This version includes corrections to errors that were introduced after the proof stage.

Admissibility of Forensic Voice Comparison Testimony in England and Wales

Geoffrey Stewart Morrison

Reader in Forensic Speech Science, Centre for Forensic Linguistics, Aston University, Birmingham

Admissibility; Criminal evidence; Expert evidence; Voice recognition

In 2015 the Criminal Practice Directions (CPD) on admissibility of expert evidence in England and Wales were revised. They emphasised the principle that "the court must be satisfied that there is a sufficiently reliable scientific basis *for the evidence to be admitted". The present paper aims to assist courts in understanding from a scientific perspective what would be necessary to demonstrate the validity of testimony based on forensic voice comparison. We describe different technical approaches to forensic voice comparison that have been used in the UK, and critically review the case law on their admissibility. We conclude that courts have been inconsistent in their reasoning. In line with the CPD, we recommend that courts enquire as to whether forensic practitioners have made use of data and analytical methods that are appropriate and adequate for the case under consideration, and that courts require forensic practitioners to empirically demonstrate the level of performance of their forensic voice comparison system under conditions reflecting those of the case under consideration.*

1. Introduction

A forensic voice comparison[1] is an analysis conducted by a forensic practitioner in order to assist the court to determine the identity of a speaker on an audio recording.[2] This usually involves comparing two recordings, one of a speaker of questioned identity (e.g. a recording of an intercepted telephone call), and the other of a speaker of known identity (e.g. a recording of an interview with a suspect). There is a growing body of literature discussing problems in multiple branches of forensic science. Many of the arguments presented here would be applicable in other branches of forensic science, but for brevity this article discusses only forensic voice comparison.

Forensic voice comparison is challenging because recordings of human voices are highly variable. Speaking styles can be more casual or more formal, speakers can be excited, or calm, or tired, they can even whisper, shout, or be suffering

---

[1] "Forensic voice comparison" has also been called by other names, including "forensic speaker comparison", "forensic speaker recognition", and "forensic speaker identification".

[2] Analyses may also be conducted for investigative purposes, or their results may be used in negotiations to resolve cases without going to trial.

from a cold, and speakers say different words and phrases on different occasions. Recording conditions in forensic cases are also highly variable and often degrade speaker-dependent information. This includes transmission through landline or mobile telephones, different kinds and loudness of background noise, echoes, recordings being saved in compressed formats such as MP3, and recordings being very short. Mismatches in speaking style and recording conditions can make recordings of the same speaker more different than they would otherwise be, and can mask differences between recordings of different speakers. A forensic practitioner has to assess whether the differences between the known- and questioned-speaker recordings are more likely to occur if they were produced by the same speaker or if they were produced by different speakers.

Over the past 20 years, there have been substantial developments in automatic speaker recognition and its application to forensic voice comparison. Implementations of automatic approaches based on relevant data, quantitative measurements, and statistical models are poised to replace subjective-judgement-based auditory and auditory-acoustic-phonetic approaches.

Below we discuss criteria for admissibility of expert evidence, then review the case law on admissibility of forensic voice comparison in England and Wales, with an excursion to Northern Ireland.[3] We successively discuss the admissibility of auditory, auditory-acoustic-phonetic, and automatic approaches.[4]

2. Admissibility criteria

Whatever approach is used, we argue that admissibility of forensic voice comparison testimony should be determined via a rigorous application of the criteria set out in the Criminal Practice Directions (CPD) at para.19A[5]:

> "It is essential to recall the principle which is applicable, namely in determining the issue of admissibility, the court must be satisfied that there is a *sufficiently reliable scientific basis* for the evidence to be admitted."[6]

The CPD at paras 19A.5–19A.6 state that

---

[3] A review of admissibility of forensic voice comparison in the United States appears in G.S. Morrison and W.C. Thompson, "Assessing the admissibility of a new generation of forensic voice comparison testimony" (2017) 18 *Columbia Science and Technology Law Review* 326. Readers are referred to the latter publication for much more extensive coverage of (jurisdictionally neutral) technical matters than is presented in the present paper. The CPD para.19A admissibility criteria have substantial parallels with those of United States Federal Rule of Evidence 702 (as amended 17 April 2000, eff. 1 December 2000; 26 April 2011, eff. 1 December 2011 (FRE 702)) and the US Supreme Court ruling in *Daubert v Merrell Dow Pharmaceuticals* 509 U.S. 579 (1993). The Law Commission Report which led to the introduction of the CPD on expert evidence explicitly drew on FRE 702 (Law Commission, *Expert Evidence in Criminal Proceedings in England and Wales*" (2011), Law Com. No.325, para.3.33) and referred to *Daubert* as "the equivalent reliability test in the United States" (at para.5.91). See also: T. Ward, "An English Daubert? Law, forensic science and epistemic deference" (2015) 15 *Journal of Philosophy, Science & Law* 26.

[4] It should be noted that, unlike in the US, spectrographic or aural-spectrographic approaches do not appear to have been used in the UK except as a supplement to an auditory-acoustic-phonetic approach. Unfortunately, some law review articles have confused spectrographic or aural-spectrographic approaches with other approaches, and thus have in-part been misdirected (D. Ormerod "Sounding out expert voice identification" (2002) Crim. L.R. 771; C. Singh "Quis custodiet ipsos custodes? Should justice beware: A review of voice identification evidence in light of advances in biometric voice identification technology" (2013) 11 Int. Comment. Evid. 1).

[5] *Criminal Practice Directions* [2015] EWCA Crim 1567 Consolidated with Amendment No.2 [2016] EWCA Crim 1714 at para.19A. The current wording of the section of interest was first introduced in [2014] EWCA Crim 1569 at para.33A.

[6] CPD, para.19A.4 quoting from *Dlugosz et al* [2013] EWCA Crim 2; [2013] 1 Cr. App. R. 32 (p.425) (emphasis added).

> "factors which the court may take into account in determining the reliability of expert opinion, and especially of expert scientific opinion, include: … the extent and quality of the data on which the expert's opinion is based, … whether the opinion properly explains how safe or unsafe the inference is (whether by reference to statistical significance or in other appropriate terms); [and whether the] examination, technique, method or process … was … properly carried out or applied, [and] appropriate for use in the particular case"

We think that these factors can be most clearly addressed by a transparent implementation of the automatic approach, and that they would be difficult to address using the older more subjective approaches. An implementation of the automatic approach and the data used can be described in sufficient detail that another suitably qualified practitioner can replicate what was done. In contrast, the actual process of the formation of a subjective judgment is not open to inspection, and is more susceptible to cognitive bias. Cognitive bias is of increasing concern in forensic science.[7]

From a scientific perspective, we believe that the most important CPD paras 19A.5–19A.6 criteria for determining whether proffered testimony has "a sufficiently reliable scientific basis" are

> "whether the opinion takes proper account of matters, such as the degree of precision or margin of uncertainty, affecting the accuracy or reliability of [the] results; … [and whether it has] been subjected to sufficient scrutiny (including, where appropriate, experimental or other testing), [and whether it has stood] up to scrutiny"

From a scientific perspective, the only way to demonstrate how well a forensic analysis system actually works is via empirical testing (empirical validation). As President Obama's Council of Advisors on Science and Technology noted

> "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."[8]

---

[7] See, for example: D.M. Risinger et al. "The Daubert/Kumho implications of observer effects in forensic science: Hidden problems of expectation and suggestion" (2002) 90(1) *California Law Review* 1; B. Found "Deciphering the human condition: The rise of cognitive forensics" (2015) 47 *Australian Journal of Forensic Sciences* 386; National Commission on Forensic Science, *Ensuring that forensic analysis is based upon task-relevant information* (2015).

[8] President's Council of Advisors on Science and Technology, *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods* (2016), p.6, emphasis in original.

The idea is not new; there have been calls from the 1960s onward for the performance of forensic voice comparison systems to be empirically validated under casework conditions.[9]

Empirical validation treats the system as a black box; i.e. empirical validation does not concern itself with how the system works, only with how well it works. To test a forensic voice comparison system, it must accept a pair of voice recordings as input, and it must output a strength of evidence value.[10] The tester, but not the system being tested, must know whether the input is a same-speaker pair of recordings or a different-speaker pair of recordings. The tester assesses how good the output is based on their knowledge of whether the input was a same- or different-speaker pair. The tester presents the system with a large number of same-speaker pairs and a large number of different-speaker pairs, assesses how good the output is for each pair, and then averages how good the system's performance is over all test pairs.

For the results of the tests to be a meaningful assessment of how well the system is expected to perform under the conditions of the case, the test data must be sufficiently representative of the relevant population and sufficiently reflective of the speaking styles and recording conditions in the case, and the number of test pairs must be large enough for the results to be a potentially convincing estimate of the level of actual performance. In the first instance, the forensic practitioner will make a judgement on whether the test data are sufficient, but ultimately the court will either accept or reject that judgement.[11]

We believe that empirical validation should be required, irrespective of the approach. As the Law Commission stated:

> "Whilst in broad terms we agree … that 'forensic analyses which are more objective and whose reliability can be quantitatively demonstrated should be preferred over more subjective analyses for which it is harder to quantify reliability', we also believe that if a subjective analysis can be tested in controlled circumstances, and opinion evidence founded on such an approach can thereby [be] shown to be reliable, there is no reason why such opinion evidence should be excluded."[12]

3. Admissibility of the auditory approach

The auditory approach, (also called the aural approach) is based on listening. Practitioners usually have training in auditory phonetics, which includes learning a phonetic alphabet which allows them to document the details of the speech they hear. The practitioner listens to the known-speaker recording and to the

---

[9] See review in: G.S. Morrison "Distinguishing between forensic science and forensic pseudoscience: Testing of validity and reliability, and approaches to forensic voice comparison" (2014) 54 *Science & Justice* 245.

[10] We use the term "strength of evidence" to refer to the conclusion reached by the forensic practitioner or the forensic analysis system, not the evidential weight assigned by the trier of fact to the forensic practitioner's conclusion.

[11] Testing of the validity and reliability of forensic analysis systems that output likelihood ratios is discussed in greater detail in: G.S. Morrison "Measuring the validity and reliability of forensic likelihood-ratio systems" (2011) 51 *Science & Justice* 91; A. Drygajlo et al. *Methodological guidelines for best practice in forensic semiautomatic and automatic speaker recognition, including guidance on the conduct of proficiency testing and collaborative exercises* (European Network of Forensic Science Institutes, 2015); D. Meuwly, D. Ramos, R. Haraksim "A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation" (2017) 276 *Forensic Science International* 142.

[12] Law Commission, *Expert Evidence in Criminal Proceedings in England and Wales* (2011), para.5.85.

questioned-speaker recording and attempts to identify properties of the speech that are similar and which the practitioner would not expect to be similar if the recordings were of different speakers, and properties of the speech that are different and which the practitioner would not expect to be different if the recordings were of the same speaker. Conclusions are based on the practitioner's subjective judgement.

Robb[13] involved a dispute as to the identity of the speaker on recordings of telephoned ransom demands. A forensic practitioner (B) conducted an auditory forensic voice comparison. The defence argued at voir dire, at trial, and at appeal, that unless auditory techniques were supplemented by acoustic analysis they were "worthless".[14] During cross-examination, the practitioner agreed that the majority of those with expertise in forensic voice comparison were of the opinion that auditory techniques were unreliable unless supplemented and verified by acoustic analysis. In addition: "He had published no material which would allow his methods to be tested or his results checked. He had conducted no experiments or tests on the accuracy of his own conclusions."[15] The appeal was concerned with whether the practitioner's auditory-based testimony should have been admitted.

The appellant cited from the judgment in the Scottish case of *Davie*[16] that the duty of an expert witness

> "is to furnish the Judge or jury with the *necessary scientific criteria for testing the accuracy of their conclusions*, so as to enable the Judge or jury to form their own independent judgment by the application of these criteria to the facts proved in evidence. … the bare *ipse dixit* of a scientist, however eminent, upon the issue in controversy, will normally carry little weight, for it *cannot be tested by cross-examination nor independently appraised*, and the parties have invoked the decision of a judicial tribunal and not an oracular pronouncement by an expert."[17]

The appellant complained that B "had not set out his criteria, so there was no way of testing the accuracy of his conclusions."[18] The court responded that:

> "We do not consider this complaint to be sound. Dr. [B] described the features of the human voice to which he paid attention. He testified that he found no significant difference between the voice on the disputed tapes and the voice on the control tape. Had he found differences he could no doubt have identified the differences."[19]

From a scientific perspective we think the appellant's argument was sound. The Court of Appeal failed to understand, ignored, or dismissed the need for empirical validation. None of what the practitioner was reported to have done constituted a demonstration of how well his implementation of the auditory approach worked under the conditions of the case. Ormerod comments

---

[13] *Robb* (1991) 93 Cr. App. R. 161; [1991] Crim. L.R. 539.
[14] *Robb* (1991) 93 Cr. App. R. 161 at 165.
[15] *Robb* (1991) 93 Cr. App. R. 161 at 165.
[16] *Davie v Edinburgh Magistrates* 1953 S.C. 34; 1953 S.L.T. 54.
[17] *Davie* 1953 S.C. 34 at 40, as quoted (with ellipsis) in *Robb* (1991) 93 Cr. App. R. 161 at 166. We have added emphasis
[18] *Robb* (1991) 93 Cr. App. R. 161 at 166.
[19] *Robb* (1991) 93 Cr. App. R. 161 at 166.

"it is surprising that the evidence was accepted by the Court of Appeal, except of course that the court would have found it difficult to reject it on the basis of a lack of reliability without having to disclose how that assessment of reliability was conducted, which would undermine the very clear denial of any specific scrutiny for admissibility of expert evidence made in that case."[20]

"If English law also required explicitly that the technique was demonstrated to meet a standard of reliability, admissibility would be extremely unlikely."[21]

O'Doherty[22] (a Northern Irish case) involved a dispute as to the identity of a speaker on a recording of a telephone call made to emergency services. A forensic practitioner (M) performed an auditory forensic voice comparison. On appeal M and three other practitioners all agreed that the majority of practitioners used auditory-acoustic-phonetic approaches, and that those who used auditory-only approaches were in the minority. One of the practitioners (N) argued that

"while auditory phonetic analysis is good at telling us whether two samples have the same accent, once it is established that two samples have the same accent, and generally similar voice quality, only quantitative acoustic analysis can go further and come anywhere near to determining whether the two samples of the same accent come from the same individual."[23]

The Northern Ireland Court of Appeal concluded that the conviction was unsafe, that no prosecution should be brought that depended on testimony based on an auditory-only approach, and that acoustic analysis was also necessary.[24]

Ormerod[25] argued that courts in England and Wales should follow the example set in Northern Ireland, but in a postscript in *Flynn*[26] the England and Wales Court of Appeal stated that:

"Nothing in this judgment should be taken as casting doubt on the admissibility of evidence given by properly qualified experts in this field. On the material before us we think it neither possible nor desirable to go as far as the Northern Ireland Court of Criminal Appeal in *O'Doherty* which ruled that auditory analysis evidence given by experts in this field was inadmissible unless supported by expert evidence of acoustic analysis."[27]

Flynn involved earwitness identification, not forensic voice comparison, and the ruling contained no other mention of the admissibility of forensic voice comparison nor material brought before the court relevant to this topic. The Law Commission Report criticized the ruling in *Flynn* as "in line with the [then] current *laissez-faire* approach to the admissibility of expert evidence in criminal proceedings".[28] Despite this criticism, and despite the subsequent revision of the CPD on admissibility of

---

[20] D. Ormerod "Sounding out expert voice identification" [2002] Crim. L.R. 771, 776.
[21] Ormerod "Sounding out expert voice identification" [2002] Crim. L.R. 771, 778.
[22] *O'Doherty* [2002] NICA 20.
[23] *O'Doherty* [2002] NICA 20 at [12].
[24] *O'Doherty* [2002] NICA 20 at [60] listed some exceptions to its ban on auditory-only approaches. Ormerod "Sounding out expert voice identification" [2002] Crim. L.R. 771, 786, however, argued that these exceptions "will not easily withstand challenge."
[25] Ormerod "Sounding out expert voice identification" [2002] Crim. L.R. 771, 774.
[26] *Flynn* [2008] EWCA Crim 970; [2008] 2 Cr. App. R. 20 (p.266).
[27] *Flynn* [2008] EWCA Crim 970; [2008] 2 Cr. App. R. 20 (p.266) at [62].
[28] Law Commission, *Expert Evidence in Criminal Proceedings in England and Wales* (2011), p.82, fn.83.

expert evidence, in *Slade*[29] the England and Wales Court of Appeal uncritically echoed *Robb* and *Flynn*. In *Slade*, testimony based on an auditory-only analysis (by M) had been admitted at trial.

4. Admissibility of the auditory-acoustic-phonetic approach

The acoustic-phonetic approach involves making quantitative measurements of acoustic properties of speech. Practitioners usually have training in acoustic phonetics. The most popular types of measurements are fundamental frequency and formant frequencies of vowels.[30] The former relate to the vibration of the vocal folds and the latter to the resonance properties of the vocal tract. Practitioners may also measure acoustic properties related to consonant sounds, intonation patterns, speaking rate, etc. These measurements could be used as input to statistical models, but common practice in the UK is to make tables or draw plots of the numbers and make a subjective judgement related to whether known- and questioned-speaker recordings were produced by the same speaker or not. Common practice in the UK is also to combine the subjective judgement made on the basis of the acoustic-phonetic analysis with a subjective judgement made on the basis of an auditory analysis, hence we call this an auditory-acoustic-phonetic approach.[31]

In the appeal in *O'Doherty*,[32] both parties adduced testimony based on auditory-acoustic-phonetic analyses. The practitioner called by the appellant (N) observed a number of differences between the voices on the known- and questioned-speaker recordings and concluded that they were produced by different speakers.[33] The practitioner called by the Crown (F) also found differences between the voices on the recordings, but attributed them to differences in speaking style and recording conditions, and concluded that it was "rather more likely than not" that the voice on the questioned-speaker recording was that of the known speaker.[34]

As previously mentioned, the Court of Appeal in *O'Doherty* ruled the auditory-only approach inadmissible, but the auditory-acoustic-phonetic approach admissible. From our perspective, however, both suffer from the same problems: the conclusion as to the strength of evidence is based directly on subjective judgement, which is non-transparent, not replicable, and susceptible to cognitive bias, and the performance of systems based on these approaches are not routinely (and were not in this case) empirically tested under casework conditions. We therefore do not regard the decision in *O'Doherty* as effective in preventing the admission of unreliable testimony: an untested implementation of one subjective approach was ruled inadmissible, but untested implementations of another subjective approach were ruled admissible.

---

[29] *Slade* [2015] EWCA Crim 71 at [143].

[30] E. Gold and J.P. French "International practices in forensic speaker comparison" (2011) 18 *International Journal of Speech, Language and the Law* 143.

[31] For descriptions of auditory and auditory-acoustic-phonetic approaches, see, for example: F. Nolan "Speaker recognition and forensic phonetics" in W.J. Hardcastle and J. Laver *The Handbook of Phonetic Sciences* (Oxford: Blackwell, 1997), pp.744–767; P.J. Rose *Forensic Speaker Identification* (Taylor and Francis, 2002); M. Jessen "Forensic phonetics" (2008) 2 *Language and Linguistics Compass* 671; H. Hollien "An approach to speaker identification" (2016) 61 *Journal of Forensic Sciences* 334.

[32] *O'Doherty* [2002] NICA 20.

[33] *O'Doherty* [2002] NICA 20 at [22].

[34] *O'Doherty* [2002] NICA 20 at [39].

The difference between the conclusions of N and F highlights the problem. Both practitioners had qualifications and experience that the court considered relevant and adequate, and both used the same approach, so how is the court to deal with their contrasting conclusions? Ultimately, each amounts to no more than *ipse dixit*; an assertion without proof. Other than by conducting empirical tests of the performance of each of N's and F's implementations of the auditory-acoustic-phonetic approach under conditions reflecting those of the case under investigation, there is no way to assess the validity of either of their conclusions.

Similarly, Ormerod expressed concerns that testimony based on acoustic approaches "will be too readily admitted … when in fact it is of questionable reliability."[35] His underlying concern was about "the need for adequate reliability in expert voice-identification evidence"[36] irrespective of the particular approach, and he stated that "It is especially important that the expert is subjected to detailed examination on his methodology and as to the error rates, sample size, etc".[37]

5. Admissibility of the automatic approach

The automatic approach is based on signal processing (a branch of engineering). It makes quantitative acoustic measurements of voice recordings and uses those measurements as input to statistical models. The most common type of measurements are mel frequency cepstral coefficients (MFCCs), which capture fine detail about the frequency components of the speech on the recording. There are important roles for the practitioner in determining appropriate questions to ask and selecting appropriate data and appropriate statistical models to answer those questions (to avoid putting garbage in and getting garbage out of the software), but making the acoustic measurements and applying the statistical models is then automatic. A great deal of research in automatic speaker recognition has focussed on developing statistical techniques which compensate for mismatches in speaking styles and recording conditions. Research and development has been driven by attempts to achieve empirically demonstrable improvements in performance.[38]

Usually, the statistical models quantify the probability of obtaining the acoustic properties of the questioned-speaker recording had it been produced by the known-speaker (how similar is the questioned-speaker recording to the known speaker?) versus the probability of obtaining the acoustic properties of the questioned-speaker recording had it been produced by some other speaker selected at random from the relevant population (how typical is the questioned-speaker recording with respect to the relevant population?). The latter requires use of a sample consisting of recordings of speakers from the relevant population. The relevant population will usually be restricted to speakers of a particular sex,

---

[35] Ormerod "Sounding out expert voice identification" [2002] Crim. L.R. 771, 779.
[36] Ormerod "Sounding out expert voice identification" [2002] Crim. L.R. 771, 785.
[37] Ormerod "Sounding out expert voice identification" [2002] Crim. L.R. 771, 783.
[38] For descriptions of the automatic approach, see, for example: D. Meuwly *Reconnaissance de locuteurs en sciences forensiques: L'apport d'une approche automatique* (University of Lausanne PhD dissertation, 2001); J.H.L. Hansen and T. Hasan "Speaker recognition by machines and humans: A tutorial review" (2015, November) *IEEE Signal Processing Magazine* 74–99; E. Enzinger et al., "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" (2016) 56 *Science & Justice* 42; E. Enzinger, "Implementation of forensic voice comparison within the new paradigm for the evaluation of forensic evidence" (University of New South Wales PhD dissertation, 2016).

speaking a particular language, and a particular accent of that language.[39] The practitioner has to choose an appropriate relevant population and obtain a sample of that population. Using the wrong population will bias the results. Also, if the court were not informed of what population the sample represented, the answer could be highly misleading. For example, even if they were different speakers, the probability of the acoustic properties of a male voice on a questioned-speaker recording calculated using a model trained on data from a known-speaker recording of a male speaker would be much higher than the probability calculated using a model trained on data from a sample of female speaker recordings. This would be misleading if the court assumed that the second probability was with respect to a population of male speakers.[40]

Slade[41] involved a dispute as to the identity of speakers on audio recordings made covertly in a car. The recordings were of poor quality, and included engine, traffic, and other noises. The Court of Appeal ruled the convictions of defendants Slade, Pearman, and Baxter unsafe for reasons unrelated to the voice evidence, but chose nonetheless to respond to a submission to adduce fresh forensic voice comparison evidence in part because the court's "views may perhaps in some respects be of relevance for wider purposes".[42] The fresh evidence was a new forensic voice comparison analysis based in-part on an automatic approach. The Court of Appeal did not explicitly reference the CPD, but does appear to have taken the CPD para.19A criteria into account. Of reported cases dealing with the admissibility of forensic voice comparison, the appeal in *Slade* is the only one to postdate the revision of the CPD on admissibility of expert evidence.[43]

The practitioner who conducted the automatic analyses (F, assisted by H) used commercial forensic voice comparison software known as Batvox. An initial analysis was conducted using one version of Batvox, and a later analysis was conducted using a newer version. Based on the initial analysis the evidence was 37–38 times more likely if the questioned-speaker were someone other than Pearman than if he were Pearman. For the subsequent analysis F stated that "We consider that … Pearman and … Slade can be eliminated with an extremely high degree of confidence. This is effectively a categorical statement of elimination".[44] And Baxter "could be eliminated with 'a fairly high degree of confidence'".[45]

[39] G.S. Morrison et al., "Refining the relevant population in forensic voice comparison—A response to Hicks et alli (2015) The importance of distinguishing information from evidence/observations when formulating propositions" (2016) 56 *Science & Justice* 492.

[40] What we have outlined in the previous paragraph is known as the likelihood ratio framework. We have described an empirical version of the framework in the context of the automatic approach to forensic voice comparison, but it can also be combined with other approaches, and can be implemented in a more subjective manner. The logic of the likelihood ratio framework is applicable across different branches of forensic science. The likelihood ratio framework is recommended by many forensic statisticians and relevant organisations as the logically correct framework for the evaluation of forensic evidence. See, for example: Association of Forensic Science Providers "Standards for the formulation of evaluative forensic science expert opinion" (2009) 49 *Science & Justice* 161; C.G.G. Aitken et al., "Expressing evaluative opinions: A position statement" (2011) 51 *Science & Justice* 1; S.M. Willis et al. *ENFSI guideline for evaluative reporting in forensic science* (European Network of Forensic Science Institutes, 2015).

[41] *Slade* [2015] EWCA Crim 71.

[42] *Slade* [2015] EWCA Crim 71 at [122].

[43] The original trial in *Slade* [2015] EWCA Crim 71 predated the revision of the CPD on admissibility of expert evidence.

[44] *Slade* [2015] EWCA Crim 71 at [153].

[45] *Slade* [2015] EWCA Crim 71 at [154]. The first statement is consistent with the likelihood ratio framework, but the latter two statements are not. They are, however, consistent with another framework known as the UK framework (J.P. French and P. Harrison, "Position Statement concerning use of impressionistic likelihood terms in forensic speaker comparison cases" (2007) 14 *International Journal of Speech, Language and the Law* 137–144; J.P French et al. "The UK position statement on forensic speaker comparison: a rejoinder to Rose and Morrison" (2010) 17

F used a reference database of studio-quality recordings of 100 speakers of "standard southern British English" aged 18–25 who were recorded in a simulated police interview setting. From these, Batvox selected the 35 that were most similar to the known-speaker recording according to its in-built algorithm. The Court of Appeal noted that F

> "and his colleague had initially considered that ideally the reference sample should comprise speakers who came (like the appellants) from West Yorkshire; but ultimately they decided that was not necessary because 'the system considered the reference population appropriate for the tasks'."

And H

> "said that there is no regional variation in vocal tracts, and therefore it was not particularly important to know where the persons in the reference sample came from. He also said that their work had been peer-reviewed; and there had been no criticism of the size of the reference population."[46]

In our opinion, the reference sample used in this case was not appropriate. MFCCs do not simply capture information about the physical properties of a speaker's vocal tract, they are also influenced by behaviour and by recording conditions. MFCC values can be influenced by language and accent. Mismatches between the language or accent spoken in the recordings intended to represent the relevant population and the language or accent spoken in the known- and/or questioned-speaker recordings can affect the calculated likelihood ratio value.[47] We would argue that the 100 speakers were not representative of any population that could reasonably be considered relevant for this case.[48] We submit that there

---

International Journal of Speech, Language and the Law 143). The UK framework was associated with the auditory-acoustic-phonetic non-statistical approach, and conclusions were reached "informally via the analyst's experience and general linguistic knowledge rather than formally and quantitatively" (French et al., 2010, at 141). The UK framework was criticised (P.J. Rose and G.S. Morrison "A response to the UK position statement on forensic speaker comparison" (2009) 16 International Journal of Speech, Language and the Law 139; G.S. Morrison, "Forensic voice comparison and the paradigm shift" (2009) 49 Science & Justice 298; G.S. Morrison, "Forensic voice comparison", Ch. 99 in I. Freckelton and H. Selby (eds), Expert Evidence (Sydney: Thomson Reuters, 2010); G.S. Morrison, "Distinguishing between forensic science and forensic pseudoscience: Testing of validity and reliability, and approaches to forensic voice comparison" (2014) 54 Science & Justice 245). By 2015 (but apparently after the Slade appeal) the lead authors of the UK position statement abandoned it in favour of the likelihood ratio framework (a public announcement to this effect was made by Dr Philip Harrison on 7 September 2015 at the Interspeech conference in Dresden, Germany).

[46] Slade [2015] EWCA Crim 71 at [160].

[47] See, for example: G.S. Morrison et al., "Refining the relevant population in forensic voice comparison — A response to Hicks et al (2015) The importance of distinguishing information from evidence/observations when formulating propositions" (2016) 56 Science & Justice 492; E. Enzinger, "Implementation of forensic voice comparison within the new paradigm for the evaluation of forensic evidence" (University of New South Wales PhD dissertation, 2016), Ch.4; A. Misra and J.H.L. Hansen, "Spoken language mismatch in speaker verification: An investigation with NIST-SRE and CRSS Bi-Ling corpora" (2014) Proceedings of the IEEE Workshop on Spoken Language Technology 273; V. Hughes and P. Foulkes, "Regional variation and the definition of the relevant population in likelihood ratio-based forensic voice comparison using cepstral coefficients" (2014) Proceedings of the Speech Science & Technology Conference 24. The results of the latter study were somewhat erratic, but this could be due to sampling variability—for each accent, recordings of only 28 speakers were used for training.

[48] In addition, unless a forensic practitioner can justify the choices made by an algorithm which automatically selects the data to represent the relevant population, we would argue that the practitioner has abdicated their responsibility to select data which they consider sufficiently representative of the relevant population and to communicate how they did this so that the appropriateness of their decisions and actions can be debated before the judge at an admissibility hearing.

were severe problems with "the extent and quality of the data on which the expert's opinion is based",[49] and that the analysis was "based on flawed data".[50]

Further, a priori it seems unlikely that a sample as small as 35 would be sufficiently representative of a relevant population. The results of one empirical study suggest that it would not.[51]

F conducted empirical tests of system performance using covert recordings made of the defendants in other cars. For those recordings, the identities of the speakers were known, and a priori we expect that one could reasonably argue that they were sufficiently similar to the conditions of the questioned-speaker recordings that the results of tests would be meaningful for the case. The Court of Appeal judgment reported that recordings of 27 other speakers were compared with the covert recordings made in the other cars. It provided no information as to the population that these speakers came from, nor the speaking style or recording conditions, thus we cannot judge their appropriateness.

Recordings of 27 speakers would seem to be a relatively small size for a set of test data. Would it be large enough to convince a judge that the results of testing would be meaningful? Whether the test recordings were sufficiently representative of the relevant population and sufficiently reflective of the speaking style and recording conditions in the case is also in question.

For the different-speaker test comparisons, 37% gave likelihood ratios greater than 1, i.e. the evidence was more likely if the same-speaker proposition were true than if the different-speaker proposition were true.[52] The Court of Appeal judgment quoted the F's report as stating that

> "the system obtains the correct result for all same speaker comparisons and for the majority of different speaker comparisons. When it does make an error it is biased towards making false identifications rather than false rejections."[53]

F was reported as saying that because of the potential for the automatic system to make errors if used stand-alone, he used it in combination with auditory and acoustic-phonetic analyses. He was also reported as saying that in conjunction with auditory and acoustic-phonetic analyses he regarded the automatic system as reliable for excluding a suspect's voice, but not reliable for making a positive identification.

Just counting whether a likelihood ratio is greater than or less than 1 in response to each test pair, is not an appropriate way of assessing the performance of a system that outputs likelihood ratios. It is fundamentally at odds with the likelihood ratio framework. The higher the value of the likelihood ratio the greater the support it provides for the same-speaker hypothesis over the different-speaker hypothesis, and the lower the value of the likelihood ratio the greater the support for the different-speaker hypothesis over the same-speaker hypothesis. Thus, if we know that the test pair is a same-speaker pair and the resulting likelihood ratio is a lot lower than 1, this is worse than if it is just a little lower than 1. Likewise, if we

know that the test pair is a different-speaker pair, a likelihood ratio value much greater than 1 is worse than a value just above 1.

Also, if the output of an automatic system is not directly reported, but instead combined with the output of other analyses using other approaches, it is the performance of the output of the combined system (the system that is actually used to evaluate the strength of the evidence in the case) that must be empirically tested, not the performance of one component of that system[54]—one component could work well but the system as a whole work poorly.

We have argued that the training and test data were inappropriate and inadequate. We would therefore argue that this should be sufficient for the court to rule the proffered testimony inadmissible. The data were flawed,[55] thus the method was not properly applied and the conclusions not properly reached,[56] and the method was not subjected to sufficient scrutiny in the form of appropriate and adequate empirical testing.[57]

If a court concluded that the training and test data were appropriate and adequate, then the court would have to consider whether the demonstrated level of performance was good enough to warrant admission.[58]

Although the Court of Appeal in *Slade* did not explicitly reference the CPD on admissibility of expert evidence, its reasoning does seem to have reflected criteria such as "the extent and quality of the data on which the expert's opinion is based",[59] whether it had been "subjected to sufficient scrutiny (including, where appropriate, experimental or other testing)",[60] and "the accuracy or reliability of those results".[61] The Court of Appeal did explicitly reference *R v T*.[62] There has been much criticism of the ruling in *T* by forensic scientists, forensic statisticians, and legal scholars,[63] and even by the Law Commission.[64] A clear (and we believe scientifically legitimate) concern raised in *T*, however, is whether there is a proper statistical foundation for the calculation of a likelihood ratio, including whether appropriate and sufficient data have been used.

The Court of Appeal in *Slade* stated that it had not been "sufficiently convincingly demonstrated in this appeal that a group of 20 or 30 speakers"[65] constituted adequate data (from context, the reference appears to be to the recordings of 35 speakers selected by Batvox to represent the relevant population). Nor did it consider "comparison with [recordings of] the voices of 20 or 30 speakers whose ages and accents may differ substantially from those of the suspect"[66] to be adequate empirical testing. It also did not consider adequate the level of

---

[54] Forensic Science Regulator. *Guidance: Validation* (FSR-G-201 Issue 1), 2014, para.3.3.
[55] CPD, para.19A.6(c).
[56] CPD, paras 19A.6(d) and 19A.6(e).
[57] CPD, para.19A.6(a).
[58] CPD, para.19A.6(a).
[59] CPD, para.19A.5(a).
[60] CPD, para.19A.6(a).
[61] CPD, para.19A.5(c).
[62] *T* [2010] EWCA Crim 2439; [2011] 1 Cr. App. R. 9 (p.85).
[63] For example: C.E.H. Berger et al., "Evidence evaluation: A response to the Court of Appeal judgment in R v T" (2011) 51 *Science & Justice* 43; M. Redmayne et al., "Forensic science evidence in question" [2011] Crim. L.R. 347; B. Robertson et al., "Extending the confusion about Bayes" (2011) 74 Modern L.R. 444; G.S. Morrison "The likelihood-ratio framework and forensic evidence in court: A response to R v T" (2012) 16 E. & P. 1.
[64] Law Commission, *Expert Evidence in Criminal Proceedings in England and Wales* (2011), Law Com. No.325, p.86, fn.94.
[65] *Slade* [2015] EWCA Crim 71 at [178].
[66] *Slade* [2015] EWCA Crim 71 at [178].

performance demonstrated using those test data. It did not accept the argument made by F that, because the system was apparently biased towards false identifications, it was reliable for making exclusions. The Court of Appeal opined

"in a number of respects it seems to us that the evidence ultimately amounts to little more than a bare assertion that the software is so designed as to ensure the right results: with no explanation of how the court can be confident that is so. For example, the selection by the software of the subset of voices from the reference population has not been explained; and no clear reason has been shown why the court should simply accept the assertion that the system has made the best choices. It does not seem to us to be a sufficient answer to this concern to say that it is only proposed that [automatic speaker recognition] should be used in conjunction with other forms of analysis."[67]

The Court ruled the proffered automatic-approach-based testimony inadmissible.[68] The Court indicated that this decision was case specific:

"In view of our overall decision on other grounds of appeal, however, it is neither necessary nor appropriate for us to make any definitive ruling in this case as to whether such evidence can ever be admissible, or as to what the position might be in the future in the light of any further scientific advance."[69]

We agree with the court's reasoning on these matters.[70]

6. Conclusion

The Criminal Practice Directions (CPD) on admissibility of expert evidence establish scientific validity as a prerequisite for admissibility.[71] In the present article we have discussed what we believe would be necessary to demonstrate scientific validity for a forensic voice comparison analysis, i.e. empirical testing of the performance of the forensic voice comparison system using test data which are 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 under investigation. We have argued that the same criteria should be applied irrespective of the approach used to evaluate the strength of the evidence. We have reviewed existing case law on admissibility of forensic voice comparison testimony, which mostly predates the revision of the CPD. We have found the published rulings to be inconsistent in their treatment of different approaches to forensic voice comparison. It remains to be seen whether

---

[67] *Slade* [2015] EWCA Crim 71 at [179].

[68] *Slade* [2015] EWCA Crim 71 at [177] and [183]. The testimony was not admitted as fresh evidence for the purpose of an appeal. The Court of Appeal had already ruled the convictions unsafe for reasons unrelated to the forensic voice comparison testimony.

[69] *Slade* [2015] EWCA Crim 71 at [177].

[70] To us, however, it seems inconsistent that the Court of Appeal in *Slade* [2015] EWCA Crim 71 ruled inadmissible testimony based on an automatic approach which had been subject to (some) empirical testing, but did not perceive a problem with trial testimony based on an implementation of the auditory-only approach which had not been subject to any empirical testing at all. It seems implicit that the Court of Appeal would also not have objected to the trial testimony based on untested implementations of the auditory-acoustic-phonetic non-statistical approach. The Court of Appeal was not asked to rule on the admissibility of the auditory-only and the auditory-acoustic-phonetic based forensic voice comparison testimony presented at trial. It seems clear to us, however, that the trial testimony would not have fared well had been examined with respect to the CPD, para.19A criteria.

[71] CPD, para.19A.

in future cases involving forensic voice comparison testimony (and testimony related to other branches of forensic science) courts in England and Wales will diligently and consistently apply the CPD para.19A admissibility criteria.