Collection and analysis of multi-condition audio recordings for forensic automatic speaker recognition

Katharina Klug, Michael Jessen, Isolde Wagner
Bundeskriminalamt, Germany
{katharina.klug|michael.jessen|isolde.wagner}@bka.bund.de

The research project introduced here intends to expand the application of Forensic Automatic Speaker Recognition (FASR) systems in forensic voice comparison cases by validating the systems using case-specific material. The project is conducted by the German Federal Criminal Police Office (Bundeskriminalamt, BKA), funded by the EU Internal Security Fund (ISF).

Background

The performance of FASR systems improved considerably within the last two decades. Current automatic systems are capable of analyzing even degraded audio recordings forensic audio experts typically face. Therefore, FASR systems have gained wide attention as a component within voice comparison casework among practitioners worldwide (Gold & French 2019).

When case material meets the requirements for FASR application, forensic speech and audio experts at the BKA combine the auditory-acoustic approach with FASR (Wagner 2019). Both approaches follow the principles of similarity and typicality, developed within the Bayesian approach to voice comparison (Rose 2002). The difference between the approaches is that whereas FASR (as well as semiautomatic speaker recognition) generates numerical strength-of-evidence results in the form of likelihood ratios, the auditory-acoustic approach treats similarity and typicality on a more qualitative level (Drygajlo et al. 2015; Jessen 2018). Therefore, if combined, the FASR approach adds quantitative information to the predominantly qualitative information gained from the traditional auditory-acoustic approach.

So far, the case scenarios in which FASR has been applied at the BKA were mainly limited to telephone interceptions (and other forms of natural telephone conversations), with strongest emphasis on speakers of German. These cases generally occur under so-called matching conditions, i.e. both, the recording of the questioned speaker and the one of the suspected speaker (or several of each category), derive from fairly regular telephone conversations (potentially with increased stress and emotion). FASR systems need to be validated on material reflecting case conditions before being used to assess the strength of evidence of a case (Drygajlo et al. 2015). Therefore, the performance of FASR systems in matching telephone interception conditions has been tested at the BKA and in laboratories with similar casework (van der Vloed 2014; Solewicz et al. 2017).

Frequently, however, casework occurs under mismatched conditions, i.e. the technical or behavioural conditions of the questioned speaker's recording(s) differ systematically from those of the suspected speaker's recording(s). Moreover, some casework occurs under matched conditions, but the conditions are not telephone-based or they are telephone-based but have further complications (e.g. high noise level, compression, reverberation). Mismatched conditions and those outside regular telephone conversations are not well represented in validations, which currently limits the scope of cases in which FASR can be used. The few casework-oriented studies on mismatched or unusual conditions that have been conducted recently are based on dedicated recorded speech corpora. Morrison & Enzinger processed high-quality speech recordings in accordance with the mismatched characteristics of a real forensic case (Morrison & Enzinger 2019 for a summary of research that was based on this material). The approach taken in van der Vloed et al. (2020) was to record telephone conversations but under very different technical and environmental conditions that can lead to mismatch or reflect specific limitations.

The research project described in the present paper takes the approach to validate FASR systems using recordings that have occurred in real forensic audio material or as part of related investigative work.

Project

Challenges encountered regarding the compilation of a real forensic audio corpus include:

- highly sensitive data, involving the need to anonymize personal data,
- less-than-complete certainty in determining speaker identification,
- unbalanced number of recordings per speaker,
- unbalanced data for languages and/or conditions of interest.

The table shows the conditions and languages for which audio material has been collected so far.

Table 1. Conditions and languages to be tested with FASR systems (provided sufficient amount of real forensic audio material)

Conditions	Language(s)
Telephone interception	German, Arabic, Turkish, Russian
Video	German, Arabic
Interior surveillance (car/living space)	German, Turkish
Voice message	German

A minimum of 20 male adult speakers per condition and language are collected, providing two to six recordings per speaker. The net duration of speech ranges from 10 to 60 seconds. Additionally, when available, training data sets of independent speakers per condition and language are collected providing only one recording per speaker to create relevant populations. The data preparation includes segmentation into net speech and anonymization of personal information (e.g. names, telephone numbers, addresses).

Using the available data, the performances are being tested of:

- various commercial speaker recognition systems,
- several generations of approaches (GMM/UBM, i-vector, x-vector),
- different methods of score normalization and adaptation.

Current results will be shown during the presentation.

Expanding the FASR application in forensic voice comparison cases will be important for the mismatch issue and for the investigation of matching conditions in non-telephone conditions. The authors hypothesize that the conditions 'video' and 'interior surveillance' will challenge the FASR application most, as speaking styles and recording conditions often vary strikingly within these conditions and typically differ quite strongly from other conditions.

References

Drygajlo, A., Jessen, M., Gfroerer, S., Wagner, I., Vermeulen, J. & Niemi, T. (2015). *Methodological guidelines for best practice in forensic semiautomatic and automatic speaker recognition.* Frankfurt: Verlag für Polizeiwissenschaft. [http://enfsi.eu/wp-content/uploads/2016/09/ guidelines_fasr_and_fsasr_0.pdf]

Gold, E. & French, P. (2019). International practices in forensic speaker comparisons: second survey. *International Journal of Speech, Language and the Law 26:* 1-20.

Jessen, M. (2018). Forensic voice comparison. In: J. Visconti (Ed.) *Handbook of communication in the legal sphere.* Berlin: Mouton de Gruyter, pp. 219-255.

Morrison, G. S. & Enzinger, E. (2019). Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (forensic_eval_01)-Conclusion. *Speech Communication* 112: 37-39.

Rose, P. (200*2). Forensic speaker identification.* London: Taylor & Francis.

Solewicz, Y. A., Jessen, M. & van der Vloed, D. (2017). Null-hypothesis LLR: A proposal for forensic automatic speaker recognition. *Proceedings of Interspeech 2017* (Stockholm, Sweden), pp. 2849-2853.

Van der Vloed, D., Bouten, J. & Van Leeuwen, D. A. (2014). NFI-FRITS: A forensic speaker recognition database and some first experiments. *Proceedings of ODYSSEY 2014* (Joensuu, Finland), pp. 6-13.

Van der Vloed, D., Kelly, F. & Alexander, A. (2020). Exploring the effects of device variability on forensic speaker comparison using VOCALISE and NFI-FRIDA, a forensically realistic database. *Proceedings of ODYSSEY 2020* (Tokyo, Japan), pp. 402-407.

Wagner, I. (2019). Examples of casework in forensic speaker comparison. *Proceedings of the 19th International Congress of Phonetic Sciences* (Melbourne, Australia), pp. 721-725.