Speaker Verification Based on Formants Using Gaussian Mixture Models

Timo Becker[1], Michael Jessen[2], Catalin Grigoras[3]

[1] *Acoustics Research Institute of the Austrian Academy of Sciences, Austria, Email: timo.becker@oeaw.ac.at*

[2] *KT 54 Bundeskriminalamt, Germany, Email: michael.jessen@bka.bund.de*

[3] *Ministry of Justice, Romania, Email: forensicav@techemail.com*

Introduction

In automatic speaker verification, the Gaussian Mixture Model (GMM) approach based on cepstral features has been successfully applied during recent years. In the approach presented in this paper, the well known UBM-GMM framework is combined with formant features which have shown a high discrimination ability among speakers in the acoustic-phonetic approach of manual speaker verification. We provide an easy to understand modeling process where the model parameters are related directly to the speakers' typical vocal tract configurations. Additionally, the within-speaker variability is reflected in these parameters. The complexity of the speaker models is low because of a reduced dimensionality compared to standard automatic speaker verification approaches.

Automatic Speaker Verification Using Formants

The approach presented in this paper originates from the idea of modeling long-term distributions of formant features as first proposed by Nolan and Grigoras [13]. Here, formant center frequencies were modeled independently using *long-term formant distributions (LTF distributions)* based on Gaussian kernel density estimation. Apart from such global approaches where the overall distribution of formants is modeled, several approaches dealing with specific phoneme categories exist [9, 12, 19, 20]. The advantage of global approaches is that they might be applied to insufficiently described languages where not enough data about the background population is available.

Formant features, as well as the corresponding bandwidths, correlate with each other [17, 18, 20]. Hence, they must not be modeled separately in a Bayesian framework where a combination of likelihood ratio scores should result from independent observations. The approach presented here models the distribution of the multivariate formant features using the well known UBM-GMM framework [16]. The common cepstral features are replaced by formant center frequency features F1, F2 and F3 and the corresponding bandwidths BW1, BW2 and BW3. This system was first described by Becker et al.[3] and revealed equal error rates ranging from 3% to 10.5%, depending on the features chosen. It was shown that the best results can be achieved using three formants and the corresponding bandwidths. However, here, we chose to focus on this optimal condition (using F1, F2, F3, BW1, BW2, BW3) and also on the common forensic condition, where no reliable F1 features can be obtained due to the proximity of F1 to the lower telephone pass band in some vowels (using F2, F3, BW2, BW3). Henceforth, those conditions will be called *optimal* and *reduced* conditions respectively. The system used is described as follows:

The feature extraction step is accomplished by compiling all n feature vectors for every speaker recording

$$X = \{x_1, \ldots, x_n\}, \tag{1}$$

where every vector

$$x_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{id} \end{pmatrix} \tag{2}$$

is a feature vector of length d[1]. Based on X, speaker models are generated as follows:

The d-variate Gaussian function is given by

$$f(x; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \, e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}, \tag{3}$$

where μ is the mean and Σ is the covariance matrix. d is a positive integer whose value is the feature vector dimensionality. A Gaussian mixture density is a weighted sum of M Gaussian distribution functions $f(x; \mu_i, \Sigma_i)$

$$f(x; \mu_1, \ldots, \mu_M, \Sigma_1, \ldots, \Sigma_M) = \sum_{i=1}^{M} p_i \, f(x; \mu_i, \Sigma_i),$$

$$\sum_{i=1}^{M} p_i = 1, \tag{4}$$

where $p_i \geq 0$ are the mixture weights. Now, a GMM consisting of M Gaussians is specified by

$$\lambda := (p_i, \mu_i, \Sigma_i)_{i=1,\ldots,M} \,. \tag{5}$$

(p_i, μ_i, Σ_i) is a tuple consisting of the model parameters. M is the number of mixture components. Every distribution of d-variate feature vectors is thus specified by λ. Full covariance matrices Σ represent formant feature correlations and account for within-speaker variability. Additionally, a *universal background model (UBM)* is generated from a collection of feature vectors from

[1] $d = 6$ for optimal, $d = 4$ for reduced

many different speakers. For model generation, the free statistical software *R* [14] and the *mclust* package [2, 4, 5, 6] were used. The similarity of feature vectors X from one speaker and a speaker model λ are expressed by the product of the Gaussian mixture density (see Equation 4), the *likelihood*. For feature vectors X (see Equation 1) and a speaker model λ (see Equation 5), the likelihood that the feature vectors come from this model is measured via computation of

$$P(X \mid \lambda) = \prod_{i=1}^{n} f(x_i \mid \lambda), \tag{6}$$

where $f(x_i \mid \lambda)$ is the Gaussian mixture density function for the specified model λ.

Every comparison of a test and training recording is a comparison of the likelihood of the test feature vectors in the speaker model and the UBM. This is expressed in the likelihood ratio

$$LR = \frac{P(X \mid \lambda_{\text{speaker}})}{P(X \mid \lambda_{\text{UBM}})}. \tag{7}$$

The likelihood for λ_{speaker} accounts for similarity, while λ_{UBM} accounts for typicality. A high likelihood ratio supports the hypothesis that the test feature vectors come from the same speaker while a small likelihood ratio supports the hypothesis that the test feature vectors come from different speakers. Here, the log likelihood ratio was used.

Data Base

This study is based on a corpus of speech produced by 68 male adult speakers of German. In this speech corpus, referred to as Pool 2010 [11], read and spontaneous speech was elicited in a neutral condition, a telephone condition and a Lombard condition; only the data from the neutral condition were used. In the spontaneous speech task, which is at the focus of the present investigation, subjects described a series of pictures in a dialogue situation where they had to avoid certain words. The recordings, which were originally made under studio conditions, were played and transmitted through real mobile phone connections. The recordings were edited manually by selecting vowels in which formants 1 to 3 were visible with sufficient clarity. LPC-based formant tracking was applied to this resulting material, and any remaining formant tracking errors were corrected manually. For signal editing and formant tracking, the software Wavesurfer [1] was used. The LPC analysis was set to detect 4 formants, of which only the first three were used[2]. The analysis window length (Hamming) was 0.049 seconds, the LPC order was 12, the preemphasis factor was 0.7, and formant values were obtained every 10 ms. The method of scanning formant data over the entire course of a recording and across all the different vowels is called LTF analysis (long-term formant distribution) and has been proposed in [13]. As a practical advantage of this method, LTF analysis can be applied to languages that are not spoken by the user, because no segmentation into vowels or other phonological units is necessary. All that is required is correct identification of the formants, which is possible cross-linguistically based on a general knowledge of acoustic phonetics.

Formant feature measurements for each speaker were halved to create training and test sets. The test features were additionally halved to increase the number of comparisons and simulate short test recordings. The average duration of the training signals was about 22 seconds and about 11 seconds for the test set.

Population Size

In [3], 18 speakers' formant measurements were used to create the universal background model (UBM) by pooling all measurements together. The UBM parameters were then estimated based on these measurements. The number of mixtures was $M = 8$ for both UBM and single speaker models. This value was determined experimentally. To evaluate the influence of the speaker number in the UBM, all 68 speakers were used in a cross-validation experiment where, for each comparison of two recordings, the UBM was compiled from the non-involved speakers. There were $136 \times 68 = 9248$ tests. 136 of those were same-speaker tests. We varied the number of UBM speakers $U = \{5, 10, 20, 30, 40, 50\}$. The results can be seen in Table 1 and Figure 1.

Table 1: Equal error rates for different numbers U of UBM speakers

U	optimal	reduced
5	0.052	0.081
10	0.039	0.066
20	0.033	0.059
30	0.029	0.047
40	0.031	0.049
50	0.030	0.052

It can be seen that the discrimination ability can be optimised to about 3% EER for the optimal condition and to about 5% EER for the reduced condition, when using 30 or more speakers for the UBM. For $U \geq 40$, the EERs are slightly higher. The differences are quite small and might be random. These results are in accordance with those from Ishihara and Kinoshita [10], where a significant improvement of the EER was observed for up to 30 speakers, when using multivariate F0 features. For this reason, we henceforth use $U = 30$.

[2] Although F4 was mostly unreliable due to the limited telephone pass band, F4 detection turned out to be useful as a safeguard to prevent errors in the automatic tracking of F3.

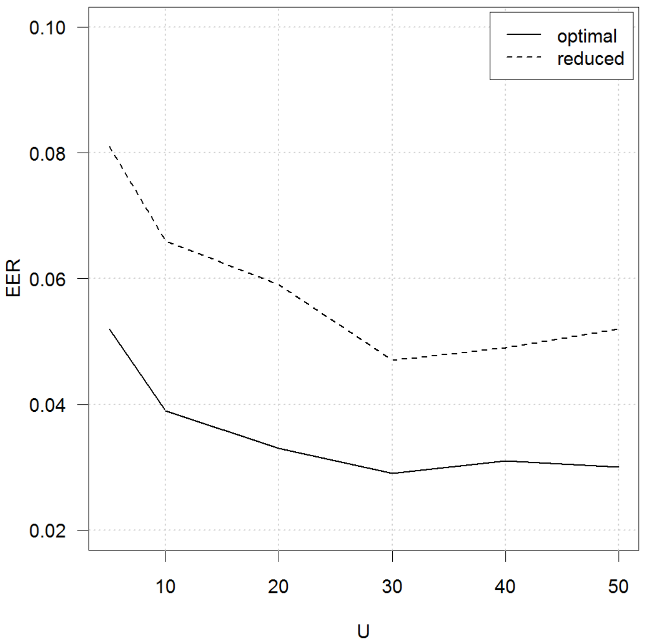

Figure 1: Equal error rates for different numbers of UBM speakers U

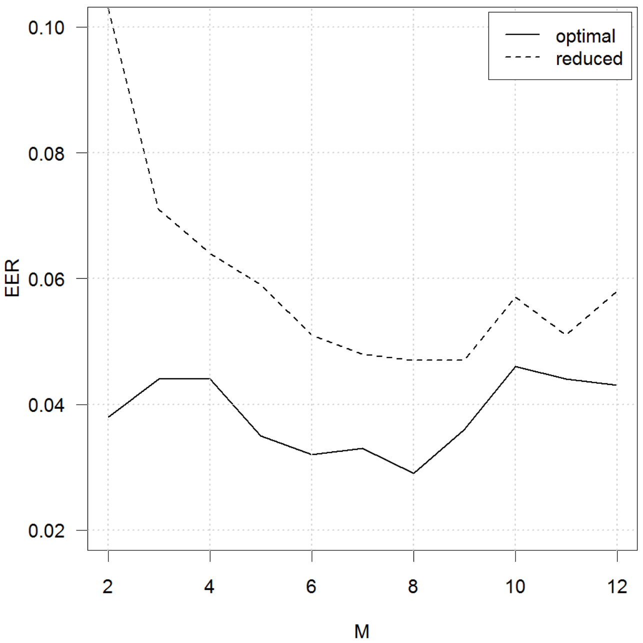

Figure 2: Equal error rates for different numbers of mixture components M

Number of Mixture Components

The number of mixture components $M = 8$ was determined experimentally in Becker et al.[3]. Here, we used $M = \{2, \ldots, 12\}$ to look at the influence of this parameter. As in the experiment concerning the number of UBM speakers U, there were 9248 tests in total. See Table 2 and Figure 2 for the results.

Table 2: Equal error rates for different M

M	optimal	reduced
2	0.038	0.103
3	0.044	0.071
4	0.044	0.064
5	0.035	0.059
6	0.032	0.051
7	0.033	0.048
8	0.029	0.047
9	0.036	0.047
10	0.046	0.057
11	0.044	0.051
12	0.043	0.058

It can be seen that the discrimination ability reaches a minimum at about $M = 8$ for both the optimal and reduced condition. The parameter M could not be further improved. The proposed best framework hence uses a model with a low complexity (state-of-the-art automatic speaker verification systems usually rely on feature vectors with about 38 dimensions, and models with 2048 mixture components [15, 7, 8]).

Conclusions

A new approach using the well known automatic UBM-GMM framework with semi-automatically extracted formant features was explored. In both, the optimal and reduced conditions, the system revealed high discriminatory abilities for male German speakers. Using models with 8 mixture components and a collection of at least 30 speakers lead to optimal results. The data base consisting of telephone-transmitted speech with short durations represents realistic data for forensic applications.

However, the effects of non-contemporaneous speech, as well as different speaking styles, still have to be investigated in respect to forensic applications. Calibration of scores and comparisons with other automatic approaches will have to be conducted as well.

The modeling of speakers imposes a low complexity and a high interpretability: low dimensional models (4 or 6 dimensions, depending on the application) are based on formant features and their variabilities, and hence enable a direct relation of speakers' vocal tract configurations to the models.

Acknowledgments

Anja Moos at Saarland University performed the formant measurements as part of her MA research and was supported by a research grant from the International Association for Forensic Phonetics and Acoustics (IAFPA). Thanks to Anja and the IAFPA.

References

[1] `http://www.speech.kth.se/wavesurfer/`.

[2] Jeffrey D. Banfield and Adrian E. Raftery. Model-Based Gaussian and Non-Gaussian Clustering. *Biometrics*, 49:803–821, 1993.

[3] Timo Becker, Michael Jessen, and Catalin Grigoras. Forensic Speaker Verification Using Formant Features and Gaussian Mixture Models. In *Proceedings*

of Interspeech 2008 incorporating SST'08, pages 1505–1508, Brisbane, September 2008. ISCA.

[4] Chris Fraley and Adrian E. Raftery. MCLUST: Software for model-based cluster analysis. *Journal of Classification*, 16:297–306, 1999.

[5] Chris Fraley and Adrian E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97:611–631, 2002.

[6] Chris Fraley and Adrian E. Raftery. MCLUST Version 3 for R: Normal Mixture Modeling and Model-Based Clustering. Technical Report 504, University of Washington, Department of Statistics, September 2006.

[7] J. Gonzalez-Rodriguez, J. Fiérrez-Aguilar, J. Ortega-Garcia, and J. J. Lucena-Molina8. *Biometric Identification in Forensic Cases According to the Bayesian Approach*, chapter Biometric Identification in Forensic Cases According to the Bayesian Approach, pages 177–185. Springer Berlin / Heidelberg, 2002.

[8] Joaquin Gonzalez-Rodriguez, Daniel Ramos-Castro, Doroteo Torre Toledano, Alberto Montero-Asenjo, Javier Gonzalez-Dominguez, Ignacio Lopez-Moreno, Julian Fierrez-Aguilar, Daniel Garcia-Romero, and Javier Ortega-Garcia. Speaker Recognition The ATVS-UAM System at NIST SRE 05. *IEEE Aerospace and Electronic Systems Magazine*, 22(1):15–21, January 2007.

[9] Catalin Grigoras. Forensic Voice Analysis Based on Long Term Formant Distributions. In *4th European Academy of Forensic Science Conference*, June 2006.

[10] Shunichi Ishihara and Yuko Kinoshita. How Many Do We Need? Exploration of the Population Size Effect on the Performance of Forensic Speaker Classification. In *Proceedings of Interspeech 2008 incorporating SST'08*, pages 1941–1944, 2008.

[11] Michael Jessen, Olaf Köster, and Stefan Gfroerer. Influence of vocal effort on average and variability of fundamental frequency. *Speech, Language and the Law*, 12(2):174–213, 2005.

[12] Geoffrey Stewart Morrison and Yuko Kinoshita. Automatic-Type Calibration of Traditionally Derived Likelihood Ratios: Forensic Analysis of Australian English /o/ Formant Trajectories. In *Proceedings of Interspeech 2008 incorporating SST'08*, pages 1501–1504, 2008.

[13] F. Nolan and C. Grigoras. A case for formant analysis in forensic speaker identification. *Speech, Language and the Law*, 12(2):143–173, 2005.

[14] R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[15] Douglas A. Reynolds, Joseph P. Campbell, T. Gleason, C. Quillen, D. Sturim, P. Torres-Carrasquillo, and A. Adami. The 2004 MIT Lincoln Laboratory Speaker Recognition System. In *Proc. ICASSP*, pages I–177–180, 2005.

[16] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10:19–41, 2000.

[17] Phil Rose. Forensic speaker recognition at the beginning of the twenty-first century - an overview and a demonstration. *Australian Journal of Forensic Sciences*, (37):49–72, 2005.

[18] Phil Rose. Accounting for correlation in linguistic-acoustic likelihood ratio-based forensic speaker discrimination. In *IEEE Odyssey 2006: The Speaker and Language Recognition Workshop*, 2006.

[19] Phil Rose. The intrinsic forensic discriminatory power of diphtongs. In *Proceedings of the Eleventh Australasian International Conference on Speech Science and Technology*, pages 64–69, 2006.

[20] Philip Rose. *Forensic Speaker Identification.* Taylor & Francis, 2002.