Realistic solutions for practising forensic scientists – a response to Morrison (2023)

1. Introduction

In Kirchhübel et al. (2023) [1], we address the issue of method validation for the Auditory Phonetic and Acoustic (AuPhA) approach to forensic voice comparison (FVC). We illustrate one possible solution, which centres around competency testing of the human expert implementing AuPhA. We received a response to our article by Morrison (2023) [2] which conveys the idea that there is but a single acceptable way of meaningfully demonstrating validity of FVC methods. This position not only contradicts a range of relevant literature (including Morrison et al. (2021) [3] – henceforth the "*Consensus*"), but it also defies common sense. We take this opportunity to respond to Morrison (2023) [2].

2. Clarifying the issue of validation

It appears that the main issue that Morrison (2023) [2] raises relates to the number of recordings needed to validate FVC methods. This is reflected in the title of Morrison (2023) [2] which is "*A single test pair does not a method validation make: A response to Kirchhübel et al. (2023)*". Morrison (2023) [2] recognises that it would be challenging in practical terms to validate AuPhA according to the recommendations made in the *Consensus*. However, it is not the case that it would just be challenging; it would in fact be impossible. An illustration may help here. The *Consensus* recommends that validation data should be sufficiently representative of the case at hand and contain a sufficient number of speakers. Importantly, the *Consensus* does not specify a number. This in itself indicates that it would be difficult to agree a specific number to apply to all cases, and this implies that there is a need for flexibility. Despite this, Morrison (2023) [2: 328] states that '*It may be that a practically achievable validation set consists of pairs of recordings from only upper tens of speakers to lower hundreds of speakers*.' Let's assume for the purposes of exposition that the validation set contains 100 same-speaker FVC tests and 100 different-speaker FVC tests. Using the AuPhA approach exemplified in Kirchhübel et al. (2023) [1], it would take at least 10 hours to complete one FVC test. We are therefore looking at 2,000 hours, or in other words, 50 full-time working weeks (net analysis time) to complete all the tests as part of the proposed validation exercise. And then this of course would need to be repeated for a new type of case in which the recordings are judged to be insufficiently comparable to those contained in the validation exercise (e.g., different recording types or speaker demographics involved).

To overcome these time and resource impracticalities, Morrison (2023) [2] puts forward the solution of simply discarding AuPhA for FVC and only using human-supervised automatic methods. However, we consider this proposal to be detrimental, in part, because not all casework recordings meet the quality requirements in order to apply an automatic system. Even for those cases where the recordings do meet the quality requirements to use an automatic system, there are practical hurdles to overcome. These hurdles include, for example, obtaining case-relevant recordings to form suitable validation datasets, ensuring the competence of the humans supervising the automatic systems, and navigating its reception within a legal environment. Therefore, solely relying on automatic systems would mean that a huge number of FVC cases go unanalysed and forensic speech science would become largely ineffectual as a forensic discipline. We are very much in favour of incorporating automatic systems into FVC, but this is not with a view to displace AuPhA. There is widespread recognition within the forensic speech science community that the two methods – AuPhA analysis and automatic speaker recognition – complement one another. As such, practitioners who regularly use automatic systems in FVC casework use them in conjunction with an AuPhA analysis (e.g., Jessen (2018) [4]; van der Vloed and Cambier-Langeveld (2023) [5]).

Rather than discarding AuPhA, a much more productive solution, in our view, is to find a way of validating AuPhA which is both meaningful and practically feasible. The approach contained within Kirchhübel et al. (2023) [1] exhibits both of these characteristics by carefully selecting data which are reflective of casework conditions and by including a comprehensive external review process.

In any case, Morrison (2023) [2] has misrepresented the proposal contained in Kirchhübel et al. (2023) [1] (particularly in the title of Morrison (2023) [2]) by suggesting that we have proposed a validation protocol that only includes an analysis of a single pair of recordings. When reading Kirchhübel et al. (2023) [1] in full, it becomes clear that what we are proposing is a portfolio approach to validation which involves numerous FVC and speech analysis exercises covering different accents, speaker demographics, recording characteristics, sample durations, etc. The single 1:1 comparison is presented as an example of one component within this approach. It is not intended as the sole contribution. In fact, we explicitly list “blind-grouping” (e.g., Cambier-Langeveld et al. (2014) [6]) as another possible component which would provide opportunity to include a larger number of speech samples into a validation

exercise. We propose that this validation approach is an ongoing and cumulative process throughout the careers of practitioners. Morrison (2023) [2] therefore fails to acknowledge our full proposal and appears to have taken a selective approach to his response.

3. Further Clarifications

There are further clarifications to make in response to Morrison (2023) [2]. The main ones are set out below.

3.1 Scope of the *Consensus* (Morrison et al. (2021) [3])

Morrison et al. (2021) [3] present a '*Consensus'* that describes recommendations for how method validation can be achieved using an automatic approach to FVC. Morrison (2023) [2] points out that the *Consensus* had 13 authors and an additional 7 supporters. We acknowledge the efforts in producing this work and we do not take issue with its contents (as indicated by its citation in Kirchhübel et al. (2023) [1]). However, it is important to highlight that the *Consensus* has a restricted scope, i.e., it applies to "*validation of forensic-voice comparison systems that are based on relevant data, quantitative measurements, and statistical models, and that output numeric likelihood ratios*". That is, it largely applies to forensic application of automatic speaker recognition. However, Morrison (2023) [2] appears to suggest that, "*with minor wording changes*", the scope of the *Consensus* can be extended to cover other approaches to FVC. The AuPhA method as described in Kirchhübel et al. (2023) [1] does not fall within the original scope because it is not dependent on statistical models, nor does it necessarily result in numeric likelihood ratios. Further, in Section 2 of this response, we have already explained why the *Consensus* cannot be extended to AuPhA, and "*minor wording changes*" would therefore make no difference. In any case, we do not consider it to be appropriate to attempt to extend a consensus, which by its very nature is based on explicit agreement from all contributors, to matters outside its scope.

3.2 Admissibility of automatic systems

Footnote 3 of Morrison (2023) [2] states:

"*Kirchhübel et al. (2023) claims that the auditory-acoustic-phonetic approach "is the only admissible approach in UK jurisdictions for voice comparison analysis." The implication that the human-supervised-automatic approach would not in-principle be admissible is incorrect*."

We offer two responses in relation to this comment. First, Morrison (2023) [2] misrepresents our statement by omitting some words from the beginning of the relevant sentence. It actually reads:

"*At the time of writing, AuPhA analysis is the only admissible approach in UK jurisdictions for voice comparison analysis*."

Second, we certainly did not intend for this to imply a general inadmissibility of automatic methods for FVC in UK jurisdictions. The key legal authority on this point is R v Slade [2015] EWCA Crim 71 where the court was presented with voice comparison analysis by way of an automatic speaker recognition system. The court voiced a number of concerns about this evidence and ultimately declined to admit it in that case, although the court did not make a definitive ruling as to whether automatic speaker recognition evidence can ever be admissible. On reflection, we accept that we could have presented a clearer account in our original article. We are keen to reiterate that we are very much in favour of incorporating automatic methods in FVC.

3.3 Expression of conclusions

Footnote 4 of Morrison (2023) [2] questions whether we have actually followed the logic of the likelihood ratio framework:

"Kirchhübel et al. [22] states that the practitioner and the reviewer "expressed their conclusions with reference to the scale that is recommended by the UK [sic] Association of Forensic Science Providers [25] and ENFSI [26] (however, their conclusions were not derived from a numerical likelihood ratio)." Both those scales, however, are intended to provide verbal expression corresponding to numerical ranges of likelihood ratios, and the ENFSI Guideline states that even if the numerator and denominator of a likelihood ratio are "informed by subjective probabilities using expert knowledge. These probability assignments shall still be expressed by a number between 0 and 1 rather than by an undefined qualifier (such as frequent, rare, etc.)." There is no evidence in Kirchhübel et al. [22] that the practitioner or reviewer actually followed the logic of the likelihood-ratio framework."

For the avoidance of any doubt, the practitioner and the reviewer in Kirchhübel et al. (2023) [1] did follow the logic of the likelihood ratio framework (which, of course, is not dependent on computing or estimating numerical probabilities). Instead of computing probabilities with reference to databases, probabilities were derived from knowledge and judgment, with

appropriate caution. Instead of assigning numerical values to these subjective probabilities, the practitioner and reviewer made use of verbal expressions. This is because the practitioner and reviewer also followed legal and regulatory authority that is directly relevant to UK jurisdictions when it comes to the expression of expert opinions which are the product of knowledge and experience, and not based on explicit databases. From R v Atkins and Atkins [2009] EWCA Crim 1876, it is clear that experts should not be using numerical estimates in these circumstances. This is because of the danger of misleading the jury to believe that quantitative processes have been applied when in fact they have not.

In addition, in guidance published in 2021, the UK Forensic Science Regulator (FSR) makes clear that numerical values should only be expressed if they are based on appropriate statistical data. In the absence of appropriate statistical data, the results should still be expressed probabilistically, but without numerical values. At paragraphs 7.2.7 and 8.5.13 in the FSR 2021 guidance [7] it states:

> *"In an ideal situation, the scientist would have access to relevant, high quality datasets to inform their evaluation. However, the nature of forensic science, where every case has different circumstances, means that large and directly relevant datasets are very rarely available. A LR [likelihood ratio] determined by the expert using experience alone can be no more wrong than the expert's opinion on which it is based and that opinion is admissible evidence. However, it is less easily tested than a LR based on relevant data, quantitative measurements, and statistical models' (7.2.7)*
>
> *'Where no (published or unpublished) structured data are available, a qualitative evaluation shall be reported, based on the expert's qualitative evaluation of the probability of the observations under each proposition..." (8.5.13)*

4. Conclusion

Morrison (2023) [2] paints a picture of forensic speech science being a field that has had years of opportunities to "progress" by way of analysis methods, and that these opportunities have been blocked by reluctant FVC practitioners who "*do not think like forensic scientists"*. However, Morrison (2023) [2] fails to accommodate the realities of FVC casework. By constraining ourselves to a human-supervised automatic method to FVC (simply because such a method could allow for hundreds of validation tests) would mean that very little FVC casework could be carried out. It is of little help to let a doctrinaire view of validation stifle other FVC methods, the scientific validity of which can be demonstrated through more flexible

means. In saying this, we are thinking like practising forensic scientists with experience of working within the criminal justice system. The point of Kirchhübel et al. (2023) [1] was to find real solutions to real problems while advocating for quality in FVC.

References

[1] Kirchhübel, C., Brown, G., & Foulkes, P. (2023). What does method validation look like for forensic voice comparison by a human expert? *Science & Justice*, 63, 251-257. DOI: 10.1016/j.scijus.2023.01.004.

[2] Morrison, G. S. (2023). A single test pair does not a method validation make: A response to Kirchhübel et al. (2023). *Science & Justice,* 63, 327-329. DOI: 10.1016/j.scijus.2023.03.001.

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

[4] Jessen, M. (2018). Forensic voice comparison. In Handbook of Communication in the Legal Sphere, edited by Jacqueline Visconti, Berlin, Boston: De Gruyter Mouton, pp. 219-255. DOI: 10.1515/9781614514664-012.

[5] van der Vloed, D., and Cambier-Langeveld, T. (2023). How we use automatic speaker comparison in forensic practice. *International Journal of Speech, Language and the Law*, 29(2), 201–224. DOI: 10.1558/ijsll.23955.

[6] Cambier-Langeveld, T., van Rossum, M. and Vermeulen, J. (2014). 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. Amsterdam: John Benjamins Publishing Company, 14-27.

[7] Forensic Science Regulator: Development of Evaluative Opinions. FSR-C-118 (Issue 1), 2021. URL: https://www.gov.uk/government/publications/development-of-evaluative-opinions.