

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.19A⁵:

“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.

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 J "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. [D] 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" g o r j cuku

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

Ormerod²⁵ 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.

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

“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.”²⁴⁶

F

F

International Journal of Speech, Language and the Law

F

*International Journal of Speech, Language and the Law
Science & Justice*

F

F

Expert Evidence Sydney Thomson Reuters

F

Science & Justice

Slade

Slade

Science & Justice

Proceedings of the IEEE Workshop on Spoken Language Technology

F

Proceedings of the Speech Science &

Technology Conference

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

⁴⁹ CPD, para.19A.5(a).

⁵⁰ CPD, para.19A.6(c).

⁵¹ D. van der Vloed, “Evaluation of Batvox 4.1 under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01)” (2016) 85 *Speech Communication* 127–130, errata in (2017) 92 *Speech Communication* 23.

⁵² Using the convention of the same-speaker likelihood in the numerator and the different-speaker likelihood in the denominator, the likelihood ratio value reported earlier would have been 1/37 to 1/38.

⁵³ *Slade* [2015] EWCA Crim 71 at [161].

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 system⁵⁴—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.