, in light of other evidence, the expert is likely to be correct or incorrect. See also Risinger *et al*. 2002 note 91 *supra*.

[122] For a detailed discussion of the implications of cognitive and contextual bias for the admissibility of forensic science evidence under *Daubert* and *Kumho Tire*, see Risinger *et al*. 2002 note 91 *supra*.

[123] With respect to forensic voice comparison, that admissibility should be decided on a case by case basis because the application of science will be case specific is also the interpretation that Faigman *et al*. 2015 note 2 *supra* infer from the rulings in *State v. Coon*, 974 P.2d 386 (Alaska 1999) and in *Angleton*. *Coon* states: "The dissent [in *Contreras v. State*, 718 P.2d 129, 136 (Alaska 1986)] reaches a different conclusion because it begins with the premise that the scientific validity of a technique is a legal issue which does not turn on case-sensitive facts. This premise does not adequately take account of the reality of the judicial process and the variable state of science. ... the state of science is not constant; it progresses daily. ... We recognize that different trial judges, in exercising their discretion, may reach different conclusions about scientific reliability. ... The principal reason for adopting the *Daubert* standard is to give the courts greater flexibility in determining the admissibility of expert testimony, so as to keep pace with science as it evolves." 974 P.2d at 399. *Angleton* states: "The potential rate of error of the aural spectrographic method is unknown and may vary considerably, depending on the conditions of the particular

6.2 Empirical testing of validity and reliability

In considering the admissibility of expert testimony, the *Daubert* ruling instructs the trial judge to consider "whether the reasoning or methodology underlying the testimony is scientifically valid and ... whether that reasoning or methodology properly can be applied to the facts in issue."[124] It goes on to state that "a key question to be answered in determining whether a theory or technique is scientific knowledge that will assist the trier of fact will be whether it can be (and has been) tested. ... '[T]he statements constituting a scientific explanation must be capable of empirical test'". Later it states that "in the case of a particular scientific technique, the court ordinarily should consider the known or potential rate of error". We interpret these statements as requiring the forensic scientist to empirically test the degree of validity and reliability of their system and provide the results of such tests to the judge so that the judge can take them into consideration when deciding on admissibility. We further interpret "properly ... applied to the facts in issue" to imply that the tests of validity and reliability must be conducted under conditions which reflect those of the case under investigation (see discussion on "fit" in Section 6.1 above).

A key question is who decides whether the test data are sufficiently representative of the relevant population and sufficiently reflective of the conditions of the known-speaker and questioned-speaker recordings? In the first instance the tester must be satisfied. The tester may be the forensic practitioner if the approach is based on relevant data, quantitative measurements, and statistical models, in which case the testing is actually automated, or another member of the forensic laboratory or an outside party if the approach is based on subjective judgment. But ultimately it is the judge at an admissibility hearing and/or the trier of fact at trial who must be satisfied. The tester must therefore explain to the judge / trier of fact how they have sampled data from the relevant population and how they have selected or simulated data which reflect the conditions of the case. If the judge / trier of fact is not satisfied with what the tester has done then they need proceed no further and should rule the proffered testimony inadmissible / ignore whatever strength of evidence statement is produced by the system. If the judge / trier of fact is satisfied with what the tester has done, then they can consider the outcome of the tests as being representative of how the system will be expected to perform under the conditions of the case. In the latter case the judge should then consider whether the test results are good enough that testimony based on the system can be admitted, and if it is admitted the trier of fact can consider, based on the test results, the degree to which they will trust the output of the system.

It is worth noting that the Daubert opinion cited two prior appellate cases involving the admissibility of

application." 269 F.Supp.2d at 902.

[124] *Daubert*, 509 U.S. at 592–3.

forensic voice comparison testimony: *United States v. Williams* (1978)[125] and *United States v. Smith* (1989).[126] Both cases had been decided under the *Frye* standard, but in each case the court had gone beyond counting of scientific supporters and considered whether the proponents of the expert testimony had laid a "proper foundation" for establishing that the testimony was "reliable" and not likely to mislead.[127] These rulings were cited in *Daubert* in support of the court's assertion that the judge at an admissibility hearing "should consider the known or potential rate of error".[128,129]

During the *Daubert* hearing in *Angleton* the court considered research literature on auditory-spectrographic approaches covering a period of over 30 years.[130] Some earlier rulings on the admissibility of spectrographic evidence had acknowledged the criticisms of this approach found in the literature, but dismissed them as raising issues going to weight rather than admissibility. In contrast, the court in *Angleton* viewed these criticisms as going to the heart of matter – the scientific validity of the testimony. With respect to testing and error rates, the court in *Angleton* concluded that:

> The evidence and testimony show that there is great dispute among researchers and the few practitioners in the field over the accuracy and reliability of voice spectrographic analysis to determine the identity of recorded speakers. ... The *post-Daubert* case law casts doubt on the

[125] *United States v Williams*, 583 F. 2d 1194,1198 (CA2 1978)

[126] *United States v Smith*, 869 F. 2d 348, 353–354 (CA7 1989)

[127] Both appeal rulings related to the legal question of whether it was appropriate for the lower courts to take these factors into consideration and whether they had properly taken them into consideration. Absent any abuse of discretion by the lower courts, the appeal rulings did not question the conclusions that the lower courts reached on the basis of consideration of these factors. Both the lower courts had found auditory-spectrographic testimony admissible.

[128] *Daubert*, 509 U.S. at 594.

[129] In both *Williams* and *Smith* the courts had considered published research related to the validity of forensic voice comparison using spectrographic approaches. One publication considered in both cases was Oscar Tosi, Herbert Oyer, William Lashbrook, Charles Pedrey, Julie Nicol, & Ernest Nash, *Experiment on voice identification*, 51 Journal of the Acoustical Society of America 2030–2043 (1972) http://dx.doi.org/10.1121/1.1913064 (hereinafter Tosi *et al*. 1972), which reported on two studies which tested the performance of students and practitioners using the spectrographic and auditory-spectrographic approaches respectively. The first was the most extensive empirical test of spectrographic or auditory-spectrographic approaches ever conducted. Immediately after its publication, however, Tosi *et al*. 1972 was criticized by Bolt *et al*. 1973 note 5 *supra*, who argued that the first study was methodologically flawed. The first study was conducted under laboratory conditions (high quality audio recordings, no background noise, no transmission through communication channels, *etc*.) and not under forensically realistic conditions. Among other criticisms of its methodology, Bolt *et al*. 1973 therefore argued that the results of the first study were not informative as to how implementations of the approach would perform under casework conditions. The second study in Tosi *et al*. 1972 was a review of actual casework, comparing the conclusion of each forensic voice comparison analysis with the verdict or plea accepted in the case. This has been criticized on the grounds that validation requires the tester to know the truth, and the outcome of a legal case cannot be substituted for the truth, especially when the object being tested may have contributed to the outcome of the case – if a forensic scientist testifies that the defendant is the speaker on the questioned-speaker recording, and on the basis of the forensic scientist's testimony the trier of fact finds the defendant guilty, the forensic scientist cannot legitimately cite the verdict as evidence that their testimony was correct – the argument is circular. For additional criticisms of Tosi *et al*. 1972 see NRC 1979 note 6 *supra*, Gruber & Poza 1995 note 8 *supra*, and Morrison 2014 note 8 *supra*.

[130] That literature included the aforementioned Tosi *et al*. 1972 note 129 *supra*, Bolt *et al*. 1973 note 5 *supra*, and NRC 1979 note 6 *supra*

> reliability and admissibility of voice spectrograph analysis.[131]
>
> The potential rate of error of the aural spectrographic method is unknown and may vary considerably, depending on the conditions of the particular application.[132]
>
> [The expert's] testimony is unreliable under Rule 702. He is applying a technique that, in general, lacks the reliability necessary for admission under Rule 702. His application of the technique was flawed ... [His] testimony does not meet the standards necessary for admission. It is properly excluded as unhelpful and confusing to the jury.[133]

As previously mentioned, since *Angleton* there are no reported cases in which testimony based on the spectrographic approach has overcome a *Daubert* challenge.

6.3 Peer review and publication

Daubert also states that "Another pertinent consideration is whether the theory or technique has been subjected to peer review and publication."[134] It goes on to state that "submission to the scrutiny of the scientific community is a component of 'good science,' in part because it increases the likelihood that substantive flaws in methodology will be detected."[135] But it warns that peer-reviewed publication is not necessarily an indication of scientific validity. "The fact of publication (or lack thereof) in a peer-reviewed journal thus will be a relevant, though not dispositive, consideration in assessing the scientific validity of a particular technique or methodology on which an opinion is premised."[136]

The NCFS has stated that what counts as foundational scientific literature supportive of forensic practice must have been published in peer-reviewed archival venues, for example, articles published in respected scientific journals, *i.e.*, "a journal that utilizes rigorous peer review with independent external reviewers to validate the accuracy in its publications and their overall consistency with scientific norms of practice".[137] The NCFS further stated that non-peer reviewed publications and ephemera such as conference presentations do not count for this purpose.

We are probably more pessimistic than the *Daubert* ruling as to the quality of many papers that are

---

131 *Angleton*, 269 F.Supp.2d at 905.

132 *Ibid*. at 902.

133 *Ibid*. at 904.

134 *Daubert*, 509 U.S. at 593.

135 *Ibid*.

136 *Ibid*. at 594.

137 National Commission on Forensic Science, Scientific literature in support of forensic science and practice (2015) http://www.justice.gov/ncfs/file/786591/download (hereinafter NCFS 2015 *literature*)

accepted for publication after peer review. Courts should be aware that the quality of peer-reviewed publications may vary from subject area to subject area. We believe that a substantial proportion of papers published in forensic science in general and forensic voice comparison in particular suffer from major methodological flaws, including the use of databases which are very small and which do not represent forensically realistic conditions. There is also nothing to prevent a group of supporters of a particular approach from forming an association and sponsoring their own journal in which they peer review each other's papers and exclude dissenting voices.[138] Even the quality of review in respected peer-reviewed journals can fail – there is a high element of chance due to the difficulty of finding reviewers who are qualified, who have time, and who are willing to review papers on a volunteer basis. The NCFS has released a document which provides a list of criteria for assessing forensic science research literature.[139] The list is actually a list of things which should all be part of the peer review process,[140] but the NCFS recommends that the criteria be applied in assessing literature which has already been published in peer reviewed journals.

Given all these problems, we would recommend that courts considering admissibility not be overly impressed by the mere existence of a peer-reviewed paper supporting a particular technique, unless the judge is able to obtain a competent independent assessment of the scientific quality of that paper (or they are able to perform their own assessment of scientific quality). Any such independent assessment should also consider the extent to which the results of a published paper are actually applicable to the conditions of the particular case under investigation.

As previously mentioned, in *Angleton* the court considered research literature on auditory-spectrographic approaches covering a period of over 30 years. The court concluded that:

> Although aspects of the voice spectrographic method have been subject to review in published studies, many of the studies conclude that voice spectrographic analysis is of questionable scientific validity as a method of identifying an unknown speaker.[141]

[138] The quality of the review process for peer-reviewed conference proceedings, as opposed to peer-reviewed journal articles, is often particularly poor, each reviewer often being asked to review up to 10 submissions in a short amount of time. Another problem is the rise and proliferation of so-called predatory journals, journals which have the trappings of peer reviewed journals but which will publish essentially anything if the authors are willing to pay: Jeffrey Beall, Predatory publishing is just one of the consequences of gold open access, 26 Learned Publishing 79–84 (2013) http://dx.doi.org/10.1087/20130203; John Bohannon, *Who's afraid of peer review?* 342(6154) Science 60–65 (2013, Oct 4) http://dx.doi.org/10.1126/science.342.6154.60. It may not be immediately obvious whether a cited paper is a genuine peer reviewed paper published in a respected journal, or whether it was published in a predatory journal.

[139] National Commission on Forensic, Establishing the foundational literature within the forensic science disciplines (2015) http://www.justice.gov/ncfs/file/795096/download

[140] Although we think that requiring strict adherence to every point would be overzealous.

[141] *Angleton*, 269 F.Supp 2d at 899.

6.4 Standards

The *Daubert* ruling also states that "in the case of a particular scientific technique, the court ordinarily should consider ... the existence and maintenance of standards controlling the technique's operation."[142]

National and International Standards[143] are published by Standards organizations. They have procedures for developing Standards which include an opportunity for public comment on drafts, through which stakeholders can provide feedback.[144] A laboratory that wishes to be accredited has to demonstrate that it follows one or more relevant National or International Standards.[145] Clients may require a laboratory to be accredited to a particular Standard before contracting services from that laboratory. The FSR in England & Wales, and the NRC and NCFS in the U.S. have respectively mandated and recommended that forensic science providers be accredited.[146] For many forensic laboratories this is accreditation to International Standard ISO/IEC 17025.[147] This Standard requires laboratories to develop written policies and procedures and produce documentation which demonstrates that they follow those policies and procedures. The Standard defines what the policies and procedures should cover, but the laboratory has substantial discretion as to the details. The Standard includes a requirement to document validation of implementations of methods used by the laboratory (empirically demonstrate the validity and reliability of systems based on approaches). Accreditation bodies will inspect the laboratory and award accreditation if the laboratory has demonstrated that it is in compliance with the Standard. It should be remembered that being accredited and following a Standard is no more than a guarantee that a Standard has been followed and that there is documentation to show that it has been followed, but it does not guarantee that the results of an analysis will necessarily be correct, especially if the Standard or the validation procedures are not actually fit for purpose.

The *Daubert* ruling does not, however, define what the Supreme Court meant by the term *standard*, and in interpreting *Daubert*, it is clear that courts consider the term *standard* to be much broader than National and International Standards. Courts appear to accept practically any so-called standard developed by just

---

[142] *Daubert*, 509 U.S. at 594.

[143] We use initial capitalization to distinguish an official *Standard* published by a National or International Standards organization from anything else which may in common usage be called a *standard*.

[144] The procedure is not perfect in every instance. The public comment period may be short and there may be a failure to publicize it well to potential stakeholders. The committee charged with drafting the standard has to consider the comments but does not necessarily have to substantially revise the document, even in the face of major criticism.

[145] Guidelines may also be issued by National and International Standards organizations. A Guideline differs from a Standard in that documenting compliance with a Standard is essential for accreditation, but a Guideline constitutes advice which the laboratory can choose to follow or not. Compliance with Guidelines does not form part of the accreditation process.

[146] FSR 2014 note 73 *supra*, NRC 2009 note 72 *supra*, NCFS 2015 *universal* note 74 *supra*

[147] International Standard Organization / International Electrotechnical Commission, ISO/IEC 17025 General requirements for the competence of testing and calibration laboratories (2nd ed. 2005-05-01). The Standard is actually designed for testing and calibration laboratories rather than forensic laboratories, and may therefore not be ideal.

about any organization without necessarily going through the procedures which would be needed to develop a National or International Standard. The standards that have been mentioned in court rulings include standards developed by the International Association of Voice Identification (hereinafter IAVI),[148] the International Association for Identification (hereinafter IAI),[149] and the American Board of Recorded Evidence (hereinafter ABRE).[150,151,152] These groups were, however, all formed by practitioners of the auditory-spectrographic approach and the standards were written by practitioners of the auditory-spectrographic approach. The existence of standards should not be taken as conveying any credibility on claims made by the supporters of an approach when those standards simply assume that the approach is valid and reliable rather than requiring demonstration of degree of validity and reliability if implementations of the approach under casework conditions. How would a neutral observer choose between mutually contradictory standards or position statements issued by rival associations when what each has to say amounts to no more than *ipse dixit*?

the IAI "does not support or approve the use of any other voice identification technique [other

[148] IAVI 1979 note 28 *supra*

[149] IAI 1991 note 28 *supra*

[150] ABRE 1999 note 28 *supra*

[151] For example in: *United States v Williams*, 583 F. 2d 1194,1198 (CA2 1978); *United States v Smith*, 869 F. 2d 348, 353–354 (CA7 1989); *State v. Coon*, 974 P.2d 386 (Alaska 1999); *United States v Robert N Angleton*, 2003, 269 F Supp 2nd 892 S D TX; *State of Vermont v Gregory S Forty*, 2009 VT 118.

Under *Daubert*, the lower court in *Coon* found that the auditory-spectrographic approach had been tested and had a low error rate when the IAI standards (IAI 1991note 28 *supra*) were followed, and that the practitioner in that case had followed those standards. The appeal court found that the lower court had not abused its discretion.

The court in *Angleton* found that "The IAI, ..., has ceased certifying aural spectrographic examiners" 269 F.Supp.2d at 902, and that the forensic practitioner in that case failed to follow all requirements of the IAI and ABRE standards (note 28 *supra*).

In *Forty* the defense sought to have forensic voice comparison testimony based on an auditory-spectrographic approach admitted. The lower court held an admissibility hearing in which the defense expert testified. The prosecution did not call a forensic expert of its own and did not challenge the defense expert with respect to his claims regarding error rates. Instead the prosecution attacked the testimony on the grounds that the forensic practitioner had not followed the ABRE standards. The ABRE standards required at least 10 words to be examined, but the recording of the offender was short and the forensic practitioner had only examined 8 words. Citing previous rulings, the lower court ruled that the testimony failed to satisfy the criteria of *Daubert* and Vermont Rule 702, and was thus inadmissible. The appeal court sympathized with the lower court's predicament of having to rule in a situation in which the prosecution failed to present adequate arguments related to the evidence proffered by the defense; however, it found that the lower court was in error in relying on rulings made by other courts rather than on the evidence and arguments presented to it by the parties. Ultimately, however, the appeal court upheld the lower court's decision to exclude the testimony on the grounds that the ABRE standards had not been followed. From a scientific perspective we find absurd the implication that the addition of two extra words would have magically rendered the auditory-spectrographic testimony acceptable. The appeal court in *Forty* appears to have found a scientifically indefensible but legally correct way to uphold the lower court's decision that the testimony was inadmissible. Arguably it did the right thing for the wrong reason.

[152] The IAVI was established as an independent association in 1971 and became part of the IAI in 1980 (see Gruber & Poza 1995 note 8 *supra* at §123). ABRE was a group which subsequently broke away from the IAI (see *Angleton* at note 3).

> than the auditory-spectrographic techniques] listed within these standards."[153]
>
> IAFPA [International Association for Forensic Phonetics and Acoustics] dissociates itself from the approach to forensic speech comparison known as the "voiceprint" or "voicegram" method ... The Association considers this approach to be without scientific foundation, and it should not be used in forensic casework.[154]

We would caution that the existence of standards, and that a practitioner follows those standards, is no guarantee of the validity and reliability of the results.[155]

The European Network of Forensic Science Institutes (ENFSI) recently published *Methodological guidelines for best practice in forensic semiautomatic and automatic speaker recognition, including guidance on the conduct of proficiency testing and collaborative exercises*.[156] This document did not go through the process to count as a National or International Standard or Guideline, and it explicitly only covers what we have called acoustic-phonetic-statistical and automatic approaches. In contrast to the IAI, ABRE, and IAFPA documents, however, the ENFSI document recommends the use of the likelihood ratio framework and empirical testing of the validity of forensic analysis systems.

With respect to frameworks for evaluation of forensic evidence in general, a number of documents recently issued by national and international organizations recommend the use of the likelihood ratio framework (none of these count as National or International Standards or Guidelines):[157]

- Association of Forensic Science Providers (AFSP, United Kingdom and Republic of Ireland) *Standards for the formulation of evaluative forensic science expert opinion*[158]

---

[153] IAI 1991 note 28 *supra*

[154] International Association for Forensic Phonetics and Acoustics, Resolution on voiceprints July 24 2007 http://www.iafpa.net/voiceprintsres.htm (last visited Oct 27, 2016). IAFPA was formed in 1991. It was formed by and primarily consists of practitioners of auditory-acoustic-phonetic approaches.

[155] Standards which at first may seem to be rational can miss the mark. For example, the ABRE standards (ABRE 1999 note 28 *supra*) includes a section on the preparation of spectrograms which includes instructions as to technical settings to be used when making spectrograms. One may be willing to accept that following these instructions will produce better quality spectrograms, but be skeptical about the degree of validity and reliability of the spectrographic approach as a whole. As previously mentioned, it is not enough to validate components of a system, one has to validate the performance of the system as a whole. One may also decide that the ABRE standards now refer to obsolete technology. They refer to analogue audio recordings on magnetic tape and the use of specialized hardware for the generation of spectrograms.

[156] Andrzej Drygajlo, Michael Jessen, Stefan Gfroerer, Isolde Wagner, Jos Vermeulen and Tuija Niemi, Methodological guidelines for best practice in forensic semiautomatic and automatic speaker recognition, including guidance on the conduct of proficiency testing and collaborative exercises (2015) http://www.enfsi.eu/sites/default/files/documents/guidelines_fasr_and_fsasr_0.pdf (hereinafter ENFSI 2015 *speech*)

[157] The PCAST report (note 19 *supra*) also essentially advocates the use of a likelihood ratio framework (see §3.6 *supra*, and Morrison, Kaye, *et al*. 2016 note 57 *supra*). The desirability of using the likelihood ratio framework was also emphasized in NIST/NIJ 2012 note 92 *supra*.

[158] AFSP 2009 note 71 *supra*

- ENFSI *Guideline for evaluative reporting in forensic science*[159]
- NCFS *Views on statistical statements in forensic testimony*[160]

6.5 General acceptance

Frye states:

> 'The rule is that the opinions of experts or skilled witnesses are admissible in evidence in those cases in which the matter of inquiry is such that inexperienced persons are unlikely to prove capable of forming a correct judgment upon it, for the reason that the subject matter so far partakes of a science, art, or trade as to require a previous habit or experience or study in it, in order to acquire a knowledge of it. When the question involved does not lie within the range of common experience or common knowledge, but requires special experience or special knowledge, then the opinions of witnesses skilled in that particular science, art, or trade to which the question relates are admissible in evidence.' ... Just when a scientific principle or discovery crosses the line between the experimental and demonstrable stages is difficult to define. Somewhere in this twilight zone the evidential force of the principle must be recognized, and while courts will go a long way in admitting expert testimony deduced from a well-recognized scientific principle or discovery, *the thing from which the deduction is made must be sufficiently established to have gained general acceptance in the particular field in which it belongs*. (emphasis added)

Daubert rejected the idea that general acceptance should be the only relevant factor for determining admissibility, but stated that:

> "general acceptance" can yet have a bearing on the inquiry. A "reliability assessment does not require, although it does permit, explicit identification of a relevant scientific community and an express determination of a particular degree of acceptance within that community." ... Widespread acceptance can be an important factor in ruling particular evidence admissible, and "a known technique that has been able to attract only minimal support within the community," ... , may properly be viewed with skepticism.[161]

---

[159] Sheila M. Willis, Louise McKenna, L., Sean McDermott, Geraldine O'Donell, Aurélie Barrett, Birgitta Rasmusson, Anders Nordgaard, Charles E.H. Berger, Marjan J. Sjerps, José Juan Lucena-Molina, Grzegorz Zadora, Colin G.G. Aitken, Luan Lunt, Christophe Champod, Alex Biedermann, Tasha N. Hicks, Franco Taroni, ENFSI guideline for evaluative reporting in forensic science (2015) http://enfsi.eu/sites/default/files/documents/external_publications/m1_guideline.pdf

[160] National Commission on Forensic Science, *Views on statistical statements in forensic testimony (initial draft)* (2016) https://www.justice.gov/ncfs/file/888601/download

[161] *Daubert*, 509 U.S. at 594.

Ultimately, however, the Court in *Daubert* concluded that "'general acceptance' is not a necessary precondition to the admissibility of scientific evidence under the Federal Rules of Evidence".[162]

The outcome of a judicial inquiry into "general acceptance" often depends on the judge's determinations of *what* must be accepted and by *whom.*

What exactly is "*the thing* from which the deduction is made" in the context of forensic voice comparison? Is it the approach, the framework, the system as applied in a particular case, or, as we would argue, all of these? The assessment of "general acceptance" may well depend on the answer to this question.

In the context of forensic voice comparison, *who* must generally accept this *thing*? Is it practitioners of a particular approach, researchers who publish specifically on forensic voice comparison, researchers in the broader scientific community? The assessment of "general acceptance" will also depend on the answer to this question.

If the goal of the inquiry is to ensure the trustworthiness of forensic voice comparison evidence, then we suggest courts look to a relatively broad scientific community. Spectrographic analysis is undoubtedly accepted among the community of spectrographic analysts (just as astrology is generally accepted among astrologers and phrenology among phrenologists), but history suggest that acceptance of a particular approach among those who make their living using or promoting that approach provides little assurance that it is trustworthy.

In *Coon,*[163] when considering under *Daubert* the admissibility of testimony based on the auditory-spectrographic approach, the appeal court noted that the lower court described the relevant scientific community as "forensic scientists and scientists in acoustics and speech-related fields with experience using the technique". The lower court therefore defined the relevant scientific community narrowly, effectively excluding all critics of the approach other than potentially a few former practitioners who had subsequently changed their opinion with respect to it efficacy. The critics would have no doubt disputed the error rate claims of the proponents. The appeal court in *Coon* noted that in *Gortarez*[164] the relevant scientific community had been defined more widely as "disinterested and impartial experts in many fields, possibly including acoustical engineering, acoustics, communications electronics, linguistics, phonetics, physics, and speech communications", the latter list being non-exclusive.[165] The appeal court in *Coon* found that it was not clear whether the auditory-spectrographic approach was generally accepted within the relevant scientific community, but did not conclude that the lower court had abused its discretion in

[162] *Daubert*, 509 U.S. at 597.

[163] *State v. Coon*, 974 P.2d 386 (Alaska 1999)

[164] *State v Gortarez*, 141 Ariz. 254, 686 P.2d 1224, 1233 (1984)

[165] The appeal court in *Gortarez* reviewed a large number of research publications (43 were listed) and found that the auditory-spectrographic approach was not generally accepted by the relevant scientific community and hence inadmissible under *Frye*.

choosing a narrow definition for the relevant scientific community or in finding that the approach was generally accepted. Faigman *et al*. (2015)[166] note that, under *Frye*, when courts have adopted a broad definition of the relevant scientific community they have unanimously found the auditory-spectrographic approach to be inadmissible, whereas when they have adopted a narrow definition they have unanimously found it admissible. In its implementation, the general acceptance criterion has therefore been about choosing a relevant scientific community rather than determining whether an approach is generally accepted within the relevant scientific community.

For U.S. courts the general acceptability of the spectrographic approach appears to have waned,[167] but is there any evidence that there is currently a generally accepted approach to forensic voice comparison? There are at least two relatively recent surveys of approaches used by practitioners, and at least one relatively recent review of the research literature. We discuss these in Appendix D.

Ultimately, however, we think that general acceptance is a very poor indicator of scientific validity, especially during a period in which a paradigm shift[168] is underway, as is now the case in forensic science in general[169] and forensic voice comparison in particular.[170] One should not prefer an earth centered model of the universe, or prefer an approach to forensic voice comparison based on subjective judgment, because it is the preference of the majority of scientists and/or practitioners, one should prefer the model or approach which shows the greatest promise or which is ultimately demonstrated to make the most

[166] note 2 *supra*

[167] Although the appeal court in *U.S. v. Drones*, 218 F.3d 496, 503 (5th Cir.2000) did not actually rule on the issue of admissibility, testimony called by both parties indicated that by the year 2000 general acceptance of the auditory-spectrographic approach had waned. During a *habeas corpus* hearing an expert witness called by the defense "testified that there were published recommended procedures for conducting [auditory-spectrographic] voice identification examinations, ... [but] that there was no set of objective criteria against which to check the accuracy of a particular expert's analysis and that voice identification analysis was largely subjective" (*Drones* at 12) and that "While ... expert voice identification testimony has been used extensively in state and federal courts over the past thirty years, he also testified that he did not know if spectrographic evidence was widely accepted by the relevant scientific community." (*Drones* at 28). An expert witness called by the prosecution "testified that very little research has been done in the area of 'courtroom application of spectrographic voice identification,' largely because since the 1970's, many researchers have felt that spectrographic comparison could not produce reliable results. He stated that 'almost nobody' in the relevant scientific community uses spectrographic voice identification because there is no theoretical basis for the proposition that an individual's voice is truly unique and identifiable." (*Drones* at 16). "He further stated that [the lack of theoretical basis] has resulted in a precipitous drop in the number of expert practitioners over the past few decades, from fifty to sixty practitioners in the 1970's to roughly a dozen experts at the time of Drones's trial." (*Drones* at 28).

[168] Thomas S. Kuhn, The Copernican Revolution: Planetary Astronomy in the Development of Western Though (1957); Kuhn 1970 note 3 *supra*

[169] Michael J. Saks & Jonathan J. Koehler, *The coming paradigm shift in forensic identification science*, 309 Science 892–895 (2005) http://dx.doi.org/10.1126/science.1111565

[170] Morrison 2009 note 29 *supra*

valid and reliable predictions / strength of evidence statements.[171,172]

6.6 Conclusion with respect to admissibility

Concluding remarks of Section II C of the *Daubert* ruling include the following statements:

> The inquiry envisioned by Rule 702 is, we emphasize, a flexible one. Its overarching subject is the scientific validity—and thus the evidentiary relevance and reliability—of the principles that underlie a proposed submission. The focus, of course, must be solely on principles and methodology, not on the conclusions that they generate.[173]

Given all the above, we believe that the *Daubert* criterion which should be given greatest weight is that requiring empirical demonstration of degree of validity and reliability. Indeed, we believe that this should be both a necessary and a sufficient criterion. We believe that the other criteria (peer-reviewed publication, standards, and general acceptance) constitute secondary proxies. Although a degree of correlation with the primary criterion may be expected, none of the other criteria should be considered either necessary or sufficient, either individually or in combination.[174]

7 The forensic voice comparison testimony in *US v Ahmed*

We review and critique the testimony in *Ahmed*[175] in some detail. The reason for this is that many of the problems in this testimony are concrete examples of the sorts of problems that we expect to potentially reoccur in future attempts to have forensic voice comparison testimony admitted under *Daubert*. Understanding the specific problems in this case will therefore potentially assist forensic practitioners to

[171] According to Kuhn, scientific revolutions are precipitated by crises, such as a major problem which has repeatedly defied solution within the old paradigm. The new paradigm must at least offer the promise of a solution, but the realization of that promise need not be immediate. It actually took approximately 80 years before the empirical advantages of the Copernican revolution were realized. It took Kepler's mathematical models applied to Brahe's observations to produce more accurate predictions of planetary motion than had ever been possible before. It would not, however, have been possible without a shift from an earth-centered to a sun-centered paradigm. Even with an empirically demonstrated better solution to the problem, there was still concerted opposition to the Copernican paradigm, and it took at least another half century before it became universally accepted. Polling the scientific community during that period would not have given us the answer which later was universally accept as correct (although the paradigm has since shifted at least once again with Einstein's theories of relativity).

[172] If a subjective approach to forensic voice comparison were empirically demonstrated to have better validity and reliability under casework conditions than approaches based on relevant data, quantitative measurements, and statistical models, then we would prefer that subjective approach, but we are not aware of any such comparative tests having been performed and having produced such results.

[173] *Daubert*, 509 U.S. at 594.

[174] See also Jonathan J. Koehler, *Forensics or fauxrensics? Ascertaining accuracy in the forensic sciences*, 49 Arizona State Law Journal (2017)

[175] *U.S. v. Ali Ahmed, Madhi Hashi, & Muhamed Yusuf*, 2015 EDNY 12-CR-661(SLT)

avoid making the same types of mistakes in the future and/or assist lawyers and judges to identify, understand, and deal with the occurrence of these types of problems in the future.

7.1 Summary of the testimony

In *Ahmed* a forensic practitioner compared five recordings known to be of defendant *Yusuf* with three questioned-speaker recordings associated with terrorist activity. Four of the known-speaker recordings were of intercepted mobile telephone calls and one was of a landline telephone call. The questioned recordings consisted of one video recording of a man speaking with a bandana over his mouth and two intercepted mobile telephone calls. The quality of the questioned-speaker recordings was poor.[176] In the known-speaker recordings and in two of the three questioned-speaker recordings, the speaker spoke a mixture of Swedish, Somali, and Arabic. The other questioned-speaker recording contained no Swedish.

As described below, the practitioner compared the known- and questioned-speaker recordings[177] using four different approaches: auditory, acoustic-phonetic non-statistical, acoustic-phonetic statistical, and automatic. He then combined the results of these four analyses to reach a final conclusion as to the strength of the evidence.

---

[176] The video recording had been retrieved from the internet and showed signs of lossy compression. Compression is a procedure which reduces the size of files so that less space is taken up on storage devices and so that they can be transmitted faster or more files can be transmitted in the same time. Many compression algorithms are *lossy* in that they result in the loss of some acoustic information and some distortion of the remaining acoustic information. Additionally, there were some transient background noises on the recording (including gunshots), which the practitioner manually edited out. One of the mobile recordings contained "highly degrading electrical current noise" Report §1.3.1, which was probably *electrical hum* from an alternating current electrical supply, which runs at either 50 or 60 Hz depending on the part of the world. The other mobile recording contained "disruptive electrical pulses" Report §1.3.1 which occurred once every half second, and it also had "a very low transmission bit rate" Report §1.3.1. Mobile telephone systems use lossy compression so that less data are transmitted. The amount of compression varies depending on the demand put on the mobile telephone system, greater compression corresponds to a lower "transmission bit rate" and greater loss and distortion of the acoustic information in the signal. This leads to poorer performance from forensic voice comparison systems compared to if the recordings are landline telephone recordings (which in turn leads to poorer performance than if the recordings are direct microphone recordings); see for example Cuiling Zhang, Geoffrey Stewart Morrison, Ewald Enzinger & Felipe E. Ochoa, *Effects of telephone transmission on the performance of formant-trajectory-based forensic voice comparison – female voices*, 55 Speech Communication 796–813 (2013) http://dx.doi.org/10.1016/j.specom.2013.01.011.

[177] In his analyses, the practitioner grouped the known-speaker recordings together and compared these as a group with each individual questioned-speaker recording:

Comparison 1: known-speaker recordings versus the *video recording* (54 seconds net speech, lossy compression)

Comparison 2: known-speaker recordings versus the *longer mobile telephone recording* (82 seconds net speech, electrical pulses, very low transmission bit rate)

Comparison 3: known-speaker recordings versus the *shorter mobile telephone recording* (35 seconds net speech, electrical hum, no Swedish)

The material in parenthesis describes the conditions of the questioned-speaker recording in each comparison.

7.1.1 Relevant population

The forensic practitioner stated that he regarded "a reasonable reference population in this case to be young male Somali and Swedish speakers of the Stockholm area with a fluent ability in Swedish."[178] He stated that the known speaker spoke Swedish with a Stockholm accent. He did not say anything about the accent of the voices on the questioned-speaker recordings, but we assume that they also spoke Swedish with Somali and Stockholm accents.

7.1.2 Auditory analysis

The forensic practitioner noted auditory perceptual similarities between the pronunciation of particular Swedish vowel and consonant sounds, and particular words in the questioned-speaker recordings and in the known-speaker recordings. He also noted similar use of filler words[179] and of a grammatically incorrect phrase. He also noted similarities in tempo and intonation, and that the voices were somewhat nasal and had a raised laryngeal setting.[180] He also noted that the speakers mixed languages (a phenomenon known as *code switching*), but also noted that this was relatively common for the relevant population.

When asked about the conclusions drawn from his auditory approach, the practitioner stated: "The important thing is that, you know, it's not assessed separately every single thing. It's more of a wholistic picture in the end".[181]

7.1.3 Acoustic-phonetic non-statistical analysis

The forensic practitioner measured several acoustic-phonetic properties of the speech in the recordings. He compared measurements of fundamental frequency and of articulation rate (a measure of how fast the speaker is speaking) for the Swedish-language portions of the recordings. He found that both were within the normal range for Swedish speakers.

---

[178] Report §3.1

[179] In English filler words include *um*, *ah*, and *like*.

[180] The practitioner did not mention that Somali is a language which includes speech sounds made with a constricted pharynx; Jerold A. Edmondson, Cécile M. Padayodi, Zeki Majeed Hassan, & John H. Esling, *The laryngeal articulator: Source and resonator*, Proceedings of the 16th International Congress of Phonetic Sciences 2065–2068 (2007). This could potentially influence a bilingual Somali-Swedish speaker's Swedish pronunciation and could be perceived as a raised laryngeal setting.

[181] Transcript day 1, page 104, lines 3–5.

7.1.4 Acoustic-phonetic statistical analysis[182]

The forensic practitioner also compared long-term formant measurements, which are measurements taken over all instances of all the vowel sounds in the whole of the recording (and of the subset of consonant sounds for which formats can also be measured). In addition to comparing the questioned-speaker recordings with the known-speaker recordings, the practitioner also compared them with a set of 500 recordings of other speakers from a database. He described these as having "similar acoustics to [the recordings] in the case".[183] No details were provided as to how exactly these recordings reflected the conditions of the known-speaker or questioned-speaker recordings, or what population the speakers represented.

The statistical analysis did not involve calculation of likelihood ratios. Instead, it involved calculating scores and then making a subjective judgment based on the value of the questioned-speaker versus known-speaker score compared to the database-speakers versus known-speaker scores.[184]

7.1.5 Automatic analysis

The forensic practitioner also made use of a commercially marketed forensic voice comparison system: Batvox version 4.1,[185] produced by the company Agnitio.[186] The measurements that Batvox makes are MFCCs (see Section 2.3.4). The statistical modeling technique used by Batvox is based on *i-vectors*, which is a common approach in automatic speaker recognition systems. Batvox first produces a score for the comparison of a questioned-speaker recording and the known-speaker recordings, then converts that score to a likelihood ratio. Again the general approach is common in automatic speaker recognition systems, although some of the details of the particular implementation may be peculiar to Batvox. To calculate scores, Batvox uses a statistical model trained on data, but those data are diverse and not representative of the particular relevant population or particular conditions of the case. Instead, Batvox attempts to take account of the relevant population and conditions of the particular case during a

[182] Although we have listed this approach as an acoustic-phonetic statistical approach, it could instead be considered an automatic approach, but using a type of measurement which is traditional in acoustic phonetics rather than a type of measurement which is traditional in speech processing.

[183] Report §3.3.3

[184] The procedure was to calculate a score for the comparison of a questioned-speaker recording with the set known speaker recordings, and also calculate scores for the comparison of each of the 500 database speakers with the set of known speaker recordings. No details were supplied with respect to the algorithm which was used to calculate the scores. All the scores were ranked, and if a questioned-speaker score was ranked the highest the practitioner made a subjective judgment as to the strength of the evidence based on the magnitude of the questioned-speaker score compared to the 500 database-speaker scores.

[185] For a more technical but still brief description of this system, see David van der Vloed, *Evaluation of Batvox 4.1 under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01)*, Speech Communication (2016) http://dx.doi.org/10.1016/j.specom.2016.10.001 (hereinafter van der Vloed 2016).

[186] In November 2016 Agnitio was purchased by Nuance.

subsequent score to likelihood ratio conversion process. The user enters *reference population data* consisting of recordings of a number of speakers, and Batvox selects the recordings which it calculates to be most similar to the known speaker. The reference data which the practitioner entered consisted of approximately 6000 mobile telephone calls and 5000 landline telephone calls (the population represented by the speakers in these recordings appears to have been Swedish speakers in general).[187] The practitioner had Batvox select the 45 recordings which it calculated to be most similar to the suspect model. These 45 selected recording were then used in training a statistical model which was used to convert scores to likelihood ratios.

As is common in automatic speaker recognition systems, Batvox includes statistical techniques intended to compensate for recording condition mismatches. One of these techniques requires the user to provide Batvox with recordings which they believe reflect the recording conditions of the questioned-speaker recording. This set of recordings is known as an *imposter set*. These recordings are then used to train a statistical model intended to compensate for the mismatch in recording conditions between the known-speaker and the questioned-speaker recordings. Use of this mismatch compensation technique, and hence use of an imposter set, is optional.

For one of the comparisons (Comparison 1), the practitioner first performed an analysis without an imposter set, and then performed another analysis using an imposter set consisting of 85 recordings of "young male speakers"[188] recorded using lapel microphones. No additional information was provided regarding what population these speakers represented. For the other comparisons (Comparisons 2 and 3), no imposter set was used.

Batvox outputs numeric likelihood ratio values, but the practitioner did not report these values. Instead he used the output of the automatic system as input to a subjective judgment process and used verbal expressions to convey the strength of the evidence.[189]

---

[187] The practitioner stated that "reference population data used can come from dialect databases such as Swedia, forensic material and other in-house adapted material" (Report §2.1). The Swedia database is a database of recordings of speakers with different dialects of Swedish. The forensic practitioner estimated that there were between 10 and 20 Somali Swedish speakers in the recordings in the reference data which he entered into the automatic system – less than 1% of the total number of speakers (Transcript day 2, p 190).

[188] Transcript, day 1, page 95, lines 6–7.

[189] The practitioner offered the explanation that: "The outcome of all the tests made with any automatic system is treated as only one part in a full analysis. ... If the material in the case has been judged to be such that proper tests can be made with automatic systems, the scores or likelihood ratios are treated as an input to the analysts. It means that the experience of using a software is much more important than a score or likelihood ratio itself. For different material, different score spaces or likelihood ratio spans are expected due to the duration and or mismatch and quality between the material tested. I.e. for a mismatched test an analyst is better at judging the value of a score or a likelihood ratio than the machine itself and the analyst will together with the other results of the analysis judge where in a likelihood ratio span the outcome should be placed, i.e. on which level in the ordinal scale." (Report §2.1)

7.1.6 Combination of results and statements of conclusions

The forensic practitioner combined the results from all four approaches and expressed his conclusions on a nine-level scale. The nine-level scale (reproduced in Appendix E) was based on that used by the Swedish National Laboratory of Forensic Science, with some additions made by the practitioner.[190] The forensic practitioner did not express his final conclusions as numeric likelihood ratio values, he only provided the level numbers and verbal expressions from his nine-level scale.

No details were provided as to how the results were combined other than that they were "weighted". The report stated that: "For the [auditory] and [acoustic-]phonetic analyses a holistic likelihood ratio span is judged impressionistically and combined with the results from the automatic tests."[191] It also stated that: "The numbers representing the levels in the scale are only to a certain degree statistically based through calculation and to some extent a judgement of likelihood ratios."[192] When questioned during direct, the forensic practitioner stated that "the final conclusion is, of course, wholistic judgment based on all the tests made and the comparisons made in the whole examination".[193]

---

[190] Ordinal scales of this general type are popular in conjunction with both the likelihood ratio framework and other frameworks for the evaluations of evidence. See, for example: ABRE 1999 note 28 *supra*; IAI 1991 note 28 *supra*; Christophe Champod, Ian W. Evett, *Commentary on Broeders (1999) Some observations on the use of probability scales in forensic identification*, 7 Forensic Linguistics 238–243 (2000); AFSP 2009 note 71 *supra*; Sheila M. Willis, Louise McKenna, L., Sean McDermott, Geraldine O'Donell, Aurélie Barrett, Birgitta Rasmusson, Anders Nordgaard, Charles E.H. Berger, Marjan J. Sjerps, José Juan Lucena-Molina, Grzegorz Zadora, Colin G.G. Aitken, Luan Lunt, Christophe Champod, Alex Biedermann, Tasha N. Hicks, Franco Taroni, ENFSI guideline for evaluative reporting in forensic science (2015) http://enfsi.eu/sites/default/files/documents/external_publications/m1_guideline.pdf

[191] Report §4

[192] Note appended to the version of the nine-level scale provided in the forensic practitioner's report.

[193] Transcript, day 1, page 121, lines 22–24. The practitioner did not use the Level numbers to express the strength of evidence for each analysis and sub-analysis he performed. Instead, he used verbal expression such as "some support", "support", "strong support", "distinct similarities". Below we convert these to their corresponding Level values on the nine-level scale. We also report the Level values corresponding to the likelihood ratio values output by the automatic system. With one exception, the Level corresponding to the practitioner's verbal expression was higher than that corresponding to the numeric likelihood ratio output by the automatic system.

For Comparison 1, the forensic practitioner concluded that the strength of evidence was Level +3.

> Strengths of evidence reported for the auditory analyses correspond to Levels +2, +2, +3, +1, +3;
> for the acoustic-phonetic non-statistical analyses to Levels 0, 0;
> for the acoustic-phonetic statistical analysis to Level +2;
> and for the automatic analyses to Level +1 (LR = 35) or Level +2 ("support") when no imposter set was used, or to Level +2 (LR = 158) or Level +4 ("extremely strong support") when an imposter set was used.

For Comparison 2, the forensic practitioner concluded that the strength of evidence was Level +2.

> Strengths of evidence reported for the auditory analyses correspond to Levels +2, +1, +3, +2, +2, +1, +3;
> for the acoustic-phonetic non-statistical analyses to Levels 0, 0;
> for the acoustic-phonetic statistical analysis to Level +1;
> and for the automatic analysis to Level +1 (LR = 42) or +2 ("support").

For Comparison 3, the forensic practitioner concluded that the strength of evidence was Level 0.

> The strength of evidence reported for the automatic analysis corresponds to Level 0 (LR = 1/3.3, "no support"). The

7.2 Critique of testimony

How can we know whether the practitioner's conclusions in this case were trustworthy? How should a court evaluate the admissibility of such testimony under *Daubert*? In the present section we critique the testimony in light of Rule 702 and the *Daubert* criteria. Each section below (except the last) addresses a question which is asked using Rule 702, *Daubert*, or *Frye* terminology. The last question relates to contextual bias, an issue not explicitly identified in *Daubert* but now of increasing concern. The questions are:

- What methodology and reasoning were used?
- Was the testimony based on sufficient data and were the principles and methods reliability applied to the facts of the case?
- Has the technique been empirically tested and what is the known rate of error?
- Has the technique been subjected to peer review and publication?
- Are there standards controlling the technique's operation?
- Is the thing from which the deduction is made sufficiently established to have gained general acceptance in the particular field in which it belongs?
- Were reasonable steps taken to reduce the potential for contextual bias?

7.2.1 What methodology and reasoning were used?

The *Daubert* ruling instructs the trial judge to consider "whether the reasoning or methodology underlying the testimony is scientifically valid and ... whether that reasoning or methodology properly can be applied to the facts in issue."[194] A prerequisite to answering these questions is to understand what "methodology" and "reasoning" were used, *i.e.*, what approach and framework were used.

At first glance, it may have looked like the *Daubert* hearing was about the admissibility of an automatic approach to forensic voice comparison, but this was not the case. An automatic approach was used, but it was only one of multiple approaches employed by the forensic practitioner, and the final conclusion as to the strength of evidence depended little on the output of the automatic system.[195] The methodology

automatic analysis was the only one conducted for this comparison.

[194] *Daubert*, 509 U.S. at 592-3.

[195] With one exception, the forensic practitioner's subjective verbal expression of the strength of evidence for an automatic analysis corresponded to a Level value more favorable to the prosecution than the Level corresponding to the likelihood ratio

used by the forensic practitioner was a mixture of auditory, acoustic-phonetic non-statistical, acoustic-phonetic statistical, and automatic approaches. The output of the analysis based on each approach was either intrinsically a subjective judgment made by the practitioner, or a subjective judgment made by the practitioner based on the output of a statistical model – the output of a statistical model was not directly reported. The final conclusion as to the strength of the evidence resulting from the combination of all the approaches was also a subjective judgment made by the practitioner.[196] It is the trustworthiness of this combination of approaches which must be assessed in the context of a *Daubert* hearing. If some component parts were judged trustworthy, this would not suffice if they were combined with other components of undetermined trustworthiness, or if the procedure for combining them was of questionable trustworthiness.

Ostensibly the practitioner made use of the likelihood ratio framework. Although the report and oral testimony, and the scale of conclusions, included multiple deviations from a proper application of the

---

value output by Batvox. When the likelihood ratio value corresponded to Level +1, the verbal expression corresponded to +2, and when the likelihood ratio value corresponded to Level +2, the verbal expression corresponded to +4. The latter corresponding to a likelihood ratio value of 1 million or more, when the likelihood ratio value output by Batvox was only 158!

The difference between the actual likelihood ratio value output by the automatic system and the practitioner's verbal expression of the strength of evidence based on the output of the automatic system was due to the practitioner using his experience and also taking into consideration another analysis he had conducted. Transcript day 2, p 269, line 9ff, emphasis added:
Q: But you disagreed with the outcome that Batvox arrived at, didn't you?
A: [... for] Mismatched. Yes, yes.
Q: And you [dis]agreed because of your own *personal experience*?
A: *Yes*. On how it evaluates for these kind of mismatched conditions *in combination with a phonetic analysis*.
Q: And so again, this is a place where *we should just take your word for it* that your score is more representative of what really happened than the score of Batvox?
A: *Yes*.

The Advisory Committee's commentary on Rule 702 notes that "The trial court's gatekeeping function requires more than simply 'taking the expert's word for it.'"

[196] There are fusion procedures which use explicit weights calculated by statistical models trained on relevant data. See for example: Stéphane Pigeon, Pascal Druyts, and Patrick Verlinde, *Applying logistic regression to the fusion of the NIST'99 1-speaker submissions*, 10 Digital Signal Processing 237–248 (2000) http://dx.doi.org/10.1006/dspr.1999.0358; Joaquín González-Rodríguez, Philip J. Rose, Daniel Ramos, Doroteo T. Toledano, & Javier Ortega-García, *Emulating DNA - Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition*, 15 IEEE Transactions on Audio, Speech, and Language Processing 2104–2115 (2007) http://dx.doi.org/10.1109/TASL.2007.902747; Geoffrey Stewart Morrison, *Tutorial on logistic-regression calibration and fusion: Converting a score to a likelihood ratio*, 45 Australian Journal of Forensic Sciences 173–197 (2013) http://dx.doi.org/10.1080/00450618.2012.733025. Such procedures are transparent and replicable. Output from multiple systems may contain some overlapping (correlated) information, but the output of each system may also contain some information which is independent of (uncorrelated with) the information from the other systems. Statistical models can take account of correlation between the output of different systems and avoid the bias that would result from counting the same information multiple times (a potential problem that often goes under the name of *double counting*). Any improvement due to fusion via such a statistical model will be due to combining the independent (uncorrelated) information provided by the different systems. In contrast, the practitioner's "holistic judgment" based final conclusion was not transparent, and we have no guarantee that it did not count the same information multiple times. (What the practitioner reported as the result of his automatic analysis had already been influenced by the result of his phonetic analysis, see note 195 *supra*, and therefore even that did not constitute independent information.)

likelihood ratio framework, for the sake of brevity we do not discuss those here.[197]

7.2.2 Was the testimony based on sufficient data and were the principles and methods reliability applied to the facts of the case?

Rule 702 requires that "(b) the testimony is based on sufficient facts or data; (c) the testimony is the product of reliable principles and methods; and (d) the expert has reliably applied the principles and methods to the facts of the case." Given the discussion above, and as previously stated in Section 6, we believe that these conditions can be met if the forensic practitioner's calculation of the strength of evidence makes use of a sample of voice recordings which are representative of the relevant population and which reflect the conditions of the case under investigation, and if the forensic practitioner empirically demonstrates, under conditions reflecting those of the case, the degree of validity and reliability of their implementation of their approach to forensic voice comparison.

Let us accept the practitioner's proposal as to the appropriate relevant population: young adult male Somali speakers who are fluent in Swedish, but who have Somali accents in Swedish. It appears that the forensic practitioner did not actually have access to a sample which would be representative of the population he specified. When asked during cross if he had recorded Somali-Swedish speakers to use in his automatic analysis, he stated that he had not.[198] In his automatic analysis, the practitioner entered several thousand recordings, of which he estimated 10 to 20 (less than 1%) were of Somali-Swedish speakers. He had Batvox select 45 of those recordings to use as a sample of the relevant population. Even if Batvox included all the Somali-Swedish speakers, they would still have represented less than half the 45 used as the sample of the relevant population, the rest presumably being Swedish speakers without Somali accents (we have no information about which particular speakers' recordings were actually included). Even if the results of the statistical model had been directly reported, the output of the automatic analysis would not therefore have answered the question implied by what the practitioner stated as being the relevant population:

> What is the probability of obtaining the measured acoustic properties of the questioned-speaker recording (in which the speaker is a young male Somali-accented Swedish speaker) if it were produced by the defendant (who is a young male Somali-accented Swedish speaker)?
>
> versus
>
> What is the probability of obtaining the measured acoustic properties of the questioned-speaker recording (in which the speaker is a young male Somali-accented Swedish speaker) if it were

[197] Lack of consistency with the likelihood ratio framework was admitted by the practitioner at multiple point in his testimony. Transcript day 1, p 102, line 21; p 109, line 10; p 119 lines 19–20; day 2, p 234, line 10; p 236, lines 16ff.

[198] Transcript day 2, p 190, line 19ff

> produced by some other young male Somali-accented Swedish speaker?

Instead they would have been answering a question which would have been much closer to the following (fully the following if no Somali-accented speakers were included in the 45 selected by Batvox):

> What is the probability of obtaining the measured acoustic properties of the questioned-speaker recording (in which the speaker is a young male Somali-accented Swedish speaker) if it were produced by the defendant (who is a young male Somali-accented Swedish speaker)?
>
> versus
>
> What is the probability of obtaining the measured acoustic properties of the questioned-speaker recording (in which the speaker is a young male Somali-accented Swedish speaker) if it were produced by a Swedish speaker who does not have a Somali accent?

We contend that the latter is a nonsensical question, and hence (even allowing for the question effectively asked to be somewhere between the two above) that the data were not "sufficient" and that the practitioner did not "reliably appl[y] the principles and methods to the facts of the case."

7.2.3 Has the technique been empirically tested and what is the known rate of error?

The *Daubert* ruling states that "a key question to be answered in determining whether a theory or technique is scientific knowledge that will assist the trier of fact will be whether it can be (and has been) tested. ... '[T]he statements constituting a scientific explanation must be capable of empirical test'"[199] And that "in the case of a particular scientific technique, the court ordinarily should consider the known or potential rate of error". Combined with the requirement to consider "whether that reasoning or methodology properly can be applied to the facts in issue." we interpret this as implying that empirical tests of validity and reliability of the forensic analysis system must be conducted under conditions which reflect those of the case under investigation, and that the results must be reported. If the judge is first satisfied that the data used to test the system are sufficiently representative of the relevant population and sufficiently reflective of the conditions of the known-and questioned-speaker recordings in the case, the judge can then consider whether the demonstrated degree of validity and reliability is sufficient.

For the forensic practitioner's analysis in *Ahmed*, what needed to be tested was the entire system: the conglomerate of his auditory subsystem, his acoustic-phonetic non-statistical subsystem, his acoustic-phonetic statistical subsystem, his automatic subsystem, and his procedure for combining the results of these subsystems, and including the forensic practitioner himself as an integral part of the system. As we have previously discussed (Section 4), knowing the performance of a subsystem or a component of a subsystem would not be sufficient. The final strength of evidence conclusion is produced by the system

[199] *Daubert*, 509 U.S. at 593.

as a whole, and it is the performance of the system as a whole which needs to be considered by the judge.

The forensic practitioner did not provide any results of empirical testing of the performance of his system as a whole. Indeed, it does not seem that the practitioner's system as a whole has ever been empirically tested under any forensically realistic conditions. During cross we have the following exchange:

> Q Have you ever been tested to see what your accuracy rate is when you didn't know the answer in advance?
>
> A So the NFC [National Forensic Center] can provide blind tests for us whenever they want. And they don't have to tell us. I presume that it's not very often because it costs them money basically. But that's the only way. ...
>
> Q But have you ever been tested, that you know of, by the NFC and given the results in a blind test?
>
> A No.
>
> Q And so you rendered results in these 350 or 400 cases and in those cases like here you say we should rely on your expertise, right?
>
> A Yes.[200]

We therefore conclude that the forensic practitioner's method has not been tested, that the degree of validity and reliability of the implementation of his method has not been empirically demonstrated under conditions reflecting those of this case, and that the practitioner's method and its implementation would not therefore satisfy this *Daubert* criterion.

Although, as we have argued, demonstrating the performance of a subsystem would not be sufficient to satisfy the criterion, and the forensic practitioner's expression of the strength of evidence depended little on the output of his automatic system anyway, the practitioner and the prosecution argued for the scientific validity of the automatic system.[201] The practitioner's report referenced a number of papers which ostensibly tested Batvox. We could critique each in turn, but here we provide only an overview. In some papers the evaluations reported were not independent evaluations, but evaluations conducted by Agnitio employees or others linked to the company (although this does not itself invalidate the results, one should be aware that the evaluations were not conducted by an independent third party). In some papers it is not clear whether the system being tested was actually the commercial Batvox version 4.1

[200] Transcript day 2, p 266, line 2ff

[201] For brevity, we do not provide here a discussion of information in the report and transcript with respect to testing of the validity and reliability of the practitioner's auditory and acoustic-phonetic analyses, but we found no evidence that they had been empirically tested using data which we believe could be deemed sufficiently representative of the relevant population and sufficiently reflective of the speaking styles and recording conditions of the known- and questioned-speaker recordings in the case.

used by the practitioner, or a different Agnitio system optimized for the particular test. Descriptions in some papers suggest that the latter was the case. In other papers it is not clear whether Batvox, or any Agnitio system, was being tested at all. Some papers seem to be describing other systems, and in some papers if Batvox is used it is one of several anonymized systems and we do not known which results correspond to those from Batvox.

The relevant population in this case was young male Somali Swedish bilinguals with Somali-Stockholm accents in Swedish. The technical recording conditions in the case included using three mobile telephone recordings and one landline telephone recordings to train a known-speaker model, and questioned-speaker recordings that were from a video with lossy compression, from a mobile telephone call with electrical pulses and very low transmission bit rate, and from a short (35 seconds net speech) mobile telephone call with electrical hum, and there being a particular mixture of Swedish, Somali, and Arabic on the recordings. We do not believe that any of the cited papers reported tests of Batvox under conditions which could reasonably be deemed sufficiently representative of this population or sufficiently reflective of these conditions for the results to be considered informative as to the expected performance of the system in this case.

In a pre-hearing submission[202] the prosecution claimed that Batvox had been tested by independent organizations and academic institutions, and in particular described its performance in the 2012 Speaker Recognition Evaluation (SRE) run by the National Institute of Standards and Technology (NIST). The prosecution claimed that Batvox was ranked first or second in four out of five conditions and in the top ten overall.[203] We would contend that this is not relevant since the system submitted by Agnitio was a research system, not the commercial version of Batvox used by the forensic practitioner,[204] and the conditions tested in the NIST SRE did not represent the conditions of the forensic case under investigation. NIST explicitly states that SRE results should not be used to make decisions as to which system is best for a particular application,[205] and that the SRE not intended to be representative of

[202] Case 1:12-cr-00661-SLT-LB. Document 222

[203] The rules of the NIST SRE prohibit participants from making the sort of claims made by the prosecution: "Participants may not publish or otherwise disseminate their own comparisons of their performance results with those of other participants without the explicit written permission of each such participant. Furthermore, publicly claiming to 'win' the evaluation is strictly prohibited. Participants violating this rule will be excluded from future evaluations." National Institute of Standards and Technology, *The NIST year 2012 speaker recognition evaluation plan* (2012) http://www.nist.gov/itl/iad/mig/upload/NIST_SRE12_evalplan-v17-r1.pdf (hereinafter NIST 2012) at p. 5

[204] Personal communication 28 January 2016 from Dr Niko Brümmer, Chief Scientist, Agnitio.

[205] NIST includes the following disclaimer on its website (http://www.nist.gov/itl/iad/mig/sre12results.cfm):

> These results are not to be construed, or represented as endorsements of any participant's system or commercial product, or as official findings on the part of NIST or the U.S. Government. Note that the results submitted by developers of commercial SR products were generally from research systems, not commercially available products. ... The systems themselves were not independently evaluated by NIST.
>
> The data, protocols, and metrics employed in this evaluation were chosen to support SR research and should not be construed as indicating how well these systems would perform in applications. While changes in the data domain, or

forensic conditions.[206] When asked by the judge, the forensic practitioner claimed that the NIST SRE 2012 test material were "very similar to the audio in this case."[207] The practitioner did not, however, give a detailed explanation of how particular recordings conditions and particular populations represented in the NIST SRE data were similar to the particular conditions of the known- and questioned-speaker recordings in the case and the relevant population in the case. We consider the populations and conditions tested in the SRE to be very different from those in the *Ahmed* case.

7.2.4 Has the technique been subjected to peer review and publication?

Daubert also states that "Another pertinent consideration is whether the theory or technique has been subjected to peer review and publication."[208]

The forensic practitioner's report included 41 references. A substantial proportion of these, however, were not peer-reviewed articles published in archival venues. Nine (22%) were conference presentations which were not accepted in the basis of peer review of a paper. Two of these have associated non-peer-reviewed conference proceedings papers, and some of the rest have accessible abstracts, but for some there is no currently accessible information about their content. One was a conference presentation which was cancelled, and hence not actually presented. In scientific research, it is papers published in peer reviewed archival venues that count. Ephemeral presentations are not considered publications and referencing them is generally discourage by reviewers and editors, and (as previously mentioned in Section 6.3) by the NCFS.[209] Despite this, the previously mentioned pre-hearing submission[210] copied the entire list of references from the report as evidence of peer review, including the 22% that are not peer reviewed publications!

> changes in the amount of data used to build a system, can greatly influence system performance, changing the task protocols could indicate different performance strengths and weaknesses for these same systems.
>
> Because of the above reasons, this should not be interpreted as a product testing exercise and the results should not be used to make conclusions regarding which commercial products are best for a particular application.

[206] This is explicitly stated with respect to a Human Assisted Speaker Recognition (HASR) test:

> Forensic applications are among the applications that the HASR test serves to inform, but the HASR test should not be considered to be a true or representative "forensic" test. This is because many of the factors that influence speaker recognition performance and that are at play in forensic applications are controlled in the HASR test data, ... (NIST 2012 note 203 *supra* at p. 6)

[207] Transcript day 1, p 138

[208] *Daubert*, 509 U.S. at 593.

[209] NCFS 2015 *scientific literature* note 137 *supra*

[210] note 202 *supra*

7.2.5 Are there standards controlling the technique's operation?

The *Daubert* ruling states that "in the case of a particular scientific technique, the court ordinarily should consider ... the existence and maintenance of standards controlling the technique's operation."[211]

In her questioning of the forensic practitioner, the prosecutor appeared to take a very broad interpretation of the term *standard*. Mention was made of a protocol written by the practitioner and agreed to by the Swedish National Forensic Center (NFC),[212] the International Association for Forensic Phonetics and Acoustics (IAFPA) *Code of practice,*[213] and the European Network of Forensic Science Institutes (ENFSI) *Methodological guidelines for best practice in forensic semiautomatic and automatic speaker recognition*.[214] Nothing that was referred to as a standard was a National or International Standard, and the practitioner's laboratory was not accredited.[215] There is no evidence that the practitioner actually followed any standards which we would consider positive indicators of the quality of his work.

7.2.6 Is the thing from which the deduction is made sufficiently established to have gained general acceptance in the particular field in which it belongs?

Frye states: "the thing from which the deduction is made must be sufficiently established to have gained general acceptance in the particular field in which it belongs." The *Daubert* ruling states that: "'general acceptance' can yet have a bearing on the inquiry. A 'reliability assessment does not require, although it does permit, explicit identification of a relevant scientific community and an express determination of a particular degree of acceptance within that community.'"[216]

In their pre-hearing submission, the prosecution argued that:

> With respect to the fourth *Daubert* factor, "general acceptance," while biometric speaker recognition is a relatively new forensic tool, the science has gained a significant level of acceptance around the world and BATVOX is the "de facto standard" for forensic biometric voice recognition, utilized by numerous law enforcement agencies including the Federal

[211] *Daubert*, 509 U.S. at 594.

[212] The practitioner clearly followed this protocol. We think, however, that this protocol led the practitioner to follow certain operating procedures which we consider unwise, see §7.2.7 *infra*.

[213] http://www.iafpa.net/code.htm. This code of practice is very general, it addresses some ethical issues, but it says nothing about which approach or framework to use or how to use them, and it says nothing about testing validity and reliability.

[214] ENFSI 2015 *speech* note 156 *supra*. The guidelines had not been published at the time of the *Daubert* hearing, but a near-final draft was available. There is no evidence that the practitioner actually followed these guidelines.

[215] Transcript day 2, p 205, p 240

[216] *Daubert*, 509 U.S. at 594.

Bureau of Investigation.[217]

They also provided a list of 55 organizations around the world who own a copy of Batvox, including law enforcement agencies and private laboratories.[218] Even if the prosecution's contention regarding Batvox being the de facto standard were true, it would be irrelevant since the forensic practitioner's analysis and conclusions were not directly or primarily based on the output of Batvox.

During direct, we find the following exchange:

> Q Is there a consensus in the scientific community about the best way to conduct a phonetic speaker comparison?
>
> A Yes.
>
> Q What is that consensus?
>
> A It is to do it basically in the same way as expressed in the report, to go through those different areas and define first the known, the known samples, to create a kind of linguistics model, phonetic model of the speaker before you compare those features of similarities that you find, and you find similarities and you find differences. Take all those into account and compare that to defined reference population for the speech involved in the case.
>
> Q So just to clarify, there's a consensus in the community that the three methods that you used in this case is the best way to conduct a forensic voice comparison?
>
> ...
>
> A Yeah, there are some people that will only do automatic and there are some people still that only will do phonetic analyses, but the majority, the consensus is to use all of the methods available.[219]

The practitioner's assertion that there is a consensus as to the best way to do forensic voice comparison was just that, an assertion, he did not back it up with evidence. As we found in Appendix D, it may be the case that approaches based on subjective judgment are generally accepted among practitioners, but among researchers, overwhelmingly the norm is the use of data, quantitative measurements, statistical models, and empirical testing of validity and reliability.

---

[217] note 202 *supra* at p 16. Note that the FBI has a policy of not providing forensic voice comparison testimony in court.

[218] When questioned, the forensic practitioner said that "almost all major forensic governmental forensic agencies in European countries have this software". (Transcript day 1, p 38). In the INTERPOL survey (note 52 *supra*), of 26 respondents who reported using a named automatic system, Batvox was the most used system, used by 12 respondents. We would caution, however, that commercial success should not be taken as an indicator of scientific validity.

[219] Transcript day 1, p 41

7.2.7 Were reasonable steps taken to reduce the potential for contextual bias?

As we have already established, in *Ahmed* the forensic practitioner's conclusion as to the strength of the evidence was based primarily and directly on subjective judgment. Whether he took any precautions to attempt to reduce the potential for contextual bias is therefore a legitimate question to ask.

From reviewing the report and testimony it does not appear to us that the practitioner took any effective steps to shield himself from potentially biasing task irrelevant information.

During cross, the practitioner was asked whether he had a second practitioner independently perform the forensic analysis. He responded that he had a colleague do that to "some extent" and then they reached a consensus. The defense attorney did not press the matter, but it is unclear to us from the transcript whether a truly independent analysis was conducted by a second practitioner.[220] Even if the analyses were independent and they reached the same conclusion, this would not necessarily reduce the potential for contextual bias if both examiners were exposed to the same potentially biasing task irrelevant information.

We believe that the practitioner is a man of integrity and he did not deliberately set out to act in a biased manner, but he followed a number of procedures which we consider unwise because they expose him to potential allegations that he was influenced by contextual bias. Examples include the following:

- With one exception, what the practitioner stated as the strength of evidence arising from an automatic analysis was more favorable to the prosecution than the likelihood ratio value output by the automatic system itself.
- The exception (Comparison 3) was a likelihood ratio which slightly favored the defense. This he discounted, however, because the recording conditions were poor. If a forensic practitioner believes that conditions are too poor to obtain valid results, they should state so at the beginning (or after having run empirical tests of the validity and reliability of the system under those conditions), and they should not run the analysis of the actual known- and questioned-speaker recordings. A forensic practitioner should not get the results of such an analysis and then try to explain them away.
- The practitioner conducted one automatic analysis (Comparison 1), got a result, then modified the system (by adding an imposter set), reran the analysis, and got a result more favorable to the

[220] If independent analyses are conducted, these should be fully documented along with any differences in conclusions. Two genuinely independent analyses reaching the same conclusion may add greater certainty to that conclusion. If the analyses are not independent, *i.e.*, the two practitioners discuss the analysis as they go along and reach a consensus, this may reduce the potential for procedural errors, but it does not provide the same sort of greater certainty that independent analyses do. If the first practitioner tells the second their conclusion, the second practitioner is subject to a confirmation bias and they tend to agree with the first examiner to a greater extent than if they had conducted a truly independent analysis. See Thompson 2016 note 96 *supra*.

prosecution than the first result. A forensic practitioner should avoid acting in a way that could give the impression that they are cherry picking results, *i.e.*, that they tested multiple systems and then selected the one which was most favorable to the party instructing them.[221]

7.3 Conclusion with respect to the *Ahmed* testimony

The forensic practitioner in *Ahmed* used a mixture of approaches: auditory, acoustic-phonetic, and automatic. The results of all of the analyses were subjective judgments. Even for the automatic subsystem, which calculated likelihood ratios using quantitative measurements and statistical models, the practitioner did not directly report the calculated values, but instead used them as inputs to making a subjective decision. The way the results from each analysis were combined was also a subjective judgment. In general the procedures were not transparent, and were not described in sufficient detail that they could be replicated by another suitably qualified practitioner.

With respect to the *Daubert* factors, the practitioner did not empirically test the validity and reliability of his system under conditions reflecting those of the case under investigation. There is no evidence that he followed any standards which we would consider indicators of trustworthiness. Although there were some peer-reviewed publications supporting some aspects of his approach, their relevance for assessing the trustworthiness of his overall conclusions was limited. Whether his approach could be considered generally accepted in the relevant scientific community is unclear. Indeed, whether *any* particular approach is generally accepted at this time is unclear. While his approach may be in line with common practice among practitioners, it is not in line with current practice in the scientific research community. Clearly, we believe that the testimony did not satisfy the *Daubert* criteria and should not have been admitted.

Shortly after the hearing the prosecution offered what the defense viewed as a favorable plea bargain and the case was resolved with a negotiated plea, rendering the admissibility issue moot. Although some might interpret this development as evidence that the prosecution feared losing the *Daubert* hearing and the case, there is no way to know how the court would have ruled. It remains to be seen how courts will view voice comparison evidence when evidence of this type is offered in future cases.

8 Meeting the *Daubert* standard: What would a potentially admissible forensic voice

[221] To his credit, the practitioner was transparent about what he did in this instance – doing this and hiding it would obviously be even worse. Undoubtedly the practitioner's reasoning for running the second analysis was a genuine belief that it would give better results than the first analysis (see, for example, van der Vloed 2016 note 185 *supra*) and was therefore a better analysis to conduct (this may not have occurred to him at the time he ran the first analysis or he may not have had imposter data available at the time he ran the first analysis). One could potentially always think that there might be a better system. The question of interest for admissibility, however, is not whether the best possible system has been used, the question is whether the system that has been used is sufficiently scientifically valid. We think it better to choose one system, test it, then use it.

comparison analysis look like?

Our critique of the testimony presented in *Ahmed* has been overwhelmingly negative. This does not, however, mean that we believe that forensic voice comparison testimony could never be found admissible under *Daubert*. We think that, in practice, only approaches based on relevant data, quantitative measurement, and statistical models would be able to satisfy the *Daubert* criteria. Below we outline how we believe a forensic voice comparison would have to be conducted in order to produce testimony which could potentially be found admissible under *Daubert*. For more concrete examples based on actual cases and including technical details, see Enzinger & Morrison (2015),[222] Enzinger *et al*. (2016),[223] Enzinger (2016) Chapter 5,[224] and Zhang *et al*. (2016).[225]

1. To facilitate transparency and replicability, the forensic practitioner should document in chronological bench notes all decisions they make and all actions they take. All parties should be made aware of the existence of these notes, and they should be provided to all parties upon request. All substantial decisions and actions should also be documented in the forensic report. On the basis of the report, bench notes, and a copy of the practitioner's standard operating procedures and other appendices, another suitably qualified forensic practitioner (or researcher) should be able to critique the first practitioner's decisions and actions and potentially replicate what the first forensic practitioner did. If anything is unclear in the report and appendices, the second practitioner should be able to find the answer in the first practitioner's notes. The second forensic practitioner should not have to guess what the first forensic practitioner actually did.

2. To reduce the potential for contextual bias, the practitioner should take steps to avoid being exposed to task irrelevant information, *i.e.*, information about the case which is not necessary for them to perform their forensic voice comparison analysis. In large laboratories, a case manager may be assigned to handle communication with the client and other parties, and only pass on to the practitioner task relevant information. In smaller laboratories the practitioner should ask the client up front not provide task irrelevant information.

3. Based on an examination of the questioned-speaker recording, and relevant information provided by the client and other parties as may be appropriate given the circumstances of the case, the practitioner should formulate the details of the same-speaker hypothesis and the different-speaker hypotheses

[222] Ewald Enzinger, & Geoffrey Stewart Morrison, *Mismatched distances from speakers to telephone in a forensic-voice-comparison case*, 70 Speech Communication 28–41 (2015) http://dx.doi.org/10.1016/j.specom.2015.03.001

[223] Ewald Enzinger, Geoffrey Stewart Morrison, Felipe E. Ochoa, *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*, 56 Science & Justice 42–57 (2016) http://dx.doi.org/10.1016/j.scijus.2015.06.005

[224] note 32 *supra*

[225] Cuiling Zhang, Geoffrey Stewart Morrison, & Ewald Enzinger, *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 (2016) http://dx.doi.org/10.1016/j.forsciint.2016.08.017

that they plan to assess. The different-speaker hypothesis must include the definition of the relevant population. Before proceeding, the suitability of these hypotheses should be confirmed with the client and other parties as may be appropriate given the circumstances of the case. The hypotheses, including the relevant population, should be clearly described in the report.

4. Based on an examination of the known- and questioned-speaker recordings, and relevant information provided by the client and other parties as may be appropriate given the circumstances of the case, the practitioner should describe what they understand to be the speaking styles and recording conditions of the known-speaker recording and the questioned-speaker recording. All reasonable enquiries should be made to obtain technical details about recording systems, *etc*. These conditions should be clearly described in the report.

5. If the practitioner believes that *a priori* the conditions of the recordings are so poor that the performance of their forensic voice comparison system will be so poor that it is unlikely to be of value to the court, they should inform the client of this before proceeding. The client may still request that the practitioner proceed, but this will be an informed decision. If the client decides not to have the practitioner proceed with a particular comparison, this should be documented in the report, and no further analyses should be conducted on the relevant recordings.

6. The known- and questioned-speaker recordings should be prepared by selecting only portions of the recordings which actually contain speech of the speaker of interest. Interlocutor speech, transient noises, and stretches of silence or background noise should be excluded from the analysis. (This will reveal one aspect of the recording conditions, the net durations of the know-speaker and the questioned-speaker speech.)

7. The practitioner should then obtain a sample of voice recordings representative of the relevant population and reflecting the speaking styles and recording conditions of the known-speaker recording and the questioned-speaker recording. The sample may come from an existing database, or new data may need to be collected. The practitioner must themself be satisfied that the sample recordings are sufficiently representative and reflective of the relevant population, speaking styles, and recording conditions. The report must explain how the forensic practitioner sampled the speakers, and how they replicated or simulated the conditions. Sufficient detail must be provided so that the judge at an admissibility hearing has a basis on which to consider whether the recordings are sufficiently representative and reflective. This is a topic on which we would expect the opposing parties to seek expert advice, and to debate before the judge during an admissibility hearing. (If the testimony is admitted, this topic may also be argued before the trier of fact in relation to weight.)

8. The relevant population sample recordings should be prepared in the same manner as the known- and questioned-speaker recordings.

9. The practitioner should split their data into at least two separate parts: a training set and a test set.

Statistical models should not be trained and tested on the same data.[226]

10. To reduce the potential for contextual bias, the practitioner should use a system based on relevant data, quantitative measurements (*e.g.*, measurements of acoustic properties of the voice recordings), and statistical models. The output of the statistical model should be directly reported, it should not be used as input to a subsequent a subjective judgment process.

11. The system should be trained and optimized using the training data, which reflect the relevant population, speaking styles, and recording conditions of the case. Ideally, a second forensic practitioner should check the first forensic practitioner's work at this stage in search of any potential mistakes. Once the forensic practitioner is satisfied with the training and optimization of the system, the system should be frozen, *i.e.*, no subsequent changes to the system will be allowed.[227]

12. The practitioner should then use the test data to empirically assess the performance of their system. The system as a whole should be tested, including any components depending on the particular human operator. The system which is tested should be the same system which will actually be used to compare the known- and questioned-speaker recordings. The results of the tests should be documented in the report, and an explanation of how to interpret any numeric or graphical results should be provided in the report or in an appendix. Sufficient detail should be provided to assist the judge at an admissibility hearing to decide if system performance is sufficient to warrant admission of the testimony. (If the testimony is admitted, this question may also be argued before the trier of fact in relation to weight.) Ideally, a second forensic practitioner should check the first forensic practitioner's work at this stage in search of any potential mistakes. Once the tests have been conducted, they should not be repeated in search of better results. The system should not be altered and then retested on the same data set.[228]

13. The last step in the analysis should be to actually compare the known- and questioned-speaker recordings. The numeric likelihood ratio generated by the system should be reported as the strength of evidence statement. The report, or an appendix, should include an explanation of the likelihood ratio framework so that the judge at an admissibility hearing and the trier of fact at trial can

[226] As an alternative to two completely separate sets of data, a procedure known as cross-validation can be used to ensure that the same data are not used for training and testing. Statistical models perform better on the same data that was used to train them versus on new data. It is performance on new data that matters, the actual known- and questioned-speaker recordings will be new data. Training and testing on the same data would give an overly optimistic assessment of the expected performance on the system on the actual known- and questioned-speaker recordings.

[227] The only exception will be if a genuine mistake is discovered at a later stage. Any such change must be fully documented in the report.

[228] Altering the system and retesting on the same data leads to optimization on the test data and the results will be overly optimistic with respect to performance on new data, *i.e.*, the actual known- and questioned-speaker recordings. Optimization should be performed on training data, not on test data. Training data may be split into an initial training set and an optimization set, or cross-validation may be used.

> understand how to appropriately interpret the result. Once the likelihood ratio for the comparison of the known- and questioned-speaker recordings has been obtained, the system should not be altered or retested, and the likelihood ratio should not be recalculated in search of a better answer.[229]

Such procedures would, we believe, be potentially admissible under *Daubert* because they are logically correct, robust to cognitive bias, transparent and replicable, and include demonstration of degree of validity and reliability under conditions reflecting those of the case under investigation. If the judge at an admissibility hearing is satisfied that the test data are sufficiently representative of the relevant population and sufficiently reflective of the speaking styles and recording conditions of the known-speaker recording and the questioned-speaker recording, and that the empirically demonstrated degree of validity and reliability of the system under these conditions is adequate, then the system will have passed what we consider to be the most important *Daubert* criterion, *i.e.*, "whether the reasoning or methodology underlying the testimony is scientifically valid and ... whether that reasoning or methodology properly can be applied to the facts in issue", including "whether it can be (and has been) [empirically] tested", and "in the case of a particular scientific technique, ... consider[ation of] the known ... rate of error".[230]

9 Conclusion

We have argued that the most important *Daubert* criterion for deciding the admissibility of an implementation of any approach to forensic voice comparison (be it auditory, acoustic-phonetic non-statistical, acoustic-phonetic statistical, or automatic) is whether it has been empirically tested under conditions reflecting those of the particular case under investigation, and found to be sufficiently valid and reliable. We see this as the direct primary indicator of scientific validity, and the other *Daubert* criteria as secondary proxy indicators. If the judge accepts that the test data are sufficiently representative of the relevant population and sufficiently reflective of the conditions of the case under investigation, they can then consider whether the empirically demonstrated performance of the system under those conditions is sufficient to warrant admission. We have also argued that, because of the substantial case-to-case variability in relevant population, speaking styles, and recording conditions, system performance will need to be empirically assessed on a case by case basis, and admissibility will need to be considered on a case by case basis.

Although we have concentrated on admissibility under FRE 702 and *Daubert*, and to a lesser extent *Frye*, our arguments are founded on what we consider to be good scientific practice, and, from a scientific perspective, these should be relevant irrespective if the details of codified legal standards for

[229] Again, the only exception will be if a genuine mistake is discovered at a later stage. Any such change must be fully documented in the report.

[230] *Daubert*, 509 U.S. at 592–3.

admissibility.

Although our focus has been on the admissibility of forensic voice comparison testimony, we believe that it would be logically consistent to apply the same criteria in considering the admissibility of testimony based on comparison of other items of forensic interest.

10 Acknowledgments

This work was partially funded by two visiting fellowships from the Simons Foundation, one awarded to each author. The authors would like to thank the Isaac Newton Institute for Mathematical Sciences for its hospitality during the program Probability and Statistics in Forensic Science which was supported by EPSRC Grant Number EP/K032208/1. The second author also received support from National Institute of Justice Award 2014-DN-BX-K032.

11 Appendix A: Combing a prior probability and a likelihood ratio to calculate a posterior probability (Bayes' Theorem)

Logically, in order to arrive at a *posterior probability* one has to combine two things: a *prior probability* and a *likelihood ratio*. That this is logically true can be formally proven and is not disputed among logicians or statisticians. This piece of logic is called *Bayes' Theorem*, and formal descriptions of this logic date back to the mid 1700s. Popular literature on this topic includes McGrayne (2011)[231] and Lindley (2006).[232]

To explain how a *prior probability* and a *likelihood ratio* are combined to arrive at a *posterior probability* we will introduce a mathematical formula, but it is not complex, it only involves multiplication and division. First the formula (which is called the *odds form of Bayes' Theorem*), then an explanation of what it means embedded in a concrete example.

Odds form of Bayes' Theorem:

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

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

$$\frac{1}{1000} \times 4 = \frac{1}{250}$$

[231] note 50 *supra*

[232] note 51 *supra*

Imagine that a crime has been committed on an island, and the population of the island is about 1000. A suspect is arrested before anyone has had the opportunity to leave the island. The suspect is put on trial.

The prosecution contends that the defendant is guilty, *i.e.*, the prosecution hypothesis, H_p, is that the offender is the defendant. Before any other information is presented, what is the probability that the offender is the defendant? We could, perhaps, interpret the legal doctrine of innocent until proven guilty as meaning that at the beginning of the trial the trier of fact should assume that the defendant is no more or less likely to be guilty than any other person selected at random from the population. Since there are about 1000 people on the island, this would convert to a probability of guilt of about 1 in 1000, which we can write as 1/1000. The prior probability that the offender is the defendant, $p(H_p)$, we therefore set at 1/1000.[233]

The defense contends that the defendant is not guilty, *i.e.*, the defense hypothesis, H_d, is that the offender is not the defendant but someone else from the population. The two hypotheses are exhaustive (there are no other options other than the offender is the defendant or the offender is someone else on the island), and they are mutually exclusive: if H_p is true then H_d is false and *vice versa*, they cannot both be true or both be false. Under these conditions, the total probability, the sum of the two prior probabilities, must be 1 (100%), *i.e.*, $p(H_p) + p(H_d) = 1$, therefore $p(H_d) = 1 - \frac{1}{1000} = \frac{999}{1000}$. The first term in the equation above is the prior probability of the prosecution hypothesis divided by the prior probability of the defense hypothesis: $\frac{p(H_p)}{p(H_d)} = \frac{1/1000}{999/1000} = \frac{1}{999} \approx \frac{1}{1000}$.[234] This is called the *prior odds*. What do the prior odds mean? In this example, they mean that before any (additional) evidence has been presented, the trier of fact believes that the defense hypothesis is about 1000 times more probable than the prosecution

[233] Our example of one divided by the size of the population is easy to understand, but overly simplistic. Perhaps some portion of the population are children who could not have committed the crime, perhaps people who live or work near the scene of the crime are more likely to have committed it than people who live in a remote part of the island. The prior probability is whatever the trier of fact believes it to be. Factors such as the size of the population of people who could potentially have committed the crime will affect that belief, but in general how the trier of fact forms that belief is not dictated by any prescribed formula. The trier of fact should not, however, use the fact that the defendant is on trial to form their prior probability belief. That the defendant is on trial is the end point of a consideration of evidence and a reasoning process conducted by the police and the prosecutor. The prosecutor must believe that the defendant is guilty, otherwise it would be unethical of them to proceed with the prosecution. The prosecutor must present to the trier of fact the evidence and the reasoning that led the prosecutor to that conclusion (or a version of that reasoning and a subset of the evidence according to the constraints of what they are practically able and legally allowed to present). The trier of fact must therefore form an initial prior probability which corresponds to the start of that reasoning process, not to its end. See *Bell v. Wolfish*, 441 U.S. 520, 533: "The presumption of innocence… may serve as an admonishment to the jury to judge an accused's guilt or innocence solely on the evidence adduced at trial, and not on the basis of suspicions that may arise from the fact of his arrest, indictment, or custody, or from other matters not introduced as proof at trial."

[234] The symbol $\approx$ means "approximately equal to". 1/999 is approximately 1/1000. We originally said that the population of the island was *about* 1000, so it does not make sense to use a value as exact as 1/999.

hypothesis.

Returning to our previous example involving hair color, all the eyewitnesses agree that the offender had blond hair, and it also turns out that the defendant has blond hair. Using the same simplifications as before, the probability that the offender would have blond hair if they were the defendant[235] is 1 (100%), *i.e.*, $p(E|H_p) = 1$. In this part of the formula *E* is the evidence (*E* = the offender has blond hair), H_p is the prosecution hypothesis, and $p(E|H_p)$ is the probability of the evidence if the prosecution hypothesis were true. A forensic practitioner obtains a random sample of 100 people on the island. 25 of them have blond hair, so the practitioner estimates that the probability that the offender would have blond hair if they were not the defendant but someone selected at random from the population of the island is about 25/100 = 1/4, *i.e.*, $p(E|H_d) = 1/4$. In this part of the formula *E* is the evidence (*E* = the offender has blond hair), H_d is the defense hypothesis, and $p(E|H_d)$ is the probability of the evidence if the defense hypothesis were true. The forensic practitioner therefore calculates that the *likelihood ratio* is $\frac{p(E|H_p)}{p(E|H_d)} = \frac{1}{1/4} = 4$. The probability that the offender would have blond hair if they were the defendant is about 4 times greater than if they were someone selected at random from the population of the island.

If the trier of fact were to follow the normative logic of Bayes' Theorem, they would use the formula above, and in this example they would multiply their prior odds $\left(\frac{p(H_p)}{p(H_d)} = \frac{1}{1000}\right)$ by the likelihood ratio $\left(\frac{p(E|H_p)}{p(E|H_d)} = 4\right)$ to arrive at *posterior odds* $\left(\frac{p(H_p|E)}{p(H_d|E)}\right)$ of $\frac{1}{1000} \times 4 = \frac{4}{1000} = \frac{1}{250}$. The posterior odds are the relative probabilities of the prosecution hypothesis being true versus the defense hypothesis being true after having considered the strength of the evidence. What do the posterior odds mean? In this example, they mean that after having heard the hair-color testimony the trier of fact believes that the defense hypothesis is about 250 times more probable than the prosecution hypothesis.[236]

If another piece of testimony, *e.g.*, testimony based on forensic voice comparison evidence, is subsequently presented, the trier of fact's posterior odds after hearing the hair-color testimony become their prior odds before hearing the forensic voice comparison testimony.[237] If all the testimony combined

[235] Note that we have reformulated the hypotheses here, to the offender is the defendant versus they are not the defendant, rather than guilty versus not guilty. In fact, we should reformulate this even further to be the person observed by the witnesses is the defendant versus they are not the defendant. In general, the likelihood ratio generated by the forensic scientist will not (and should not) directly address the issue of guilt. In the present example, the defense could argue that the defendant was at the crime scene at the time of the crime, but they were not the one who committed the crime. In that case, the likelihood ratio we are calculating in the present example in relation to hair color would be irrelevant, and would not assist the trier of fact to reach a decision on the question of guilt.

[236] We could do some additional mathematics to calculate the posterior probability of the prosecution hypothesis, $p(H_p|E)$, but it is actually easier to think in terms of odds.

[237] The different evidence could be presented in any order, *e.g.*, forensic voice comparison testimony could be first and hair

leads to high enough posterior odds, then the trier of fact may decide that the case has been proven beyond a reasonable doubt.[238]

Even if they are not aware of it, if a forensic practitioner presents a posterior probability, such as a 95% probability that the voice on the questioned-speaker recording was produced by the known speaker, they must have at least implicitly used a prior probability. But unless the trier of fact tells the forensic practitioner what prior probability to use, the forensic scientist cannot calculate the appropriate posterior probability. If a forensic practitioner were to present a posterior probability, the only logically correct way for the trier of fact would use it would be for them to find out what (explicit or implicit) prior odds the forensic practitioner used, divide the forensic practitioner's posterior odds by the forensic practitioner's prior odds to calculate the likelihood ratio, then multiply the trier of fact's prior odds with the likelihood ratio.

12 Appendix B: Testing the validity of a forensic voice comparison system

A general protocol for testing the validity of a forensic voice comparison system is as follows: A pair of voice recordings is presented to the system, one recording with conditions reflecting those of the known-speaker recording and the other with conditions reflecting those of the questioned-speaker recording. The tester knows whether this pair of recordings is a same-speaker pair or a different-speaker pair, but the system being tested must not be told which of these is true. The system analyzes the two recordings and outputs a strength of evidence statement. If operating within the likelihood ratio framework, this output would be presented as a likelihood ratio. In another framework the system could, for example, output "same-speaker" or "different-speaker". Whatever the output of the system, the tester compares this output with their knowledge as to whether the input was a same-speaker pair or a different-speaker pair, and assigns a goodness score (or a badness score) to the result. For example, if the input is "same-speaker" and the output is "same-speaker" this is good, but if the input is "same-speaker" and the output is "different-speaker" this is bad. A standard procedure is to assign a *cost* (a badness score) to each answer, *e.g.*, if the output is correct the cost assigned is 0, but if the output is incorrect the cost assigned is 1 (see Table 1 for a list of correct and incorrect combinations in this framework).

color second. Mathematically, the initial prior odds and multiple likelihood ratios will produce the same result irrespective of the order in which they are multiplied together.

[238] Of course, in a real case it is unlikely that all the evidence presented will be forensic evidence with numerically quantified strength of evidence, and it may be that the trier of fact does not use normative Bayesian reasoning.

Table 1. List of input and output possibilities and corresponding correctness for a system which outputs either "same-speaker" or "different-speaker". (This is not consistent with the likelihood ratio framework.)

		output	
		same-speaker	different-speaker
input	same-speaker	correct	incorrect
	different-speaker	incorrect	correct

Within the likelihood ratio framework, the answer is not "same-speaker" versus "different-speaker", but a gradient value such that the larger the likelihood ratio value the greater the support for the same-speaker hypothesis over the different-speaker hypothesis, and the smaller the likelihood ratio value the greater the support for the different-speaker hypothesis over the same-speaker hypothesis. Within this framework, when the input is a different-speaker pair, the larger the likelihood ratio value the higher the cost assigned, and the smaller the likelihood ratio value the lower the cost assigned. Also, when the input is a same-speaker pair, the smaller the likelihood ratio value the higher the cost assigned, and the higher the likelihood ratio value the lower the cost assigned. The tester presents a large number of same-speaker test pairs and a large number of different-speaker test pairs,[239] calculates the cost for each pair, then averages over all the cost values. The smaller the average cost value the better the validity of the forensic voice comparison system.

13 Appendix C: Testing the reliability of a forensic voice comparison system

Several factors can affect the reliability (precision) of a forensic voice comparison system, including intrinsic variability at the source, sampling variability, and measurement variability. For example, using one recording of the known speaker rather than another, or using one sample of the relevant population rather than another, or re-measuring the same recordings again can result in a different value for a calculated likelihood ratio. In general, the smaller the size of the sample used to train a statistical model, the poorer the reliability of the model.[240]

There are several solutions proposed for dealing with imprecision in forensic likelihood ratios:

One proposal is not to report a specific number, but rather to report that the likelihood ratio lies within a

[239] Ultimately the size of the test set will have to be large enough to satisfy the judge at an admissibility hearing and/or the trier of fact at trial. As discussed earlier with respect to the size of samples for estimating probabilities, the larger the set of test data the better the estimate of the performance of the system, but the higher the financial cost. Irrespective of financial cost, we would expect the total number of same-speaker test pairs to need to be at least in the tens and total number of different-speaker test pairs to need to be at least in the hundreds for the judge / trier of fact to be satisfied.

[240] In the context of forensic voice comparison smaller sample sizes could be due to having shorter known-speaker and/or questioned-speaker recordings, fewer recordings of the known speaker, or fewer recordings of speakers representative of the relevant population.

range, *e.g.*, between 10 and 100, or between 100 and 1000, and to give verbal expressions to each range, *e.g.*, "moderate support for one hypothesis over the other", "moderately strong support for one hypothesis over the other". Some practitioners whose strength of evidence statements are based on subjective judgments proceed directly to picking one of the verbal expressions in a predefined scale and never calculate a numeric likelihood ratio value.

Another proposal is to use additional test data to numerically estimate and report the degree of precision of the system. Some protocols for doing this are similar to the protocol described above for testing validity, but somewhat more complex. For example: Use multiple known-speaker-condition recordings of each test speaker. Compare each known-speaker-condition recording of a given speaker with a questioned-speaker-condition recording of a given speaker (could be a same-speaker or a different-speaker comparison). Look at the variability within the resulting group of likelihood ratio values. Repeat for other combinations of same- and different-pairs of test speakers, and calculate an average of the within-group variabilities. Results of a forensic analysis may then be reported as a best estimate plus a range, *e.g.*, my best estimate for the strength of the evidence is a likelihood ratio of 1000 and based on the results of tests of the reliability of my system I am 98% certain that it is greater than 100 and less than 10,000. Results may also be reported as a best estimate and the end of the range closest to the neutral likelihood ratio value of 1, *e.g.*, my best estimate for the strength of the evidence is a likelihood ratio of 1000 and based on the results of tests of the reliability of my system I am 99% certain that it is greater than 100.

Other researchers and practitioners have philosophical objections to the whole idea of measuring the precision of likelihood ratios. They report a single value, but may adopt statistical procedures which result in likelihood ratio values which are closer to the neutral value of 1 than would otherwise be the case.

Whether forensic practitioner should or should not assess the precision of the output of their forensic analysis systems, and if so how, is a matter of debate. Interested readers may wish to consult the papers in a 2016–2017 virtual special issue on measuring and reporting the precision of forensic likelihood ratios in the journal *Science & Justice*, and papers cited in those papers.[241]

14 Appendix D: Is there any evidence that there is currently a generally accepted approach to forensic voice comparison?

Is there any evidence that there is currently a generally accepted approach to forensic voice comparison?

If the relevant scientific community were chosen to be forensic voice comparison practitioners, there are

[241] http://www.sciencedirect.com/science/journal/13550306/vsi

two relatively recent surveys of practitioners which can be considered. Gold & French (2011)[242] published the results of a survey of forensic voice comparison practitioners, including some working in private laboratories and universities, and some working in law-enforcement and government laboratories. Of 35 respondents:

- 25 (71%) reported using an auditory–acoustic–phonetic approach.
- 7 (20%) reported using a human supervised automatic approach.
- 2 (6%) reported using an auditory-only approach.
- 1 (3%) reported using an acoustic-phonetic-only approach.
- The spectrographic approach was not mentioned.

Turning to a more recent survey conducted by INTERPOL,[243] however, the picture changes somewhat. The number of respondents who reported having speaker identification capabilities was 44, but many reported using more than one approach (hence the following values add up to more than 44 and more than 100%).

- 25 (41%) reported using an auditory-acoustic-phonetic (non-statistical) approach.
- 21 (34%) reported using a spectrographic or auditory-spectrographic approach.
- 20 (33%) reported using a human-supervised automatic approach.
- 15 (25%) reported using an auditory approach.
- 15 (25%) reported using an acoustic-phonetic-statistical approach.
- 9 (15%) reported using a fully-automatic approach.

Differences between the results of the two surveys may in part be attributable to the fact that the INTERPOL survey only solicited responses from law-enforcement agencies.[244] A particularly notable difference between the surveys was the great popularity of the spectrographic approach found by the INTERPOL survey compared to its complete absence in the reported results of the Gold & French survey.

Since auditory-acoustic-phonetic approaches were the most popular in both surveys, one could conclude

[242] Erica Gold, J. Peter French, International practices in forensic speaker comparison, 18 International Journal of Speech, Language and the Law 143–152 (2011) http://dx.doi.org/10.1558/ijsll.v18i2.293 (hereinafter Gold & French survey)

[243] note 52 *supra*

[244] Non law enforcement laboratories were only potentially included if they were contracted to work for law enforcement agencies. The INTERPOL survey also potentially included responses related to investigative applications in addition to forensic applications (the latter being related to the preparation of reports and testimony for presentation in court). The fully-automatic responses are likely to have been related to investigative applications.

that this is generally accepted. Whereas this represented a majority (71%) in the Gold & French survey, in the INTERPOL survey it was only the largest minority (41%). In the results of the INTERPOL survey no approach was used by a majority, hence there seems to be a lack of consensus among practitioners in law-enforcement agencies, and no approach appears to be generally accepted, at least if general acceptance requires a majority. *Williams*[245] and *Smith*[246] held under *Frye* that general acceptance does not require a majority, but in that case would 33%, a substantial minority, be enough? If so, then the spectrographic approach would meet this criterion, but *Angleton* under *Daubert* found it not to be generally accepted by the scientific community.[247]

If we group the approaches which by definition are based on subjective judgment (auditory, spectrographic, and acoustic-phonetic non-statistical approaches) versus the others (although, rather than being directly presented, the output of acoustic-phonetic statistical and automatic approaches can also be used as input to a subjective judgment process), the balance of responses is 61 to 44 in the INTEPROL survey and 27 to 8 in the Gold & French survey. This could suggest that, although there may be a lack of consensus among practitioners as to exactly which approach to use, approaches in which the strength of evidence statement is primarily and directly based on subjective judgment are generally accepted.

Choosing practitioners as the relevant "scientific" community, however, may be problematic. A better "scientific" community may be those who publish peer-reviewed research on forensic voice comparison. Given our previous comments on the quality of the peer-reviewed literature, however, this may also be a problematic choice. For what it is worth, in a review of forensic speech science literature published between mid 2010 and mid 2013, Morrison & Enzinger (2013)[248] found that in contrast to earlier years there had been a shift toward the vast majority of experiment-based publications empirically testing systems which used data, quantitative measurements, and statistical models to calculate likelihood ratios, *i.e.*, acoustic-phonetic statistical and/or automatic approaches combined with the use of the likelihood ratio framework (we count 33 papers in this class). Papers not using the likelihood ratio framework were in the minority (we count 4), and papers only describing approaches in which the conclusion as to the strength of evidence was based primarily or directly on subjective judgment were in the distinct minority (we count 1 of the latter 4). General acceptance in the scientific research community therefore appears to be empirical testing of systems which use data, quantitative measurements, and statistical models to calculate numeric likelihood ratios as strength of evidence statements.

---

[245] *United States v Williams*, 583 F. 2d 1194,1198 (CA2 1978)

[246] *United States v Smith*, 869 F. 2d 348, 353–354 (CA7 1989)

[247] In *Angleton* the court considered research literature on auditory-spectrographic approaches covering a period of over 30 years, and concluded that: "These articles show that neither voice spectrography nor aural spectrographic analysis has been generally accepted as a method of identifying unknown recorded speakers." 269 F.Supp.2d at 900.

[248] Geoffrey Stewart Morrison, Ewald Enzinger, *Forensic speech science – Review: 2010–2013, in* Proceedings of the 17th International Forensic Science Mangers' Symposium 616–623, 629–635 (Niamh NicDaéid 2013)

With respect to framework for evaluation of forensic evidence the INTERPOL survey results were:

- 22 (50%) identification / exclusion / inconclusive
- 10 (23%) numeric likelihood ratio
- 9 (20%) verbal likelihood ratio
- 4 (9%) verbal posterior probability
- 3 (7%) numeric posterior probability
- 3 (7%) UK framework

Among law enforcement agencies who responded to the INTERPOL survey and indicated that they have speaker recognition capabilities, identification / exclusion / inconclusive was by far the most popular framework for expressing strength of evidence. The likelihood ratio framework was second most popular (18 respondents, 41%, indicated that they used numeric or verbal likelihood ratios or both).

The vast majority of authors who regularly publish on the topic of forensic inference and statistics in refereed journals agree that the likelihood ratio framework is the logically correct framework for the evaluation of forensic evidence (we could list thousands of references). These authors may (and do) disagree on nuances, and on details of how best to implement the framework, but they agree that it is the logically correct framework.

15 Appendix E: Scale of conclusions used by the forensic practitioner in *Ahmed*

The following is the Swedish National Laboratory of Forensic Science nine-level scale of conclusions with additions to the verbal expressions made by the forensic practitioner. The additions are in italics. The equivalent likelihood ratio ranges come from Nordgaard *et al*. (2012).[249] The end of each range closest to a likelihood ratio of 1 is included in the range, the end of the range furthest from 1 is excluded. The values 1/6 and 6 are excluded from the Level 0 range.

[249] Anders Nordgaard, Ricky Ansell, Weine Drotz, & Lars Jaeger, *Scale of conclusions for the value of evidence*, 11 Law, Probability and Risk 1–24 (2012) http://dx.doi.org/10.1093/lpr/mgr020

Level	Verbal expression	Equivalent likelihood ratio range
+4	The results of the examination extremely strongly support that *the compared speech material originates from the same speaker*. The results are extremely more probable if the main hypothesis is true compared to if the alternative hypothesis is true. *Very striking and distinctive similarities revealed themselves during the comparison of recordings. Phonetically and acoustically the speech showed consistent and distinctive similarities in accordance with the main hypothesis. Even if it currently is impossible to rule out the possibility that there is some support for the alternative hypothesis (others in the population who share the relevant features from a recorded voice) I believe this possibility to be close to negligible in this case.*	1,000,000 and greater
+3	The results of the examination strongly support that *the compared speech material originates from the same speaker.* The results are much more probable if the main hypothesis is true compared to if the alternative hypothesis is true. *Several distinctive similarities revealed themselves during the comparison of the recordings. Phonetically and acoustically the speech showed several distinctive similarities in accordance with the main hypothesis. Even if it currently is impossible to rule out the possibility that there is some support for the alternative hypothesis (others in the population who share relevant features from a recorded voice) I believe this possibility to be very small in this case.*	6,000 – 1,000,000
+2	The results of the examination support that *the compared speech material originates from the same speaker*. The results are more probable if the main hypothesis is true compared to if the alternative hypothesis is true. *Several similarities were revealed during the comparison of the recordings. Phonetically and acoustically the speech showed several similarities in accordance with the main hypothesis. Even if it currently is impossible to rule out the possibility that there is some support for the alternative hypothesis (others in the population who share relevant features from a recorded voice) I believe this possibility to be small in this case.*	100 – 6,000

Level	Verbal expression	Equivalent likelihood ratio range
+1	The results of the examination support to some extent that *the compared speech material originates from the same speaker*. The results are somewhat more probable if the main hypothesis is true compared to if the alternative hypothesis is true. *No dissimilarities were revealed during the comparison of the recordings. Phonetically and acoustically the speech showed similarities in accordance with the main hypothesis. On the account that only a small number of non-distinctive similarities were revealed, a certain support for the alternative hypothesis can not be ruled out (a number of other speakers with the same regional, social and ethnical background might share some relevant features).*	6 – 100
0	The results of the examination support neither of the hypotheses that *the compared speech originates from the same or different speakers*. The results are equally probable if the main hypothesis is true compared to if the alternative hypothesis is true.	1/6 – 6
−1	The results of the examination support to some extent that *the compared speech material does not originate from the same speaker*. The results are somewhat more probable if the alternative hypothesis is true compared to if the main hypothesis is true.	1/100 – 1/6
−2	The results of the examination support that *the compared speech material does not originate from the same speaker.* The results are more probable if the alternative hypothesis is true compared to if the main hypothesis is true.	1/6,000 – 1/100
−3	The results of the examination strongly support that *the compared speech material does not originate from the same speaker.* The results are much more probable if the alternative hypothesis is true compared to if the main hypothesis is true.	1/1,000,000 – 1/6,000
−4	The results of the examination extremely strongly support that *the compared speech material does not originate from the same speaker.* The results are extremely more probable if the alternative hypothesis is true compared to if the main hypothesis is true.	1/1,000,000 and less

Note included on the Swedish National Laboratory of Forensic Science version (the version valid from 21 January 2013):

> If one of the hypotheses can be excluded other terms are used, such as ‘it is’, ‘it is not’ or ‘it can be excluded that’.

Note included on the forensic practitioner’s version:

> The numbers representing the levels in the scale are only to a certain degree statistically based through calculation and to some extent a judgement of likelihood ratios.