Comparing MVKD and GMM-UBM applied to a corpus of formant-measured segmented vowels in German

Michael Jessen
Department of Speaker Identification and Audio Analysis, Bundeskriminalamt, Germany
michael.jessen@bka.bund.de

Currently, the two common methods of obtaining likelihood ratios for the purpose of system evaluations in forensic voice comparisons are the MVKD approach, which was originally proposed by Aitken & Lucy (2004), and the GMM-UBM approach, which was originally proposed within the context of automatic speaker recognition. The MVKD approach has been developed for **token based** scenarios. For example, formant frequencies are measured at the center of about ten tokens of a vowel category per recording (e.g. Morrison et al. 2011). The GMM-UBM approach has been developed for **data-stream-based** scenarios. This applies to MFCC feature vectors used in automatic speaker recognition, which are extracted as a data stream with a sampling rate of about ten milliseconds. These data-stream-based scenarios are not limited to automatic speaker recognition but can also be used with acoustic-phonetic data, for example long-term formants, where formant feature vectors are extracted in close temporal succession across vowels (Becker et al. 2008). Occasionally, one of the methods of obtaining likelihood ratios has been used across the scenarios. For example, Morrison (2011) applied both GMM-UBM and MVKD to tokenized data (diphthong contour parameters). However, using GMM-UBM on tokenized data turned out to be not always successful (Zhang et al. 2011; Rose 2013).

In the present experiment the two methods are compared in their "natural habitat", i.e. GMM-UBM with data streams and MVKD with tokens-based data. The speech corpus used for this purpose is a mobile-phone transmitted portion of Pool 2010 (Jessen et al. 2005) in which 21 male adult speakers of the West-Central regional variety of German spoke in a spontaneous style, which was compared to them speaking in a semi-spontaneous style (Jessen et al. 2013 for further details). Recordings with net durations between 20 and 40 seconds were segmented for the vowels /I/ (short/lax i), /a/ and /@/ (schwa) and measured for F1, F2 and F3. Token-based data were extracted using the point-labeling facility of Praat (labeling a vowel at a point minimally influenced by context) and stream-based data by interval labeling (labeling a vowel from beginning to end). The label information was exported to Wavesurfer, where the formant tracking and manual correction were carried out. MVKD was applied based on the implementation by Morrison (http://geoff-morrison.net/) and GMM-UBM was applied based on VOCALISE (http://www.oxfordwaveresearch.com/j2/vocalise), including its region-conditioning tool SPARSE (Jessen et al. 2014 for examples). The likelihood scores obtained with these methods subsequently underwent calibration and fusion. Some of the results are shown in Figure 1. It shows that MVKD and GMM-UBM, when used in their "natural habitat", have similar performance, although the results of GMM-UBM were mostly better under fusion. Figure 1 also shows that different vowels yield different patterns. For example**,** schwa has the lowest performance, probably due to its strong coarticulation, hence highest intra-speaker variation. Overall, fusing different vowels leads to improvement, but less strongly than in Morrison et al. (2011). Fusion was also applied between the data shown in Figure 1 and Long-Term Formants F1, F2, F3 (Jessen et al. 2013), which have an EER of 8.85 and C_{llr} of 0.395. However, no systematic improvement in speaker discrimination was obtained.

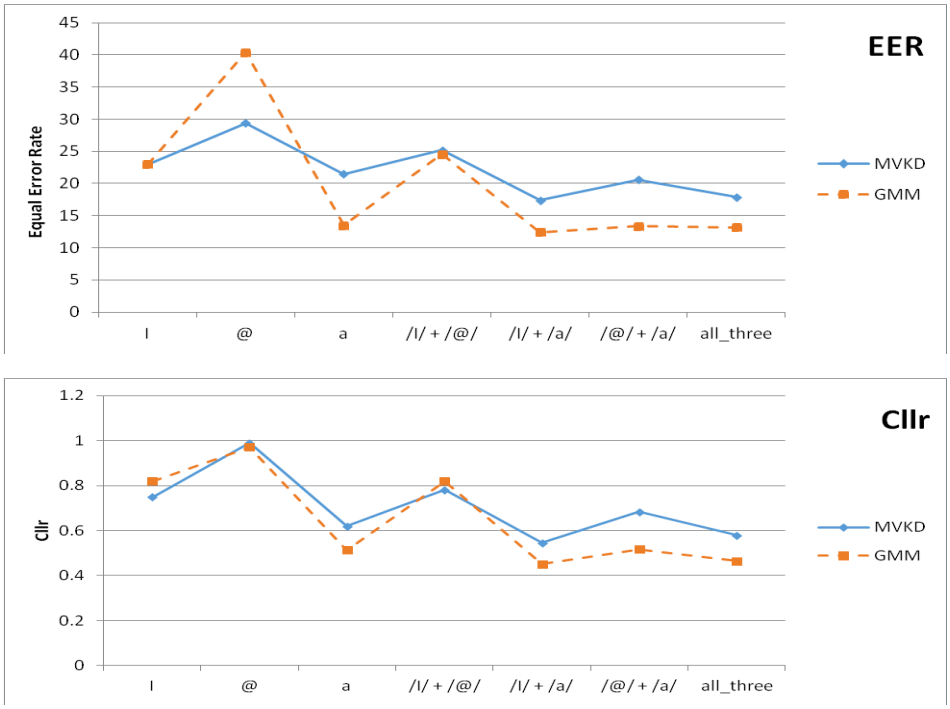

Figure 1 Equal Error rate (upper graph) and C_{llr} (lower graph) using MVKD (uninterrupted lines) and GMM-UBM (interrupted lines) for the three vowels individually (first three entries on x-axis) and fusion between different vowels (remaining entries) on vowel-segmented data from the Pool 2010 corpus.

Acknowledgement: The work of Nicola Wagner in performing vowel annotation and formant measurements and the one of Ewald Enzinger in performing MVKD analysis, calibration, fusion and related forensic-statistic procedures is gratefully acknowledged.

References

Aitken, C.G.G. and D. Lucy (2004). Evaluation of trace evidence in the form of multivariate data. *Applied Statistics*, **53**, 109–122.

Becker, T., M. Jessen and C. Grigoras (2008). Forensic speaker verification using formant features and Gaussian mixture models. *Proceedings of INTERSPEECH '08*, Brisbane, 1505–1508.

Jessen, M., O. Köster and S. Gfroerer (2005). Influence of vocal effort on average and variability of fundamental frequency. *International Journal of Speech, Language and the Law*, **12**, 174–213.

Jessen, M., E. Enzinger & M. Jessen (2013). Experiments on Long-Term Formant Analysis with Gaussian Mixture Modeling using VOCALISE. *Paper presented at the IAFPA Conference*, 2013, Tampa, Fl.

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, June 12–14, pp. 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.

Morrison, G. S., C. Zhang & P. Rose (2011). An empirical estimate of the precision of likelihood ratios from a forensic-voice-comparison system. *Forensic Science International*, **208**, 59–65.

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

Zhang, C., G.S. Morrison and T. Thiruvaran (2011). Forensic voice comparison using Chinese /iau/. *Proceedings of the International Congress of Phonetic Sciences*, Hong Kong, pp. 2280–2283.