To appear in *The International Journal of Speech, Language and the Law, Volume 26, number 1.*

International Practices in Forensic Speaker Comparisons: Second Survey

Erica Gold and Peter French

ABSTRACT

A survey relating to current practices in forensic speaker comparison testing was recently undertaken of 39 laboratories and individual practitioners across 23 countries. Questions were organised around a number of themes, including the preliminary assessment and preparation of case materials, the checking of analysts' work, frameworks used for the expression of conclusions, the use of automatic speaker recognition systems, the use of reference populations, and awareness of cognitive bias. Developmental trends in this area of forensic speech science are established by comparing responses to the present survey with those to the authors' earlier survey published in 2011.

KEYWORDS

Forensic speaker comparison, methods, practices, preparation, automatic speaker recognition (ASR) systems, conclusion frameworks, cognitive bias, reference populations, developmental trends.

1.0 INTRODUCTION

There has been a growth of national and international conferences and seminars dedicated to, or including, workshops on, forensic speech science; for example, the annual events mounted by the International Association for Forensic Phonetics and Acoustics, the Germanic Society for Forensic Linguistics, Interspeech, Speaker Odyssey and the Australian Speech Science and Technology Association. However, contributions to these events tend to focus exclusively or predominantly on specific, narrowly defined theoretical or empirical issues, rather than on the exposition of forensic speaker comparison casework practice and procedure (although see French et al, 2017). As a consequence, casework practitioners, while they might belong to professional associations and attend community events can nevertheless remain unaware of the detail of work-a-day procedures used by their colleagues nationally and internationally. A further consequence is that they may have little conception of how their own practices are situated within the wider array of practices in place elsewhere. Our previous survey of the field (Gold and French, 2011), constituted a first attempt to remedy this position and we welcome the further survey by Morrison and his colleagues on behalf of Interpol in this respect (Morrison et al. 2016). However, the rapidly changing and expanding nature of the field, in our view, necessitates a more detailed survey than either of the earlier studies, and one that includes a heavy concentration on recent advances that have previously received little or no attention. We therefore considered it necessary to conduct a new survey that addressed these developments, and, indeed, probed in more depth and detail some of the matters explored in earlier surveys.

Some principal recent developments in the field that have informed our selection of questions in the present survey are set out below.

1.1 Bayesian Reasoning and Likelihood Ratios

There has been a promulgation of Bayesian reasoning frameworks for interpreting the findings arising from speaker comparisons. This development follows a growing awareness within forensic science as a whole of the necessity of considering the results of casework examinations against competing hypotheses that represent the positions taken by the opposing parties to a legal action - prosecution *versus* defence, complainant *versus* respondent. The prevalence of this approach is reflected in the formation of associations and groups with the explicit aim of promoting understanding and information about the application of Bayes to different areas of forensic casework. Examples of this include the *Bayes and the Law* group, based around Queen Mary University of London but including contributors from a wide range of countries across the globe, and seminars on Bayesian probabilistic reasoning mounted at the *Isaac Newton Institute Cambridge Programme on Probability and Statistics in Forensic Science, UK.*

Associated with the increased awareness and acceptance of Bayesian reasoning has been the recommendation that likelihood ratios (LRs) be adopted for the expression of conclusions in casework across the gamut of forensic specialities (Saks and Koehler 2005). The proposal that all forensic speaker comparison conclusions should be formulated using LRs has been widely and vociferously advocated by Rose and Morrison (2009; Morrison 2009) in particular. Again, however, up-to-date information is needed on the extent to which, and the form in which, they have been adopted in casework.

The expression of speaker comparison conclusions in an LR framework not only involves the assessment of how similar the recorded speech samples are, but also how distinctive or typical the features in common are. In other words, one asks what is their frequency of occurrence in the relevant language/accent community? The most precise and reliable assessments of typicality are those that are empirically-based, i.e. arrived at by examination of a database, or reference population (Ross et al. 2016). The questionnaire used in our earlier survey provided some information on this matter, but there was an indication from some participants at least that their ability to consult reference data was curtailed by its availability. Subsequent developments concerning databases, coupled with a possibly growing awareness of the desirability of using this type of material, have left us in the position that we just do not know the present day situation. We therefore felt justified in including questions on the use of reference population data in speaker comparison casework in this follow-up survey.

1.2 Increased Use of Automatic Speaker Recognition Systems

There has been a growth in the deployment of automatic speaker recognition (ASR) systems in recent years. This has been in relation to both commercial security applications, such as banking and insurance, and in the criminal investigation, national security and forensic sectors. The extent to which these systems are applied for evidential as opposed to investigative and intelligence gathering purposes, however, and the details of how they are used internationally, are so far uncharted. Hence they were incorporated into the questionnaire underlying the present study.

1.3 Awareness of Cognitive Bias

'Cognitive bias' refers to a range of ways in which the objectivity of an analyst's judgments can be influenced or compromised by factors arising from earlier stages in the examinations or from external information. This effect in its various forms ('early conclusion bias', 'contextual bias', 'confirmation bias', etc.) has recently become prominent in the general forensic science literature (Dror, 2014; Nakhaeizadeh et al. 2018), and an awareness of ways of minimising this has entered the field of forensic speech science (Rhodes, 2016). However, the degree to which individual practitioners and laboratories have implemented preventative or precautionary measures is unknown. This too, then, has been included here.

1.4 Other Developments

In addition to the principal areas above, there is a broader array of more minor areas that have received attention in both research literature and the various guideline and code of practice documents produced by professional and government bodies, including the UK Forensic Regulator's *Codes of Practice and Conduct* (2017) and the Appendix adapting this specifically to Forensic Speech and Audio Services (2016). These areas include procedures for avoidance of contamination of samples, preparation of the samples for analysis, the use of assistants in the preparatory work, quality control and the checking of findings and conclusions by a second expert. These areas, too, are addressed in the study. Where issues covered in the previous exercise were revisited in the present survey, comparative numbers and/or percentages are provided below.

2.0 THE SURVEY

The current survey was administered online using SurveyGizmo, and data collection was completed in 2017. It comprised 62 questions organised around six topics, namely:

(i) background information about the participants
(ii) practices and procedures for the processing and assessment of recordings prior to analysis
(iii) use of automatic speaker recognition systems (ASRs)
(iv) checking of work by a second analyst
(v) use of reference data and frameworks for expressing conclusions
(vi) procedures for avoiding or minimising analyst bias

Participants were given the option of answering all questions or just some of them.

2.1 Participant Recruitment

Potential participants were contacted principally through lists relating to the European Network of Forensic Science Institutions (ENFSI) Forensic Speech and Audio Analysis Working Group (FSAAWG), the International Association for Forensic Phonetics and Acoustics (IAFPA), professional contacts in the field, and allowing those contacted to send the survey link on to others actively engaged in forensic speaker comparison casework. In total, 39 practising forensic speech scientists agreed to participate. While this is only a small increase in numbers over our previous survey (36), it nevertheless represents a much larger number of nations (see *3.1.1* below).

3.0 RESULTS

The following section provides all the results gathered from the survey. The results are presented in six sub-sections that correspond with the six question categories we presented to participants. Owing to the fact that the number of participants responding to questions varied across the sections, many of the results are, by necessity, presented as percentages.

3.1 Background Information

3.1.1 Countries

Respondents were from the following 23 countries: Australia, Brazil, Canada, China, the Czech Republic, Denmark, Estonia, France, Germany, Hungary, Italy, Latvia, Lithuania, Morocco, the Netherlands, Norway, Romania, Russia, Slovakia, Spain, Sweden, UK and USA. This is a 77% increase in countries over our previous survey.

3.1.2 Place of work

Participants were asked to specify their working arrangements or affiliations. Twenty-seven practised within government laboratories, 7 carried out their work in university departments, 6 worked as individual private practitioners, 2 practised in an independent facility with other staff, and 1 was located in a research institute. Four respondents identified more than one working affiliation. The biggest change here over Gold and French (2011) concerns the number of government laboratories participating in the survey. The number here is more than double the previous study.

3.1.3 Case load

Respondents were asked how many forensic speaker comparison cases they were involved in per year. Case load ranged from 2 to 175 (mean = 40.4; SD = 48). However, they were not asked about the volume of recordings within cases or the total number of cases they have worked on over their career.

3.2 Preliminary Assessment of Samples

Prior to any analysis of recordings, experts were asked about their preliminary routines in assessing the quality and suitability of the material they receive in a given case. Regardless of methodology (ASR versus auditory/acoustic), 94.6% of respondents said they carried out some form of adequacy evaluation of recordings before embarking on analysis.

3.2.1 Factors in the Assessment of Samples

Respondents who stated they did carry out preliminary assessments were asked whether they assessed duration and recording quality. With regard to the former, 57.6% said they had no minimum duration requirements. For those who did apply a minimum duration criterion, the range was from 7 seconds to 30 seconds (mean = 16.4 seconds; SD = 8.4 seconds).

In respect of whether sample quality figured in the assessments, 73% of participants said they did undertake an assessment of the technical quality of recordings (27% said they did not). Of those carrying out quality evaluations, 77.8% measured signal to noise ratio (SNR), 74.1% measured frequency range of recordings and 51.9% measured or took account of other factors. The ‘other factors’ fell into two broad categories. First, some participants specified further technical features and sources of potential impediment, including: digital compression, clipping (signal overload) and other non-linear distortions, reverberation (echo), aliasing, recording gain/level and channel mismatches. Second, there were non-technical factors related to the voices/speech. These included overlapping speech, emotion in voices, presence of voice disguise and the proportion of the phonological repertoire represented in the samples. Finally, of those who carried out preliminary assessments of adequacy, 91.7% responded that they did on occasion reject recordings for analysis. The proportion of cases rejected ranged from 1% to 60% (mean = 20.4%; SD = 15.1%).

3.2.2 Format Preferences

Participants were asked to specify their preferences concerning file formats, sampling rate (SR) and bit depth.

With regard to file format, there was an overwhelming preference for .wav format files (86.5%), with only two respondents expressing a preference for .mp3, one for each of .aac, .wma and one for 'any linear format'. For sampling rate (SR), 93.3% of respondents said they had a preferred sampling rate. Preferences were 44.1 kHz (31.3%), 11.025 kHz (9.4%), 8 kHz (37.5%) and 22.5 kHz (3.1%). 18.8% of those who had a preference added a caveat to the effect of 'you get what you get'. Finally, for bit depth, 93.3% of respondents said they had a preferred bit depth. Preferences were for 32 bit (5.3%), 16 bit (84.2%) and 8 bit (10.5%).

3*.2.3 Processing of Samples*

Prior to analysis, analysts may carry out processing of recordings in order to convert a file format, edit out extraneous material or equalize channels (i.e. render the channel characteristics of the known and questioned recordings comparable to one another by making adjustments to their frequency bandwidths and/or volume). Participants were therefore asked about their sample processing routines.

71.4% of respondents stated that on occasion they converted sound files received to another format. Of those who undertook file conversion, 40% said they only did so when necessary e.g. so that their software could handle the files. We note that conversion may involve codec conversion and/or resampling (change of SR and bit resolution), but we did not explore this further. 97.1 %; of respondents reported that they (or another member of their organisation) undertook editing of the sound files submitted for analysis. 77.8% said that they did the editing themselves, 5.6% stated that it was undertaken by a colleague of equivalent expertise, 5.6% stated that it was carried out by a technical assistant and 11.1% reported that this varied from case to case. All participants said that the editing work was checked. In terms of equalizing channels, 71.4% of respondents reported they carried out sound filtering to achieve channel parity. Of those who did undertake channel equalisation, around half stated they only did so in limited circumstances or that they used both the equalised and non-equalised recordings in the analysis.

3.3 Use of Automatic Speaker Recognition Systems

The next set of questions related to the use of automatic speaker recognition systems. Of those surveyed, 41.2% of respondents stated they used an automatic speaker recognition system (ASR); 58.8% stated they did not. This contrasts with Gold and French (2011) where it was reported that only 17% used an ASR. For those using an automatic system, participants were asked which system they currently worked with. The following responses are provided in Table 1. below, with Batvox being the most prevalent answer. Note that some participants reported using more than one type of ASR.

Table 1: Automatic Speaker Recognition Systems in Use

Product	Number of Users
Batvox	9
Nuance	3
Alize	2
iVocalise	2

Sive	2
MatLab Scripts	1
VoiceNet	1

Follow-up questions were asked of those using a system in casework. (Some of the answers are represented in the previous section (Section 3.2), but for the purposes of examining preliminary processes related only to ASR related casework, responses are considered separately here.)

Of those using an ASR, 78.6% have a system with an inbuilt function for evaluating the adequacy of the samples for analysis. Those systems with such a function expressed the outcomes of the adequacy evaluations in a variety of frameworks, including binary ('yes'/'no') decisions and various gradations of warning. ASRs varied in terms of their minimum sample duration thresholds. The reported minimum duration required for questioned samples ranged from 3 to 20 seconds (mean = 9.8 seconds). For reference samples, minimum durations ranged from 10 to 96 seconds (mean = 43.5 seconds). In terms of minimum technical quality, the SNRs ranged 10dB – 25dB, minimum high frequency values were reported as 3.4 and 4 kHz, minimum sampling rate was reported as 8 kHz by all those who responded, and 16 bit depth was reported by a single participant who included resolution information. Participants reported rejections rates based on the ASR inbuilt evaluation function ranging from as small as 10% to as great as 70% of cases (mean = 30%).

It is important to note that the majority of the ASRs used allow the user to over-ride the inbuilt decision to reject a sample (64.3%). However, one user noted that they never accepted the system evaluation, while four never, or almost never, over-rode it.

Following preliminary analysis questions in relation to ASR, respondents were asked about their use of reference populations. Table 2 provides the responses to the two main reference population questions.

Table 2: Use of reference populations in ASRs

Question	Yes	No
Does ASR have ready made reference populations?	50% (7)	50% (7)
Do you have access to other reference populations?	85.7% (12)	14.3% (2)

For those respondents in Table 2 whose ASRs include a readymade reference population, three said that they always use the supplied reference population. Two respondents never use the reference population provided, while a single respondent noted that they used it roughly half the time. That particular respondent noted that the other half of the time they use their own self-collected reference populations of anonymised recordings drawn from prior casework. Three of the seven respondents with ready made reference populations reported modifying the technical characteristics of their 'off the shelf' reference population in order to emulate casework conditions. Two of those three ASR users specifically indicated that they amended the technical characteristics of the population in terms of the channel or selected a different group of reference speakers depending on language/accent.

Those participants using their own selected reference populations in their ASR analysis were asked to identify the populations. Answers were generally vague. Four ASR users stated they were not allowed to disclose the information. The remaining respondents who did answer the question gave the following information: their own (1 respondent), different languages and varying numbers of speakers (1 respondent), and 4 different reference populations - phone, microphone, male, female – (4 respondents). When asked how they acquired their reference populations, 5 respondents stated they were unable to disclose the information. The remaining 6 had collected their reference populations largely from recording campaigns (e.g. involving volunteers) and also through previous casework. Finally, ASR users were asked to specify the circumstances under which they used their available reference populations. One respondent stated he/she was unable to reveal this while another reported using the same reference population in all cases. The remaining six respondents reported that they did this on a case by case basis, but it was largely to match the accent and channel conditions. One respondent was specific in noting that they chose the reference population that best matched the reference (suspect) recording.

ASR users were asked about whether they calibrated their systems, and whether, before undertaking the testing on the actual recordings in question, they first carried out pilot tests using similar samples in order to establish whether the system could perform reliably with material of that type. The questions and responses are reported below in Table 3.

Table 3: Testing and Calibration of ASRs

Question	Yes	No	Sometimes
Do you test ASR in relevant field conditions?	35.7% (5)	42.9% (6)	21.4% (3)
Do you calibrate your system?	57.1% (8)	42.9% (6)	

For those using an ASR, participants were asked whether their methodology included the use of human analysis alongside the ASR evaluation (e.g. acoustic and/or auditory-phonetic analysis). As with Gold and French (2011), all experts again reported the use of human analysis in conjunction with the ASR evaluation. Half the participants stated that this was carried out by phoneticians and half stated it was undertaken by engineers. Thirteen (92.1%) reported the human element as comprising both auditory and acoustic analysis, while one reported that it consisted of acoustic examinations only.

Finally, a number of participants (12) chose to comment on combining the results of their ASR analyses with those from auditory and/or acoustic analysis. Some were non-specific as to how this was done (e.g. cross-verification, complementing and confirmation were mentioned but not elaborated on). However, one reported never combining but presenting the 2 sets of results independently. Three reported combining the results and expressing an overall, single integrated conclusion on a classical probability or other, verbal scale. 64.3% of ASR users indicated that they provided numerical LRs at some point in their case reports (35.7% did not).

3.4 Checking Work and Expression of Conclusions

This section considered the checking of casework findings and conclusions, regardless of whether an expert used an ASR or not.

3.4.1 Checking

71.4% of the respondents reported that their casework was checked and agreed upon by another analyst. 20% stated that they did not have their work checked and 8.6% stated that this sometimes happened. Responses to what exactly was checked are summarized in Table 4 below.

Table 4: Aspects of Case Reports Checked by another Analyst

Aspects Checked	% of Participants
Key findings	69.2%
Conclusion	69.2%
Readability of report	15.4%
Methods	11.5%
Everything	19.2%

Table 4 indicates that most experts have their key findings and conclusions checked. However, only around 1 in 5 experts have all elements of their report checked. For the 88.4%[1] of experts who have their conclusions checked, 52% said that the checker arrives at his/her own conclusion based on the first analyst's checked findings. 16% of experts said that the analyst and checker conducted independent analyses (i.e. both completed separate comparisons for the same case). Another 16% of experts reported that the checker saw the analyst's conclusion before arriving at their own. The final 16% of experts equally responded with "other" (i.e. the checker does something different), and that what the checker did depended entirely on the case at hand. Participants who had their work checked reported that any disparity between the analyst's and checker's (or second analyst's) conclusions was normally in relation to the *strength* of the evidence, rather than the 'polarity', and typically only different by a single scalar degree.

3.4.2 Expression of Conclusions

All participants, regardless of methodology, were asked to identify the conclusion framework that they employed to express their interpretation of the findings in the case. They were given 6 frameworks to select from: binary, classical probability, UK Position Statement, support statement, verbal likelihood ratio and numerical likelihood ratio. For information about each of these, apart from the support statement, see Gold and French (2011). The support statement was set out as a way of evaluating forensic scientific evidence by Champod and Evett (2000) as a means of presenting a conclusion in a simpler and more readily digestible form for lawyers and triers of fact than the more complexly formulated likelihood ratio. It was adopted in place of the UK Position Statement by most UK forensic phoneticians in 2014. This followed from a transfer to support statement framework by the laboratory of the second author in January of that year, and his forwarding a reporting template and explanatory letter to other UK practitioners. The scale of support statements and its correspondences with both verbal and numerical likelihood ratios is set out in French (2017).

Table 5 provides the breakdown of responses from the 34 of the 39 who identified which framework they used. (The number in parentheses next to a percentage indicates the number of responses that were deemed as having fit into a given category based on the description provided by the respondent.) For comparison purposes, percentages from our previous survey are also included. The present survey adds up to just over 100% as two participants identified more than one framework (38 different respondents provided a total of 40 responses). The 2011 survey adds up to just below 100% as there was a single participant who had their own version of a conclusion framework that did not fall within any of the categories listed below.

Table 5: Conclusion Frameworks in Use

[1] Although it might appear from Table 4 that only 69.2% have their conclusions checked, this figure must be added to the 19.2% who stated that they had everything checked.

Conclusion Framework	% of Participants Present Survey	% of Participants 2011 Survey
Binary	5.3%	5.7%
Classical probability	23.7%	40%
UK Position Statement	13.2%	31.4%
Support Statement	23.7%	0%
Verbal Likelihood Ratio	26.3%	11.4%
Numerical Likelihood Ratio	13.2%	8.6%

Nearly all experts only ever use a single conclusion framework for speaker comparison casework. However, two respondents reported using different conclusion frameworks in accordance with the needs of the particular case.

3.5 Reference Data

This section concerns the use of reference data in forensic speaker comparison casework. All respondents were asked about their reference data habits irrespective of whether they used an ASR system. They were asked whether they made use of any reference data when making typicality assessments. 68.8% stated that they used some form of reference data while the remainder did not. These percentages are not materially different from those reported in our 2011 survey. The majority who responded positively made it clear that they only used reference data when it was available, and in most cases they found that it was not. A number of participants in our 2011 survey also remarked on the scarcity of useable background data.

When asked from where the reference data was taken, 19 reported drawing data from public sources, 17 had self-collected data for either a case or for research purposes, 2 used old cases, 1 used (unspecified) unpublished data, and 1 used data bought from a commercial source. For those making use of public sources (or published literature), the following were identified:

Table 6: Specific sources identified as containing reference populations or population statistics

Sources Used for Population Statistics	
Source	**Number of Mentions**
Articulation rate (Gonçalves 2014, 2017)	1
Articulation rate (Gold 2012[2], 2014)	2
Brazilian University Corpus	1
Experiments/research	2
F0 (Hudson et al. 2007)	2
Hesitation markers	1
Lithuanian Voice Database (BALSAS LTv)	1
Personal databases	2
Scientific publications/literature	7
Speech Technology Centre Database (Russia)	1
Various Sources	1
2010 NIST Speaker Recognition Workshop RSR2015 STC Database	2

Some of the items indicated in Table 6 are rather general indications of resources used (e.g. scientific publication), while others are very specific citations (e.g. Hudson et al. 2007). Any entry that has a

[2] See Gold (2018) for an updated version of Gold (2012).

multiplier following it indicates that it was mentioned more than once and how many times. For those entries that have a dated reference provided, a specific phonetic variable has already been analysed and reference data for that parameter can be found in the publication. Table 6 is not an exhaustive or even representative list of reference data available to the forensic speaker comparison community. Rather the entries are simply a sampling of reference data and more than likely the data sources that are accessed more frequently. The most noticeable change since the 2011 survey concerns the fact that analysts who do use reference data are now using a much wider set of sources. Only 4 sources were reported in the earlier study.

3.6 Cognitive Bias

The final section of survey considered cognitive bias and the role awareness of it may play in forensic casework procedures. Traditionally, cognitive bias has not figured highly in the awareness of the forensic speech science community. However, recent work by, for example Fraser (Fraser, 2018; French and Fraser, 2018) and Nakhaeizadeh et al. (2018) has shown that cognitive bias can affect the outcomes of the analysis. Cognitive bias can be considered at simple levels such as which recording is examined first (e.g. questioned or known) to those levels that are more obvious like what additional – potentially ‘steering’ - information is given to an expert when instructed to carry out a forensic speaker comparison case.

For any given case, participants were asked whether they examine the questioned (evidential) recording or the known (suspect) recording first. The relevance of this is that the known recording is normally a better quality, longer and therefore richer source of individual speech data than the questioned sample, and examination of this first can prime the analyst to look for, and find, features in the questioned sample that might not otherwise have been apparent. 55.9% of experts stated that they examine the questioned recording first; 11.8% examine the known recording. 14.7% reported that they examined the poorest quality first, and 5.9% each examine the best quality first, that they had no rules and/or went back and forth between the questioned and known recordings. Many analysts typically have strong opinions on this matter.

Practices relating to the background information made available to experts have received little previous attention. For some of the experts in our survey, the information they received is governed by local jurisdictional rules and regulations. Responses concerning what they receive are presented in Table 7 in descending order of frequency of reporting (participants could provide multiple responses).

Table 7: Background Information Provided to Analysts

Background Information Received	
Information	**Number of Mentions**
Basic information about case/suspect (e.g. age, name, place of residence)	11
Varies from case to case	10
The task being asked of the expert (including deadline to submit)	8
Almost nothing or nothing	8
Crime type	4
Recording device characteristics	4
Site and date of offense	4
Other evidence from the case	2
Previous convictions/charges of suspect	2
Desired outcome (due to wording on standard forensic lab submission forms)	1
Emotive information (due to wording on standard forensic lab submission	1

forms)	
MG21 form (particular to the UK; description of case and other potential evidence)	1
Police officer's opinion (due to wording on standard forensic lab submission forms)	1
Police report	1
Relationship between the unknown and known speakers	1

Respondents were asked if they found certain background information helpful, and if so what specifically. 64.5% of respondents said that some background information was helpful, while 29% that no background information was helpful, and an additional 6% said that it depended on the case. For those who found some background information helpful, they specifically identified the information in Table 8. Responses concerning what they receive are presented in Table 8 in descending order of frequency of reporting (participants could provide multiple responses).

Table 8: Background Information that Analysts find Helpful

Helpful Background Information	
Information	**Number of Mentions**
Technical information concerning recordings (including sampling rate and bit depth)	9
Age of suspect	4
Geographical movement/mobility of suspect	2
Languages spoken by suspect	2
Physical or emotional condition of suspect	2
Time span between recordings	2
Basic information relating to the results of the investigation	1
If any close relatives are also involved in the case	1
Information related to the perpetrator's population/defence hypothesis	1
Information the suspect provides in the recording	1
Location of offense	1
Suspect's education level	1
Suspect's ethnicity	1
Suspect's gender	1
Whether the suspect is missing teeth	1
Why something was used as reference material	1

Note that some participants who responded that having no background information was helpful in a case did not consider technical characteristics of the recording to be included in this requirement.

Those who considered certain background information liable to bias their examination were asked to identify any or all such information. A list of responses from the 30 who responded is provided in Table 9. Responses are presented in descending order from most often to least cited, with the number of responses indicated next to them.

Table 9: Background Information that Analysts Consider Potentially Biasing

Potentially Biasing Background Information	
Information	**Number of Mentions**
Everything	7
Other pieces of evidence	7
Information regarding the crime	4
Content of questioned /known samples	3

"I do not know"	3
Information about victims and/or if the suspect caused damage	3
Police officers' opinions	3
Personal information about suspect	2
None	2
Content of police interviews (including the behaviour of the suspect)	1
Errors in the names of those involved	1
Knowing which is the reference recording and which is the questioned recording	1
Language profile of the known speaker (i.e. non-native speaker)	1
Media reports about the case	1
MG21 Form (UK)	1
Previous voice comparisons	1
Reasons why the police suspect the person in custody	1
Recording quality	1
Records of court hearings	1

Responses were extremely varied in relation to what experts thought could be potentially biasing in their casework. Some experts were not aware of anything in specific that might affect their work, while other respondents were hyper-aware and thought that everything could potentially affect their analysis. Although this survey question was intended to consider forensic speaker comparison casework, there were a few responses from participants that indicated that they were more aware of potential bias when carrying out transcription work rather than speaker comparisons.

Given that some experts seem to be very aware of potential biasing factors, we were interested in whether any participants had implemented any procedures in an effort to shield themselves or their colleagues from biasing information. 36.4% stated that they had some form of protocol or procedure in place, while the majority did not. When specifically asked for details about any protocols or procedures in place for shielding bias, two respondents noted that they were currently working on the development of these. Respondents who already had protocols in place said that they carried out things such as: blind groupings of samples (see Cambier-Langeveld 2007), limiting access to investigation details, analysing the poorer quality sample first, linear sequential unmasking of information (Dror 2011, 2014; Dror and Cole 2010), two analyses by separate experts, and explicit identification of alternative hypotheses.

4.0 DISCUSSION

The present survey had two main purposes. First, it sought to update information about forensic speaker comparison practices established in the earlier survey we carried out in 2011. Second, it attempted to explore areas that have risen to prominence in the literature on forensic science generally, and which have begun to enter into the work-a-day methods and procedures of forensic speech scientists. In the latter respect, it cuts much new ground and investigates in detail many issues that have received only cursory attention elsewhere.

In pursuing these aims, a number of general trends and themes have emerged. First, there is a growing awareness of the forensic speaker comparison expert as a *human* system. This has several aspects. One concerns an acknowledgement of the need to have outcomes, results and conclusions checked by a second expert. We have seen this in relation to both preparatory work and the main analyses and comparisons. We find, for example, that all respondents who had preparatory editing of speech samples undertaken by colleagues or assistants checked their work before progressing to the main analysis. Also, in respect of the main comparisons, just less than 90% of experts had their conclusions checked by a second expert and of those more than half reported that the checker would first assess the case records and samples and arrive at their own independent conclusion before seeing that of the primary analyst. Whilst marking a growth of consciousness of the potential fallibility of the expert, we see these developments in an entirely positive light; they constitute important safeguards against erroneous results.

A further recognition of the limitations of human analysis systems concerns the awareness of the potential for cognitive bias, i.e. the potential for background information of various sorts entering, albeit unconsciously, into the analyst's mind-set and influencing the outcomes of the comparisons. While still in a minority, encouragingly, almost one third of participants put forward the view that it was helpful for the analyst to be provided with no background information about the case.

There has been some movement since our 2011 survey with respect to giving automatic systems a role in the analysis process, although, as with the earlier investigation, no respondent reported using an ASR system without some other form of analysis. Whilst this might in part be ascribed to the wider availability of the technology and to its increased reliability and accuracy achieved through development of i-vector and x-vector processing (see e.g. Kelly, 2018), it nonetheless also represents an attempt to move the analyses away from reliance on a human system alone. Again, we have seen encouraging figures concerning the testing of ASR systems in field conditions relevant to casework prior to undertaking the actual evidential tests.

A further marker of the movement away from dependence on the individual analyst and his/her experience in interpreting case data concerns the findings on the use of reference data. While the percentage of analysts using this has remained relatively stable at c. 70% since the 2011 survey, those analysts who do use it are exhibiting more diversity in the corpora and sources consulted. This may be due in part to a multiplication of data sets and relevant publications since the 2011 survey. This once more signifies a movement towards putting the comparisons on a more objective and empirical footing.

In addition to the various attempts to introduce checks and balances and move the analyses onto more objective bases, we would, finally, mention the shifts we have seen between 2011 and the present survey with respect to conclusion frameworks. In particular, the very appreciable gravitation to Bayesian likelihood ratio frameworks – verbal, numerical and the associated support statements – has followed a 'paradigm shift' (Saks and Koehler 2005) in forensic science more generally. The movement to a fully numerical likelihood ratio for expression of overall conclusions, appears a remotely distant prospect, owing to the need for manifold reference populations representing the gamut of social, regional and ethnic varieties of each language to be analysed, together with the limited but variable and unknown 'shelf-life' of the data brought about by accent-specific rates of change. We would, however, end on a positive note, in that future research may consider the use of more generic population data - e.g. 'general' British English may be established as representative enough of the majority of British English accents and statistical techniques could be applied to calculate tolerances, sensitivity measures and safety margins for cases under analysis (French et al, 2018; Gold et al. 2018).

Acknowledgements

The authors would like to thank all the survey participants for their insightful responses and making this article possible. Thanks goes to the helpful feedback provided by two anonymous reviewers and Dr Michael Jessen. Finally, we would also like to thank the School of Music, Humanities and Media at the University of Huddersfield for supporting this project with an Early Career Project Grant.

References

Cambier-Langeveld, T. (2007) Current methods in forensic speaker identification: Results of a collaborative exercise. *International Journal of Speech, Language and the Law* 14(2), 223-243.

Champod, C. & Evett, I. W. (2000). Commentary on: Broeders, A.P.A. (1999) some observations on the use of probability scales in forensic identification. *Forensic Linguistics* 7(2), 239-243.

Dror, I.E. (2011). The paradox of human expertise: Why experts can get it wrong. In N. Kapur (Ed.), *The Paradoxical Brain*. Cambridge, UK: Cambridge University Press, 177-183.

Dror, I.E. (2014) Practical solutions to cognitive and human factor challenges in forensic science. *Forensic Science Policy & Management: An International Journal* 4, 105-113.

Dror, I. E. & Cole, S. (2010) The vision in 'blind' justice: Expert perception, judgment and visual cognition in forensic pattern recognition. *Psychonomic Bulletin & Review*, 17(2), 161-167.

Forensic Regulator (2016) *Codes of Practice and Conduct: Appendix for Speech and Audio Forensic Services.* FSR-C-134, 1.

Forensic Regulator (2017) *Codes of Practice and Conduct for Forensic Science Providers and Practitioners in the Criminal Justice System,* 4.

Fraser, Helen (2018) "Assisting" listeners to hear words that aren't there: dangers in using police transcripts of indistinct covert recordings. *Australian Journal of Forensic Sciences* 50(2), 129-139.

French, P. (2017) A developmental history of forensic speaker comparison in the UK. *English Phonetics 21*, 255 – 270.

French, P. and Fraser, H. (2018) Why 'ad hoc experts' should not provide transcripts of indistinct forensic audio, and a proposal for a better approach. *Criminal Law Journal* 42(5), 298-302.

French, P., Harrison, P., Kirchhübel, C., Rhodes, R. & Wormald, J. (2017) From receipt of recordings to dispatch of report: Opening the blinds on laboratory practices. Presentation at the *26th Annual Conference of the International Association of Forensic Phonetics and Acoustics*, Split, Croatia, 9-12 July 2017.

French, P., Harrison, P., Hughes, V., Watt, D., Llamas, C. and Braun, A. (2018) Comparing apples with apples, apples with oranges and apples with apples and oranges: The effects of (mis)matching reference population accents in ASR speaker comparisons. Presentation at the *27th Annual Conference of the International Association of Forensic Phonetics and Acoustics*, University of Huddersfield, UK, 29 July – 1st August 2018.

Gold, E. (2012) Articulation rate as a discriminant in forensic speaker comparisons. *UNSW Forensic Speech Science Conference*. Sydney, Australia, 3 December 2012.

Gold, E. (2014) *Calculating likelihood ratios for forensic speaker comparisons using phonetic and linguistic parameters*. Unpublished PhD Dissertation, University of York.

Gold, E. (2018) Articulation rate as a speaker discriminant in British English. *Proceedings of Interspeech*, Hyderabad, India, 2-6 September 2018, 1828-1832.

Gold, E. and French, P (2011) International practices in forensic speaker comparison, *International Journal of Speech, Language and the Law* 18(2), 293-307.

Gold, E., Ross, S., Earnshaw, K. (2018). The 'West Yorkshire Regional English Database': investigations into the generalizability of reference populations for forensic speaker comparison casework. *Proceedings of Interspeech, Hyderabad, India, 2-6 September 2018,* 2748-2752.

Gonçalves, C. S. (2014) Taxa de Elocução e de Articulação em Corpus Forense do Português Brasileiro. *Language and Law/Linguagem e Direito* 1 (2), 114-116.

Gonçalves, C. S. (2017) Taxa de elocução e taxa de articulação em *corpus* utilizado na perícia de Comparação de Locutores. *Letras de Hoje*, Porto Alegre 52 (1), 15-25.

Hudson, T., de Jong, G., McDougall, K., Harrison P., and Nolan F. (2007) F0 Statistics for 100 Young Male Speakers of Standard Southern British English, In J. Trouvain and W. Barry (eds.), *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, 6-10 August 2007, 1809-1812.

Jessen, M. (2007) Forensic reference data on articulation rate in German. *Science and Justice* (47), 50–67.

Kelly, F. (2018) The future of the field: Automatic speaker recognition. Presentation at the *27th Annual Conference of the International Association of Forensic Phonetics and Acoustics*. Huddersfield, UK, 29 July – 1 August.

Morrison, G.S. (2009) Forensic voice comparison and the paradigm shift. *Science & Justice* 49, 298–308.

Morrison, G.S., Sahito, F.H., Jardine, G., Djokic, D., Clavet, S., Berghs, S., Goemans Dorny, C. (2016). INTERPOL survey of the use of speaker identification by law enforcement agencies. *Forensic Science International*, 263, 92–100.

Nakhaeizadeh, S., Morgan, R.M., Rando, C., and Dror, I.E. (2018) Cascading bias of initial exposure to information at the crime scene to the subsequent evaluation of skeletal remains. *Journal of Forensic Sciences* 63 (2), 403-411.

Rhodes, R. (2016) Cognitive bias in forensic speech science: risks and proposed safeguards. Presentation at the *25th Annual Conference of the International Association of Forensic Phonetics and Acoustics*, York, UK, 24-27 July 2016.

Rose, P. and Morrison, G.S. (2009). A response to the UK Position Statement on forensic speaker comparison. *International Journal of Speech Language and the Law* 16(1), 139 – 163.

Ross, S., French, P. and Foulkes, P. (2016) UK Practitioners' Estimates of the Distribution of Speech Variants. Presentation at the *25th Annual Conference of the International Association of Forensic Phonetics and Acoustics*, York, UK, 24-27 July 2016.

Saks, M.J., Koehler, J.J. (2005) The coming paradigm shift in forensic identification science. *Science* 309, 892–895.