Experiments on Long-Term Formant Analysis with Gaussian Mixture Modeling using VOCALISE

Michael Jessen[1], *Ewald Enzinger*[2] *and Marianne Jessen*[3]
[1]*Department of Speaker Identification and Audio Analysis, Bundeskriminalamt, Germany.*
michael.jessen@bka.bund.de
[2]*Acoustics Research Institute, Austrian Academy of Sciences, Vienna, Austria*
ewald.enzinger@oeaw.ac.at
[3]*Stimmenvergleich, Wiesbaden, Germany*
jessen@stimmenvergleich.de

Long-Term Formant Analysis is one of the possible ways of making use of the speaker-discriminatory abilities of formant frequencies in a forensic context (Nolan & Grigoras 2005). When plotted in the form of histograms, the LTF results for a given audio file of a speaker often have distributions that are too complex to be properly modeled by single Gaussians. This is particularly the case for F2, but in a more limited sense also for F1 and F3. A richer way of capturing this information than the use of single Gaussians is the use of Gaussian Mixture Models (GMM) (Reynolds 1992). These are also applied in multivariate fashion, thereby taking into account the correlations between the formants. LTF-modeling with GMM has been applied previously by Becker et al. (2008, 2009) and promising performance has been shown.

The current study is based on data from the Pool 2010 corpus (Jessen et al. 2005). For the experiments, speech from 22 male adult speakers of the West-Central regional variety of German was used. Questioned-speaker samples were taken from a condition of Pool 2010 (recently extracted and not previously used) where subjects expressed various thoughts and feelings about the recording situation; this condition is referred to as the **sp**ontaneous speech condition. Suspect-speaker samples were taken from a condition in which the suspects explained pictures to a conversation partner while under the instruction to avoid certain words. This condition is of a similar (reduced) level of naturalness as elicited in many forensic suspect recordings and will be called the **s**emi**sp**ontaneous speech condition. For the Universal Background model (UBM), speech from 22 further male speakers of the same variety was used, which was spoken in the ssp-condition. All the speech was recorded under high-quality conditions and subsequently transmitted through authentic mobile phone connections. After labeling all vocalic portions with visible F1, F2 and F3 in *Praat*, these portions were extracted, concatenated and saved in a separate file. In these files, formant tracking was performed using *Wavesurfer* in its default settings. Wavesurfer allows for convenient manual correction of formant tracks, which was carried out were necessary. All extraction files (pure vocalic stream) were close to ten seconds long, which is relatively realistic forensically.

GMM modeling and the calculation of likelihood scores was performed using the recently available automatic and phonetic speaker recognition system *VOCALISE* (see References). For all experiments, symmetric testing was used, i.e. the scores obtained comparing the features of the questioned speech with the model suspect speech were averaged with the scores obtained comparing the features of the suspect speech with the model of the questioned speech. The number of Gaussian components used for modeling the suspect recordings and the UBM was varied systematically from 1 to 8. Furthermore, separate analyses were performed on the formant frequencies alone (F1, F2, F3) and on the formant frequencies and their bandwidths (F1, F2, F3, B1, B2, B3). Cross-validation calibration was applied to the output scores and log likelihood ratio cost (C_{llr}) was calculated for each of the 16 system tests. A Tippett plot of one of the tests and

C_{llr}-results for all tests are shown in Fig 1.

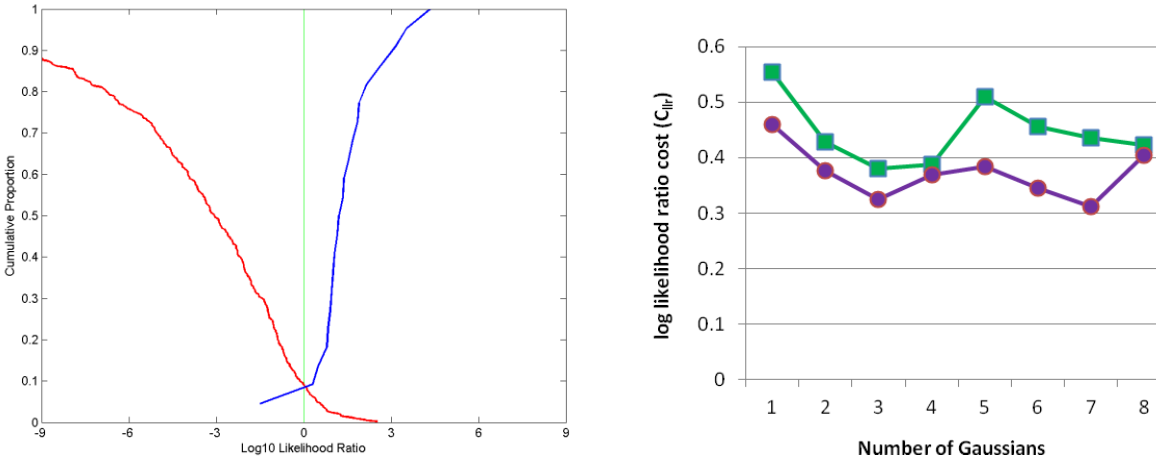

Figure 1 Left: Tippett plot of test with 3 Gaussian components and the full parameter set F1,2,3, B1,2,3. The different-speaker distribution is the one decreasing from left to right in terms of calibrated $Log_{10}LR$, the same-speaker distribution is the one increasing from left to right. **Right**: C_{llr}-values (vertical axis) for each of the 16 system tests, i.e. 8 different numbers of Gaussian components (horizontal axis) and 2 parameter sets (F1,2,3-set with squares, F1,2,3, B1,2,3-set with circles).

Results show that, firstly, the full set of parameters (formant frequencies and bandwidths) lead to better speaker-discriminatory performance in terms of C_{llr} (lower values indicating better performance) than the reduced set of parameters (just frequencies), although with some numbers of Gaussian components this difference is only small. This advantage of the full parameter set is consistent with the results of Becker et al. (2008, 2009). Secondly, overall (across parameter sets) maximum performance was reached with a number of just 3 Gaussian components. This result differs from Becker et al. (2009), where performance increased up to 8 components.

According to the present experiments, three Gaussian components are enough to capture practically all the discriminatory information in long-term formants. Future research could address whether it is advantageous to use a higher number of components for long-term F2 than long-term F1 and F3, since the distribution of the former formant tends to be more complicated impressionistically than the one of the latter two.

References

Becker, T., M. Jessen & C. Grigoras (2008). Forensic speaker verification using formant features and Gaussian Mixture Models. *Proceedings of INTERSPEECH 2008*, Brisbane, 1505–1508.

Becker, T., M. Jessen & C. Grigoras (2009). Speaker verification based on formants using Gaussian mixture models. *Proceedings of NAG/DAGA 2009,* Rotterdam, 1640–1643.

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

Nolan, F. & C. Grigoras (2005). A case for formant analysis in forensic speaker identification. *Intern. J. of Speech, Language and the Law,* **12**, 143–173.

Reynolds, D.A. (1992). *A Gaussian Mixture Modeling approach to text-independent speaker identification*. PhD dissertation, Georgia Institute of Technology.

VOCALISE (Voice Comparison and Analysis of the Likelihood of Speech Evidence), Oxford Wave Research Ltd, Oxford, United Kingdom. www.oxfordwaveresearch.com/j2/vocalise