Comparing automatic and semi-automatic methods on a corpus of forensically authentic telephone conversations

Michael Jessen[1] and Ewald Enzinger[2]
[1]*Department of Speaker Identification and Audio Analysis, BKA, Germany*
michael.jessen@bka.bund.de
[2]*Acoustics Research Institute, Austrian Academy of Sciences, Austria*
ewald.enzinger@oeaw.ac.at

In order to obtain a realistic estimate of the speaker-discriminatory performance of a method or system under the conditions of forensic casework, it is important that evaluations are performed based on speech material reflecting the actual technical and behavioural conditions which are regularly encountered in casework. For this purpose the corpus GSF 1.0 (German Forensic Speech Corpus) was compiled by the BKA (see Becker, 2012; Solewicz et al., 2012; Becker et al., 2012). A new German forensic speech corpus has been compiled more recently by the author within the frame of a research project aiming at a systematic comparison between an automatic system and various semi-automatic methods, i.e. methods based on human selected and corrected acoustic events and features, particularly single-vowel formants, long-term formants, formant dynamics and segmental cepstra (see Morrison, 2011; Rose, 2013; Gold, 2014, among others). The new corpus adopts the comparison data of nine speakers from GSF 1.0 and recruits another fourteen from speakers encountered in casework between 2010 and 2013. The evaluations that are based on this corpus consist of 23 same-speaker comparisons and 506 different-speaker comparisons. The UBM (Universal Background model) used for the evaluations consists of 25 speakers and was based on the same type of real-case telephone conversations as the recordings used in the comparisons. Selection criteria and preparation principles for this new corpus include the following:

- All recordings come from natural telephone conversations by male adult speakers, usually obtained by telephone interception.
- The language spoken in the recordings is German, including regional varieties (no strong traditional dialects) as well as foreign-accented and ethnolectal German varieties.
- The recordings contain various levels of emotional involvement and increased vocal effort, although samples with extreme levels are excluded.
- The minimum net duration was at 20 seconds for both questioned and suspect speakers.
- The recordings differ in technical quality aspects such as noise level and distortion, but recordings with extreme disturbances were excluded.

At the initial stage of the investigation, an evaluation of long-term formants using VOCALISE (http://www.oxfordwaveresearch.com/j2/products/vocalise) has been carried out. Long-term formant analysis using F1, F2, and F3 yielded 18.5% EER on the new corpus. Including the bandwidths of these formants lead to no further improvement. As can be expected, performance is lower compared to high-quality telephone-transmitted laboratory speech, where the EER was found to be about 9% for F1 to F3 and 5% when including bandwidths (Jessen et al. 2014). Further semi-automatic methods, as mentioned above, have been evaluated. The performance of the different semi-automatic methods will be compared individually and under mutual fusion as well as compared and fused with automatic speaker recognition.

Acknowledgement

The work of Felix Jungnick in performing vowel and consonant annotation is gratefully acknowledged.

References

Becker, T. (2012). *Automatischer forensischer Stimmenvergleich*. Norderstedt: Books on Demand.

Becker, T., Y. Solewicz, G. Jardine and S. Gfrörer (2012). Comparing automatic forensic voice comparison systems under forensic conditions. *Proceedings of the Audio Engineering Society 46th International Conference*, Denver, 197–202.

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

Jessen, M., A. Alexander and O. Forth (2014): Forensic voice comparisons in German with phonetic and automatic features using VOCALISE software. *Proceedings of the Audio Engineering Society 54th International Conference*, London, 28–35.

Morrison, G. S. (2011). A comparison of procedures for the calculation of forensic likelihood ratios from acoustic-phonetic data: Multivariate kernel density (MVKD) versus Gaussian mixture model - universal background model (GMM-UBM). *Speech Communication,* **53**, 242–256.

Rose, P. (2013). More is better: likelihood ratio-based forensic voice comparison with vocalic segmental cepstra frontends. *The International Journal of Speech, Language and the Law*, **20**, 77–116.

Solewicz, Y, T. Becker, G. Jardine and S. Gfroerer (2012). Comparison of speaker recognition systems on a real forensic benchmark. *Proceedings of Odyssey 2012*, Singapore.