Science & Justice 63 (2023) 251–257

Contents lists available at ScienceDirect

Science & Justice

journal homepage: www.elsevier.com/locate/scijus

Short Communication

What does method validation look like for forensic voice comparison by a human expert?

Christin Kirchhübel [a], Georgina Brown [a,b,*], Paul Foulkes [c]

[a] Soundscape Voice Evidence, Lancaster, UK
[b] Department of Linguistics and English Language, Lancaster University, UK
[c] Department of Language and Linguistic Science, University of York, UK

ARTICLE INFO

Keywords:
Forensic voice comparison
Method validation
Competency testing

ABSTRACT

Method validation has gained traction within forensic speech science. The community recognises the need to demonstrate that the analysis methods used are valid, but finding a way to do so has been more straightforward for some analysis methods than for others. This article addresses the issue of method validation for the Auditory Phonetic and Acoustic (AuPhA) approach to forensic voice comparison. Although it is possible to take inspiration from general regulatory guidance on method validation, it is clear that these cannot be transposed on to all forensic analysis methods with the same degree of success. Particularly with respect to an analysis method like AuPhA, and in a field of the size and characteristics of forensic speech science, a bespoke approach to method validation is required. In this article we address the discussions that have been taking place around method validation, and illustrate one possible solution to demonstrating the validity of voice comparison by a human expert using the AuPhA method. In doing so we consider the constraints placed on sole practitioners, which generally go unacknowledged.

1. Introduction

In this article we address the issue of method validation for forensic voice comparison by a human expert. Method validation is of importance in many fields, but forensic voice comparison poses particular problems. Moreover, ongoing regulatory changes to forensic science provision in the UK mean that a solution to the problem is a pressing concern. In view of this, a key purpose of this paper is to lay out a realistic approach to validation for forensic voice comparison by a human expert.

The provision of forensic science services to the Criminal Justice System in England and Wales is regulated by the Forensic Science Regulator (FSR). The core task carried out by forensic speech practitioners, i.e., voice comparison analysis, falls within the remit of the FSR. Voice (or speaker) comparison analysis involves the comparison of recorded voices (e.g., telephone recordings, covertly recorded material, police interviews with suspects) to assist the court in deciding whether they come from the same or different speakers. There are two main accepted methodologies for carrying out voice comparison analysis. One is known as *Auditory-Phonetic and Acoustic* (AuPhA) analysis by a human expert, and the other is voice comparison by means of an automatic speaker recognition system (usually complemented by some measure of human analysis) (see [1] for a more detailed discussion). At the time of writing, AuPhA analysis is the only admissible approach in UK jurisdictions for voice comparison analysis.

The Forensic Regulator Act 2021 [2] places the Regulator on a statutory footing, and the Regulator is currently in the process of producing a statutory code of practice and conduct (the "Code") [3]. Although only a draft of the Code is available at present, there is no doubt that certain provisions contained within this draft are going to remain part of future published versions. This includes the requirement for *method validation.* The importance of working with valid methods is also recognised in secondary legislation applicable in England and Wales. Part 19 of the Criminal Procedure Rules and the associated Criminal Practice Directions (CrimPD) [4] emphasise the court's active role in the assessment of the reliability of expert opinion evidence. CrimPD V 19A.5 [5] lists factors which the court may take into account in making such assessments and these include: "*the extent and quality of the data on which the expert's opinion is based, and the validity of the methods by which they were obtained*".

Method validation means demonstrating that a method achieves what it is claimed to achieve. While there is widespread agreement on

* Corresponding author at: Department of Linguistics and English Language County South Building Lancaster University Lancaster LA1 4YL, UK.
E-mail addresses: ck@soundscapevoice.com (C. Kirchhübel), g.brown5@lancaster.ac.uk (G. Brown), paul.foulkes@york.ac.uk (P. Foulkes).

https://doi.org/10.1016/j.scijus.2023.01.004
Received 7 October 2022; Received in revised form 19 January 2023; Accepted 23 January 2023
Available online 26 January 2023

the importance of method validation, establishing what method validation involves in the context of voice comparison using the AuPhA approach has, so far, been a problem. The Code includes broad directions on how method validation *should* be performed, but naturally, the Code does not address what method validation *actually* involves for the AuPhA approach. This is, in part, because the Code targets a broad range of forensic science disciplines. Therefore, it comes down to the forensic speech science (FSS) community to transpose the provisions contained in the Code onto this specific area. The FSS community has engaged with this topic for several years now, and a number of ideas have been put forward about what method validation means for AuPhA voice comparison. Despite having devoted thought to the issue, it is fair to say that, in the UK at least, the FSS community has been struggling to arrive at an approach to AuPhA method validation that is both theoretically cogent and achievable in practice.

Working towards a solution to validating the automatic approach to voice comparison (as opposed to AuPhA) has been less problematic. [6] present a consensus as to how method validation can be achieved in this context. However, it would not be possible to simply adopt the recommendations made in that paper for the AuPhA approach, as the two methods of analyses, i.e., automatic vs AuPhA, are not sufficiently comparable [1: 245].

One problem with validating AuPhA voice comparison may originate from the way that AuPhA has been conceptualised for the purpose of method validation within the FSS community, i.e., a conceptualisation that separates the 'method' from the 'analyst' and thereby treats them as independent components. As we go on to explain, this conceptualisation and resulting approach to method validation is a key source of the problem for AuPhA voice comparison. Further, adopting such a conceptualisation also means that following the FSR guidance on method validation is near-impossible. We argue that, for the purpose of validating AuPhA voice comparison, the method and the analyst are inseparable. Therefore, a solution to AuPhA method validation can be found in competency testing. Reassuringly, recent discussions among members of the UK FSS community (2021–2022) suggest that views are starting to converge towards this latter conceptualisation and corresponding route to method validation.

In light of these developments, in this paper we aim to consolidate the discussions that have been taking place and share experience of a possible approach to validating AuPhA voice comparison through competency testing. Section 2 provides an overview of AuPhA voice comparison, emphasising that all aspects of AuPhA depend to a large extent on the analyst. Section 3 further justifies why it is competency testing that could be used to validate AuPhA voice comparison. Sections 4 and 5 present and discuss an example of how competency testing was implemented by the authors on behalf of *Soundscape Voice Evidence*, a small forensic speech analysis provider. The approach aims to enable any practitioner to be able to integrate competency testing into their practice, irrespective of whether they are a sole practitioner, or part of a large laboratory. Section 6 offers a candid discussion and ideas around how competency testing could be integrated into the wider FSS community.

2. Auditory-Phonetic and Acoustic (AuPhA) voice comparison

AuPhA voice comparison involves two types of analysis which are carried out in parallel: (i) auditory-phonetic and (ii) acoustic analyses. Auditory-phonetic analysis involves repeated and detailed analytical listening, while acoustic analysis involves using software to visualise and measure aspects of the speech signal. In applying the AuPhA approach, the analyst makes qualitative and quantitative observations about a range of different voice and speech features in the speech samples under analysis. It is not possible to give an exhaustive list of features one might examine, but some examples of the features that are commonly analysed include vowel and consonant sounds, voice quality (timbre), voice pitch, speech rhythm, (dis)fluency, speech tempo, and lexical/grammatical usage (see, for example, [1;7] for more detail). Many of the frameworks used to analyse the features in AuPhA are largely uncontroversial and are adopted in many branches of phonetics and linguistics, e.g., the International Phonetic Alphabet. There are also some frameworks that have emerged for specific use within FSS, e.g., a modified version of the Vocal Profile Analysis protocol [8].

Although field-specific literature has given descriptions of frameworks and features that are involved in AuPhA voice comparison, there has been insufficient acknowledgement of the extent of decision making that takes place when undertaking a comparison. Although some features are routinely analysed, e.g., voice quality and vowel sounds, ultimately it is the analyst who decides which features to analyse, and how. This is done on a case-by-case basis, taking into account the analyst's decisions around the quality, quantity and comparability of the material, and the distinctiveness of the voices. For example, decision making at the feature level can be illustrated using measurements of fundamental frequency (f0), the acoustic correlate of voice pitch, which is usually expressed numerically in Hertz (Hz). First, the analyst decides whether the recordings are suitable for f0 analysis (it might not be possible in cases of poor technical quality of the recording, for example). Provided that they are, the analyst then decides which parts of the speech to include in the f0 analysis. It is typical to encounter recordings which contain speech of varying levels of vocal effort (e.g., whisper, neutral, raised voice, shouted, etc), which impacts on resulting f0 measurements. Therefore, in order to make meaningful f0 comparisons, the analyst controls for voice level within and across recordings by selecting appropriate and comparable sections of the speech. Once this has been done, f0 analysis is conducted via specialist software. At this stage, several more decisions need to be made, for example in the choice of software, settings within the software, and the specific measurements taken (e.g., mean, median, range). While the f0 analysis results are seemingly objective, then, they are heavily dependent on the decisions that the analyst has made. Similar decision making applies to all the features included in an AuPhA analysis, whether qualitative or quantitative.

Based on the overall collection of findings, the analyst then assesses the nature of the similarities and differences between the speech samples under comparison. Irrespective of whether two recordings contain the same speaker or different speakers, it is inevitable that there will be both similarities and differences in the observed voice and speech features. On the one hand the voices are likely to be similar enough for an untrained ear (such as that of an investigating police officer) for the case to be brought, and are at least likely to contain the same language. On the other hand, no two samples of speech are identical, because speech is a dynamic process and every speech event is unique. Speech contains no permanent feature and as such is not a biometric akin to fingerprints or DNA, for example.

In relation to the similarities, it is for the analyst to determine how closely matching and how unusual they are. In relation to the differences, it is for the analyst to determine whether these are more likely to be a result of the differences in recording context or the fact that different speakers are involved. Having carried out this interpretative exercise, the analyst forms an overall conclusion. This is an evaluative opinion that draws on the analyst's expertise gained from advanced training, professional development, research literature, and casework experience as well as ecological experience derived from exposure to voices on an everyday basis.

It is relevant to point out that AuPhA voice comparison is not unusual with respect to the reliance placed on analyst judgement and decision making. This is evident from the Law Commission report, entitled *Expert Evidence in Criminal Proceedings in England and Wales* [9]. In producing the report, the Commission gathered cross-disciplinary responses to their proposals in relation to the admissibility of expert opinion evidence. Input from the FSS community at the time, as well as other disciplines such as fingerprint analysis and fibre analysis, emphasised the reliance on analyst experience and interpretation.

3. Method validation for AuPhA voice comparison

Section 2 highlighted the fact that the analyst is an integral part of the AuPhA voice comparison analysis; they are at the centre of every aspect of the process, irrespective of whether this involves making qualitative observations, taking quantitative measurements or interpreting analysis findings. In view of this, AuPhA voice comparison is best conceptualised as one single interpretative method that the analyst employs in reaching a voice comparison conclusion. Therefore, AuPhA method validation could be achieved by testing the competence of the analyst.

There have been suggestions within the UK FSS community that method validation could be achieved through the running of many 'validation' tests on all of the individual voice and speech features that might contribute to an analysis. This would be based on the alternative conceptualisation of AuPhA voice comparison which separates the analyst and the method and in turn treats each individual voice and speech feature as separate methods. In practical terms this might mean, for example, that the analysis of vowel sounds is regarded as one method, the analysis of f0 as another, the analysis of voice quality as another, then the competence of the analyst being yet another method, and so on. If one were to apply this 'micro-method' conceptualisation, it is not clear what validation would involve. The answer perhaps would be to carry out many different 'validation exercises' on these individual voice and speech features. However, to insist on individually validating the full range of features available to a forensic speech practitioner to show the features' performance in discriminating speakers would be a mammoth task for any field, let alone a field bearing the size and characteristics of the FSS community. It is not only the sheer number of features that requires consideration, but also how these perform as features under a range of recording conditions and within different speech communities (e.g., different regional varieties of the language).

Aside from the impractical scale of the task, there is an inherent limitation that comes with 'micro-method' validation. Validation implies that there is ground truth to compare against. For many features included in an AuPhA analysis, it is not possible to obtain ground truth because most qualitative and quantitative elements of the analysis are *estimates*. For example, [10] and [11] provide a reminder that the aim of quantitative acoustic measurements, such as f0, is to arrive at the most representative value, i.e. an estimate, rather than a "true" value. Finally, conclusions arising from AuPhA voice comparison analyses reflect the expert's interpretation of the combined range of different features included in any specific case. Presumably, there would also have to be an overall validation exercise that combines all the individual 'validation exercises'. The task would clearly be untenably complex.

Of course, the above does not discount the importance of demonstrating the science that underpins the use of individual voice and speech features in order to differentiate among speakers. It is vital to continue to test the performance of voice and speech features in discriminating speakers, how these features are affected by recording and speaking contexts, how best to analyse these features, etc. However, rather than viewing this type of testing as method validation, it is better to view this testing as 'method development'. The Expert Working Group for Forensic Speech and Audio Analysis (FSAAWG) within the European Network of Forensic Science Institutes (ENFSI) has recognised the role of method development in this respect in Section 6.1 of their Best Practice Manual [12]:

"*The methodology of FSC has been developed within forensic laboratories for decades. The development has been conducted on the basis of scientific research in the field of forensic speech science and linguistics, through published peer-reviewed literature and empirical testing (under casework conditions), and through knowledge exchange within the community of researchers and experts* (e.g. *during conferences). The methodology consists of analysing individual speech features on different dimensions with appropriate methods, each of them having been developed and tested for its speaker discriminatory power in the aforementioned way (see the literature in Appendix 2 'Bibliography'). Individual analyses are then combined to arrive at an overall conclusion.*" (page 10).

This paragraph goes on to say that:

"*New methods of feature analyses shall also be validated as they become available*".

The use of the word "validated" in this sentence has the potential to lead to confusion. We strongly suspect that ENFSI do not intend to promote, what we have termed, 'micro-method' validation for AuPhA voice comparison as this is not what is done within ENFSI.

Instead of competency testing, as we propose in this paper, the approach adopted within ENFSI foregrounds the method that is being employed rather than the individual analyst.[1] Within and across government laboratories in Europe, AuPhA method validation is performed by laboratories taking part in collaborative exercises and proficiency tests. The names of the analysts who participate in these activities are typically anonymised. The intention is to test the method by retrieving test results from multiple analysts and generalising across them.

At first glance, the approach adopted within ENFSI appears to differ from what we are proposing in this paper. When taking a closer look, however, the difference is more of a difference in emphasis rather than a difference in substance. The approach to competency testing proposed in this paper does not just focus on analyst-level testing, but also on comparing test results from individual analysts with those produced by other analysts employing the same method.

4. Competency testing: Obstacles for AuPhA voice comparison

There are different ways of assessing and certifying the competence of forensic analysts (some of which are recognised in the draft Code). These include:

a) Designing in-house assessments using positive and negative competency tests;
b) Independent confirmation of results/opinions by another competent examiner, i.e., without prior knowledge of the first result/opinion provided;
c) Participating in inter-laboratory comparisons, e.g., collaborative exercises;
d) Proficiency testing through recognised and relevant professional organisations.

On the surface, there appears to be a pool of options. When considering the context of forensic voice comparison, however, issues soon arise with respect to their viability. Forensic voice comparison is a niche area within forensic science. In the UK context, voice comparison casework for evidential purposes is currently undertaken entirely by private providers, rather than government organisations. Many of these providers are individual practitioners (sometimes academics). To give an indication of the size of the field, to the best of our knowledge, at the time of writing, there are around five sole practitioners, some of whom work with an assistant, and one micro company with four full-time analytical staff. All of these factors – being niche, working as a private provider and, where applicable, being a sole practitioner – have meant that opportunity to engage with competency testing has been very limited.

Options a) and b) are not readily available to sole practitioners. In order to pursue these possibilities, sole practitioners would have to collaborate with external partners. This requires that there are external partners available who are not only interested in collaborating, but who also have the necessary time and expertise. Option c) lends itself well to government-funded laboratories. The Speech and Audio units of the German Bundeskriminalamt (BKA) and those of the relevant Landeskriminalämter (LKAs) participate in various forms of inter-laboratory

[1] We are grateful to an anonymous reviewer for providing this insight.

comparisons and have done so for many years [13]. It is, of course, conceivable that inter-laboratory testing could be put in place between private providers, but nothing like this currently exists in the UK. All of these options carry with them financial implications. While this is not a barrier per se, any financial expenditure needs to be proportionate to the organisational set-up. Without external funding, it is clear that these financial repercussions would have to be absorbed by clients through the costs they pay for voice comparison analysis.

In relation to option d), the first and second authors' attempts at setting up proficiency testing through relevant organisations have so far been unsuccessful. There are private organisations which advertise that they assist providers and disciplines in conducting proficiency testing. However, from experience of contacting these organisations to develop a proficiency test for forensic voice comparison, there is little appetite for the prospect. These organisations have responded to say they do not have the resource or knowledge base to develop a proficiency test for this area, and it would seem that finding a way of doing so for such a small forensic discipline is not a profitable venture.

In contrast to these private organisations, the FSAAWG within ENFSI (see Section 3 above) is an example of a professional organisation that is invested specifically in speech and audio analysis. The FSAAWG was formed in 1998 and one of its aims is to improve, develop, and evaluate methodologies used in forensic voice comparison. To this effect, the group has in previous years run inter-laboratory comparisons. The first author explored the possibility of engaging with the activities of the FSAAWG but was informed that, as a private provider, it is not open to her to join the group and participate in its activities.

Another professional organisation relevant to the area of forensic voice comparison is the International Association for Forensic Phonetics and Acoustics (IAFPA), formally established in 1991. One of its aims is to provide "*a forum for the interchange of ideas and information on practice, development and research in forensic phonetics and acoustics*" [14]. Between 2015 and 2016, there was an attempt by the IAFPA Working Group on Developing an Infrastructure for Testing to initiate competency exercises, but the take up of this testing opportunity by UK practitioners was very low. Reasons may relate to the financial and time investment required for these types of activities.

5. A route to competency testing for AuPhA voice comparison

In view of the points raised in Section 4, it is currently down to proactive and committed members of the FSS community to pave a way to competency testing. An example of a project which aimed to do just that can be seen in [15]. The author, together with colleagues at the Netherlands Forensic Institute, designed a collaborative exercise, one of the purposes of which was to gain greater insight into the different analysis methods employed when carrying out voice comparison analysis. As part of the collaborative exercise, participants were invited to compare voices in twelve recordings and report on their findings. The participants represented practitioners from within and outside of Europe, some of whom employed the AuPhA approach, and others adopted approaches involving an automatic system. We do not know whether UK practitioners volunteered and participated in this exercise. In any case, to the best of our knowledge, shared analysis exercises such as this have not been repeated by UK practitioners outside of their own laboratory or organisation since.

[15] emphasises that the collaborative exercise was not intended to be a proficiency test, and that the results, as presented in the article, do not allow for an assessment of the performance of the individual analysts involved. There is no doubt that collaborative exercises such as this are invaluable. It is through open discussion and sharing of best practice that improvements in performance in voice comparison analysis can be achieved. However, it is also fruitful to complement collaborative activities of this kind, with outright proficiency testing as this more directly taps into the competence of the analyst.

The remainder of this section outlines the way in which the authors, on behalf of *Soundscape Voice Evidence* developed and carried out a proficiency test (the "Test"). The Test was put in place from the perspective of sole practitioners and micro-organisations, as these make up the current UK forensic speech analysis provision. For sole practitioners, there is an inevitable need to reach out to individuals outside of the organisation to assist with different aspects of the Test. However, even in the case of micro-organisations where there is the potential to draw on 'in-house' expertise, calling on colleagues from outside the organisation to contribute to the Test has the added advantage of transparency and increased objectivity.

Given that the current proficiency testing exercise is, to some extent, exploring new ground, it was decided that the best strategy would be to start simple. As such, the Test was to take the form of a mock case covering the comparison of the voice and speech patterns of a questioned speaker in one recording with those of a known speaker in another recording. A one-to-one comparison of this type is a standard request in UK casework. On a superficial level, the size of the Test may somewhat limit the generalisations that can be made with respect to the analyst's performance in voice comparison analysis. However, the purpose of the Test was not just to evaluate whether the analyst can reach the right conclusion, i.e., whether the analyst's overall conclusion accords with the ground truth. The Test also sought to evaluate the process followed by the analyst in carrying out AuPhA voice comparison, including scrutiny of the analyst's skill and decision making, e.g., the analysis of f0, analysis of vowel sounds, interpretation of observed similarities and differences, etc. Furthermore, a key objective was to design a test which would provide a viable long-term option for sole practitioners in particular. As such, the amount of work involved in completing the Test had to be proportionate to the organisational setup. To illustrate, in completing the collaborative exercise designed by [15] participants spent 65 hours on average, ranging from 28 hours to 120 hours. We suspect that the participants employing the AuPhA approach were closer to the 120-hour mark. This is a significant time investment that any provider type would struggle to support on a recurring cycle, let alone a sole practitioner in the private sector.

Having consulted relevant guidance [16], the following subsections provide further detail on (i) the construction of the Test; (ii) how the Test was carried out by the analyst; and (iii) the external review.

5.1. Proficiency test construction

As the Test co-ordinator, the second author (GB) was responsible for constructing the Test, the process of which can be divided into two stages: identifying potential test data and deriving the actual Test case data from the pool of potentials.

5.1.1. Identifying potential mock case data

As a starting point, it was considered important that the data used for the mock case should meet the following criteria:

a) The speech and recording types should be reflective of forensic casework data;
b) There should be ground truth available with regards to the identities of the speakers in the two recordings, i.e., the voices in the two recordings are known to form a same-speaker pair or a different-speaker pair;
c) The Test should not be "too easy"; neither should it be "too hard" as to become unsuitable for AuPhA voice comparison.

To meet a), the ideal scenario would be to use real case data. It is only real case data that can truly fulfil this requirement because, as discussed in [17], it is impossible to fully replicate speech produced under the conditions typically reflected in evidential recordings containing speech (for example, the emotions and pressures often experienced by the speakers). However, a key problem with using real case data is that it is often not possible to accommodate requirement b). The ground truth is

generally unknown, i.e., that it is definitely the same person speaking in the recordings under analysis, or that the speakers in the recordings are definitely different people. A pragmatic solution may be to find recordings from a real case where the suspect pleaded guilty, and then rely on the guilty plea as forming ground truth. There are also shortcomings to this approach, however, including the fact that suspects may plead guilty irrespective of whether they have actually committed the wrongdoing in question, or the guilty plea was based on case information or evidence other than the voice comparison analysis. Even if it were to be assumed that a guilty plea confirmed that the same speaker is found in the recordings under analysis, this would restrict the possible mock case scenarios to just same-speaker trials and would leave no prospect of including different-speaker trials in the proficiency test.

An alternative route to obtaining data for the mock case would be to record the test samples from scratch. This is an approach that was adopted by [15] when designing the collaborative exercise discussed above. While one would have control over many of the technical and environmental characteristics of the recordings if collected in this way, this would limit the variety of candidate voices, unless multiple voices are recorded across multiple recording sessions. This of course has implications on the time and labour involved in creating a proficiency test. As a key objective was to design a process which is viable for sole practitioners currently working in the UK, it became clear that recording samples from scratch would not be a feasible option, at least initially.

GB therefore turned towards existing databases of speech recordings, in particular those which contain speech samples submitted by speakers across multiple recording sessions. Accurate metadata in such databases allow for proficiency tests supported by ground truth. While it is acknowledged that these do not contain recordings wholly reflective of casework conditions, many of the recordings are nevertheless forensically relevant.

Given that they were primarily collected for FSS research purposes, the Dynamic Variability in Speech (DyViS) database [18] and the West Yorkshire Regional English Database (WYRED) [19] were considered as sources of mock case data. Both corpora contain recordings of at least 100 young adult men speaking in different forensically relevant recording conditions and communicative settings, e.g., studio recordings, telephone transmission, conversations with different levels of formality, and monologues. Unfortunately, both corpora were judged to be unsuitable for the current proficiency testing purpose because the analyst taking the Test, i.e., the first author (CK), is already very familiar with the speech samples in these corpora through attending research talks and carrying out research projects involving the analysis of the speech contained within these corpora.

It was therefore decided that a corpus with which CK did not have previous contact would be more suitable for the Test. To this end, GB sourced a dataset from the UK Government's Defence Science and Technology Laboratory (DSTL) which contained speech from approximately 100 men and women across multiple recording sessions. The voices in this corpus also covered a range of demographics, including different accent varieties and speaker ages. As such, GB decided that this formed a suitable pool from which to select recordings for the Test.

5.1.2. Deriving the data used for the Test

Using the DSTL dataset, GB sought pairs of speech recordings that could form potential test data. Firstly, as the vast majority of evidential speech recordings contain male voices, GB decided that the Test should contain male voices. GB listened to samples of all the male speakers in the dataset and considered three main factors when selecting potential test pairs:

i) The samples are not so poor quality that they are deemed to be unsuitable for AuPhA analysis;
ii) The samples are sufficiently challenging and reflective of casework, including the mismatch between them;
iii) The voice characteristics are not too distinctive so as to make the Test "too easy".

In relation to all of i) - iii), GB relied on her exposure to evidential recordings over three years, as well as her experience working within speech science more broadly. This initial selection procedure resulted in nine potential sample pairs which were taken forward to the next stage. This next stage called on an Advisory Panel.

The purpose of the Advisory Panel was to assess the nine sample pairs and to eventually reach a single pair of recordings that could be used for the Test. The Advisory Panel consisted of GB and two researchers within speech science, both of whom have appropriate and differing experiences of working with speech databases. For the purpose of structuring the discussions, members of the Advisory Panel were asked to consider each of the nine sample pairs (at first, without knowing the ground truth) and note the following for each pair:

a) How similar or different the voices sound;
b) Any general observations about the recordings and voices;
c) How challenging the voice comparison analysis might be;
d) Whether there was an appropriate degree of mismatch in recording channel and/or speaking style;
e) How appropriate it was for a forensic voice comparison proficiency test.

After detailed discussion of all nine sample pairs, the ground truth of each pair was revealed to the Advisory Panel for further deliberations. Panel members were then asked to independently rank the nine pairs in order of how appropriate they were for the Test. There was a lot of agreement among the panel members, with all three identifying the same sample pair as the most fitting for the Test. The recordings that made up the Test were judged to be of relatively good technical quality for evidential material. There was a mismatch in recording channel, with one sample being telephone-transmitted and the other not. The speech that was not telephone-transmitted was instead captured at a distance from the microphone. There was also mismatch in communicative setting in that one of the samples featured a conversation involving a police call handler where the speaker of interest was reporting a crime. In the other sample, the speaker of interest was enquiring with a car dealership about the purchase of a car. Together, this meant that there was a matrix of mismatch between the recordings under analysis. The speech in both recordings was spontaneous and the speakers were speaking on normal voice levels. Around 60 seconds of net speech was available in one of the recordings, while around 80 seconds of net speech was available in the other. It was this sample pair that was taken forward to form the Test for CK.

5.2. Taking the Test

CK was aware that she was participating in a proficiency test and so this meant that the Test was *overt*. [20] present arguments for implementing *covert* tests in forensic laboratories, i.e., where the analyst is unaware that they are taking a proficiency test. However, particularly for the small organisational setup typical of FSS practice in the UK, covert testing would be impractical to arrange. In order for the analyst to believe that they are carrying out a real case, the proficiency test would have to replicate the whole casework pipeline, starting at the enquiry stage where a customer approaches the organisation for a quote and turnaround time.

Further, it is unclear how much there is to be gained from covert tests, as opposed to overt tests. The purpose of proficiency testing is to determine whether analysts are competent. If an analyst is not competent, it is expected that this would be revealed by both overt and covert tests. Even if overt tests lead to a change in behaviour, that change in behaviour is unlikely to be material to the question of whether they can competently carry out an analysis.

CK carried out the Test replicating the process she would ordinarily use in real casework, and arrived at an overall conclusion addressing the question of whether the same speaker or different speakers featured in the two test recordings. She made a record of her analysis and conclusion and wrote a report in the same format as for a real case.

5.3. The external review

Although performance could be measured in a broad sense by comparing CK's overall conclusion with the ground truth, the scope of this proficiency test extended to a more comprehensive evaluation of CK's AuPhA voice comparison analysis process. As such, it was important to engage with an external reviewer.

The third author (PF) was approached to be the external reviewer based on his experience within the field in both research and casework, and his independent status in relation to *Soundscape Voice Evidence*. The purpose of the external review was to carry out an "audit" of CK's work, covering a) the analysis itself, b) the conclusion, and c) the written communication of the analysis and conclusion. PF was given some possible questions to consider as part of the review:

- Whether an appropriate range of features was selected and whether the selected features were suitable for comparison given the quality of the samples, e.g., whether the samples were suitable for voice quality, f0, vowel analysis, etc.;
- Whether there were any features which should have been selected and analysed but which were not included;
- Whether appropriate tools were used in extracting the features;
- Whether the observations made about the features were appropriate, e.g., whether the reviewer agrees with the assessment of voice quality, vowel sounds, etc.;
- Whether the reviewer agrees with the interpretations made in relation to the analysis and comparisons of the features;
- Whether the conclusion is justified;
- Whether the written communication of the analysis and conclusion achieves a balance between being understandable to the non-expert while retaining the necessary technical detail for another analyst to follow.

Overall, PF was free to choose how he would like to approach the reviewing task, and he was encouraged to comment on any aspect that he felt important to highlight and discuss. In approaching the review, PF first chose to carry out his own AuPhA voice comparison on the Test recordings, with no knowledge about the materials (including whether or not the recordings were from the same speaker), prior to referring to CK's analysis notes and report. This was in an effort to increase the transparency and rigour of the review.

As part of his independent analysis, PF checked the edited sound files produced by CK and carried out an overall suitability assessment of the material for AuPhA analysis. He then based his own AuPhA voice comparison analysis on CK's edited files. PF drew on a range of voice and speech features (e.g., vowel and consonant sounds, voice pitch, voice quality, speech prosody and discourse patterns) in order to arrive at his own conclusion that addressed the question of whether the two recordings contained the same or different speakers.

PF's analysis and conclusion formed a basis for comparison with CK's analysis and conclusion. PF checked every voice and speech feature that had been mentioned by CK, but which he had not considered as part of his own analysis, or which appeared to differ from his own observations. With respect to the latter, for example, CK noted the deletion of /r/ in the word 'from', PF did not note this as an outright deletion of /r/ but rather as the /r/ being greatly reduced. Overall, CK and PF were extremely close in the selection of voice and speech features, the qualitative and quantitative observations reported, and the interpretations drawn. Specifically:

- The mean f0 estimates were identical;
- The acoustic estimates of vowel sounds were very similar;
- Perceptual evaluations of vowel and consonant sounds, voice quality, prosody and discourse-level patterns were very similar;
- There was agreement between CK and PF in relation to the distinctiveness of the set of features shared between the questioned and known voices;
- There was agreement between CK and PF in relation to the overall voice comparison conclusions; firstly, both of their conclusions were in accord with ground truth (i.e., that the evidence supports the same-speaker view), and secondly, they arrived at a similar degree of support for that view (i.e., CK arrived at *moderately strong support* and PF arrived at *moderate support*). CK and PF expressed their conclusions with reference to the scale that is recommended by the UK Association of Forensic Science Providers [21] and ENFSI [22] (however, their conclusions were not derived from a numerical likelihood ratio).

As we would expect, there were minor differences between CK's and PF's analyses. However, CK and PF worked together to review the features in question to resolve the discrepancies. This exercise revealed that the source of difference was not in whether a feature was present or absent, but rather it was the result of the descriptive terms used by each analyst. For example, both CK and PF noted underarticulation as a characteristic that was shared between the questioned and the known voices; however, while CK used the expression 'lax vocal tract' to capture this feature, PF expressed it as 'lax speech'. There were no differences in observation or interpretation that would be of any material value in a forensic voice comparison case.

6. Discussion

There are benefits to proficiency testing beyond the most basic function of revealing an individual analyst's competence. Proficiency testing could also increase confidence and trust placed on the whole FSS community by police officers, lawyers and the public. In addition, proficiency testing schemes provide a professional development framework which could fuel the morale and pride of new and existing practitioners. Naturally, this would support the recruitment and retention of forensic speech analysts, enabling the field to sustain itself.

The example of proficiency testing presented in this paper is seemingly straightforward, but the financial and time investment required should not be underestimated in the context of a small organisation. We have pointed out the challenges attached to the availability of appropriate personnel and data, and there are also substantial costs associated with the exercise. Taking into account direct costs (paying the Test coordinator, external reviewer and Advisory Panel) as well as the indirect cost in the time spent on designing and carrying out the Test (i.e., loss of earnings), this exercise amounted to around 5% of *Soundscape Voice Evidence*'s annual turnover.

The Test carried out could be criticised as being too small, as it only involved the comparison of two recordings and only two analysts, and therefore only so much can be drawn from it. However, we question the viability of even carrying out small tests like this on an annual basis. In order to ensure that competency testing becomes a more integral part of a practitioner's existence, there must be room for flexibility with respect to the nature, size and frequency of this testing. It is envisaged that a practitioner would compile a varied portfolio of competency exercises, some of which might involve groups of practising analysts coming together on an ongoing basis throughout their career. To achieve breadth within the portfolio, the competency exercises might cover different data types and would vary in nature and scope. Examples of exercises that could contribute to such a portfolio include:

- Full-blown voice comparison analyses involving different accents, speaker demographics, sample durations, recording characteristics, etc.;
- Voice comparison sub-tasks such as accent profiling (as exemplified in [23]);
- Narrowly-focused analysis tasks that aim to scrutinise and exchange best practice in vowel analysis, f0 analysis, voice quality analysis, etc.;
- A lighter touch task involving blind-grouping of multiple speech samples (as exemplified in [24]).

The more analysts involved in a single exercise, the stronger the foundation to make generalisations about AuPhA voice comparison. In the UK, we can draw on only a handful of analysts; however, for some of the narrowly-focused analysis tasks that do not rely on language specific expertise, there could be opportunity to involve analysts who practise outside of the UK.

The portfolio concept could also address other challenges that FSS faces. There is an expectation placed on forensic speech analysts that case notes and reports are reviewed by another suitably qualified analyst to ensure quality. An expectation that such reviewing takes place in every single case presents a practical barrier to those who work as individual practitioners and in turn suffocates a field that is dependent on these practitioners. Further, pressuring a niche forensic discipline to include these reviews in every single case may mean organisations are tempted to put in place surface-level checking procedures which amount to box ticking but which lack substance – possibly even leading to a false sense of security. Compiling a portfolio of the kind described in this section, which incorporates the use of external and comprehensive review of analysts' work, could provide an alternative safeguarding mechanism that carries credibility.

A recurring criticism of AuPhA voice comparison is that there are no error rates that can be referred to in order to demonstrate its reliability (unlike automatic systems used to carry out voice comparison). This inability to produce numerical error rates is even used to advocate for the use of automatic systems over the AuPhA approach implemented by a human expert [25]. A portfolio of competency exercises could go a long way to place confidence in an individual practitioner and in AuPhA voice comparison as a whole.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Christin Kirchhübel: Conceptualization, Investigation, Project administration, Writing – original draft, Writing – review & editing. **Georgina Brown:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Paul Foulkes:** Methodology, Investigation, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] M. Jessen, Forensic voice comparison, in: J. Visconti (Ed.), Handbook of Communication in the Legal Sphere, De Gruyter Mouton, Berlin, Boston, 2018, pp. 219–255, https://doi.org/10.1515/9781614514664-012.

[2] The Forensic Science Regulator Act 2021: www.legislation.gov.uk/ukpga/2021/14/contents.

[3] The Forensic Science Regulator Draft Statutory Code of Practice: https://www.gov.uk/government/consultations/forensic-science-draft-statutory-code-of-practice.

[4] The Criminal Procedure Rules 2020, Part 19: https://www.gov.uk/guidance/rules-and-practice-directions-2020.

[5] The Criminal Practice Directions 19A: https://www.gov.uk/guidance/rules-and-practice-directions-2020.

[6] G.S. Morrison, E. Enzinger, V. Hughes, M. Jessen, D. Meuwly, C. Neumann, S. Planting, W.C. Thompson, D. van der Vloed, R.J.F. Ypma, C. Zhang, A. Anonymous, B. Anonymous, Consensus on validation of forensic voice comparison, Science and Justice 61 (3) (2021) 299–309, https://doi.org/10.1016/j.scijus.2021.02.002.

[7] P. Foulkes, P. French, Forensic speaker comparison: A linguistic-acoustic perspective, in: L.M. Solan, P.M. Tiersma (Eds.), Oxford Handbook of Language and Law, Oxford University Press, Oxford, 2012, pp. 557–572.

[8] J. Laver, S. Wirz, J. Mackenzie Beck, S.M. Hiller, [1981] A Perceptual Protocol for the Analysis of Vocal Profiles, in: J. Laver (Ed.), the Gift of Speech: Papers in the Analysis of Speech and Voice, Edinburgh University Press, Edinburgh, 1991, pp. 265–280.

[9] Law Commission report. Expert Evidence in Criminal Proceedings in England and Wales [report no 325, published March 2011].

[10] P. Foulkes, G. Docherty, S. Shattuck Hufnagel, V. Hughes, Three steps forward for predictability. Consideration of methodological robustness, indexical and prosodic factors, and replication in the laboratory, Linguistics Vanguard 4 (s2) (2018) special edition on Predictability and phonology: past, present and future.

[11] Kendall, T., and Fridland, V. (2021). Sociophonetics (Key Topics in Sociolinguistics). Cambridge: Cambridge University Press. doi:10.1017/9781316809709.

[12] Best Practice Manual (BPM) for the Methodology of Forensic Speaker Comparison produced by the Expert Working Group for Forensic Speech and Audio Analysis (FSAAWG) within the European Network of Forensic Science Institutes (ENFSI) in December 2022. ENFSI-FSA-BPM-003. Version 1. Available: https://enfsi.eu/about-enfsi/structure/working-groups/documents-page/documents/best-practice-manuals/.

[13] Konrat, C. and Jessen, M. (2013). Fundamental Frequency Analysis – A Collaborative Exercise. Presentation delivered at the annual International Association for Forensic Phonetics and Acoustics conference. Tampa, Florida, USA.

[14] International Association for Forensic Phonetics and Acoustics (IAFPA) website. URL: https://www.iafpa.net/ [accessed on 4th October 2022].

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

[16] ENFSI. (2014). Guidance on the Conduct of Proficiency Tests and Collaborative Exercises within ENFSI. QCC-PT-001. Issue date: 27/06/2014.

[17] G. Brown, S. Ross, C. Kirchhübel, Voicing Concerns: The balance between data protection principles and research developments in forensic speech science, Science & Justice 61 (2021) 311–318.

[18] F. Nolan, K. McDougall, G. de Jong, T. Hudson, The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research, International Journal of Speech, Language and the Law 16 (1) (2009) 31–57.

[19] Gold, E., Ross, S., and Earnshaw, K. (2018). The 'West Yorkshire Regional English Database': Investigations into the generalizability of reference populations for forensic speaker comparison casework. Proceedings of Interspeech 2018 – 98th Annual Conference of the International Speech Communication Association, September 2-6, Hyderabad, India, pp. 2748-2752.

[20] R. Mejia, M. Cuellar, J. Salyards, Implementing blind proficiency testing in forensic laboratories: motivation, obstacles, and recommendations, Forensic Science International Synergy 2 (2020) 293–298, https://doi.org/10.1016/J.FSISYN.2020.09.002.

[21] Standards for the formulation of evaluative forensic science expert opinion. (2009). Science and Justice 49: 161–164.

[22] ENFSI guideline for evaluative reporting in forensic science: Strengthening the Evaluation of Forensic Results across Europe, version 3.0, 08/03/2015. https://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf.

[23] O. Köster, R. Kehrein, K. Masthoff, Y.H. Boubaker, The tell-tale accent: identification of regionally marked speech in German telephone conversations by forensic phoneticians, International Journal of Speech, Language and the Law 19 (1) (2012) 51–71.

[24] T. Cambier-Langeveld, M. van Rossum, J. Vermeulen, Whose voice is that? Challenges in forensic phonetics, in: J. Caspers, Y. Chen, W. Heeren, J. Pacilly, N. O. Schiller, E. van Zanten (Eds.), Above and Beyond the Segments. Experimental Linguistics and Phonetics, John Benjamins Publishing Company, Amsterdam, 2014, pp. 14–27.

[25] Morrison, Admissibility of Forensic Voice Comparison Testimony in England and Wales, Criminal Law Review 1 (2018) 20–33.