Dimension-Decoupled Gaussian Mixture Model for Short Utterance Speaker Recognition

Thilo Stadelmann and Bernd Freisleben
Dept. of Mathematics & Computer Science, University of Marburg, Germany
{*stadelmann, freisleb*}*@informatik.uni-marburg.de*

Abstract

The Gaussian Mixture Model (GMM) is often used in conjunction with Mel-frequency cepstral coefficient (MFCC) feature vectors for speaker recognition. A great challenge is to use these techniques in situations where only small sets of training and evaluation data are available, which typically results in poor statistical estimates and, finally, recognition scores. Based on the observation of marginal MFCC probability densities, we suggest to greatly reduce the number of free parameters in the GMM by modeling the single dimensions separately after proper preprocessing. Saving about 90% of the free parameters as compared to an already optimized GMM and thus making the estimates more stable, this approach considerably improves recognition accuracy over the baseline as the utterances get shorter and saves a huge amount of computing time both in training and evaluation, enabling real-time performance. The approach is easy to implement and to combine with other short-utterance approaches, and applicable to other features as well.

1. Introduction

Furui [2] stated that one of the two major challenges in automatic speaker recognition today is to cope with the lack of available data for training and evaluating speaker models. For instance, in automatic speaker indexing and diarization of multimedia documents, unsupervised speaker clustering has to deal with the output of speaker segmentation algorithms that chop the signal into chunks of typically less than 2-3 seconds. In speaker verification and identification, enrollment and evaluation data is expensive in the sense that the system should bother a user as little as possible. This is in conflict with the general finding that one needs 30-100 seconds of training data to build a state-of-the art model of high quality that can be evaluated using approximately 10 seconds of test data. This state-of-the-art model refers to the Gaussian mixture model (GMM) approach with diagonal covariance matrices used in almost all current systems, together with Mel-frequency cepstral coefficient (MFCC) feature vectors [7][3].

Several approaches exist in the literature to cope with short utterances. For instance, Merlin et al. propose to work in an explicit speaker feature space in order to overcome the intra-speaker variability omnipresent in acoustic features due to the phonetic structure of speech [6]; less ambiguity and variability in the transformed space is believed to lead to more stable model estimates with less training data. A prototypical implementation of the acoustic space transformation via projection to anchor model scores shows promising results. Larcher and his colleagues criticize the GMM (with universal background model, UBM) approach for its insufficiency for mobile appliances in terms of data demands for training and evaluation [4]. Their solution includes using temporal structure information (i.e. word dependency) and multimodal information (video) to compensate for short training and evaluation samples. Vogt et al. use a factor analysis technique to arrive at subspace models that work well with short training utterances and can be seamlessly combined with the optimal GMM-UBM model when plenty of training data is available [8]. In subsequent work, they suggest to estimate confidence intervals for speaker verification scores, leading to accurate verification decisions after only 2-10 seconds of evaluation data where usually 100 seconds are needed [10][9].

In this paper, we present a different approach to address the problem of small sets of training and evaluation data: we propose a novel way to reduce the number of necessary free parameters in the GMM with the aim of obtaining more stable statistical estimates of model parameters and likelihoods using less data. Furthermore, better—i.e., closer to truth—estimates improve recognition accuracy, and less complex models have a

strong positive effect on runtime, too. Additionally, the approach can be combined with other short utterance approaches proposed in the literature.

The paper is organized as follows. In Section 2, the motivation for our approach is explained by looking at feature distributions. Section 3 introduces the dimension-decoupled GMM (DD-GMM). In Section 4, experimental results are presented. Section 5 concludes the paper and outlines areas for future work.

2. Feature Distributions

Consider the plot[1] of a diagonal covariance GMM with 32 mixtures in Figure 1, trained on the set of 19-dimensional MFCCs extracted from 52.52 seconds of anchor speech from a German news broadcast.

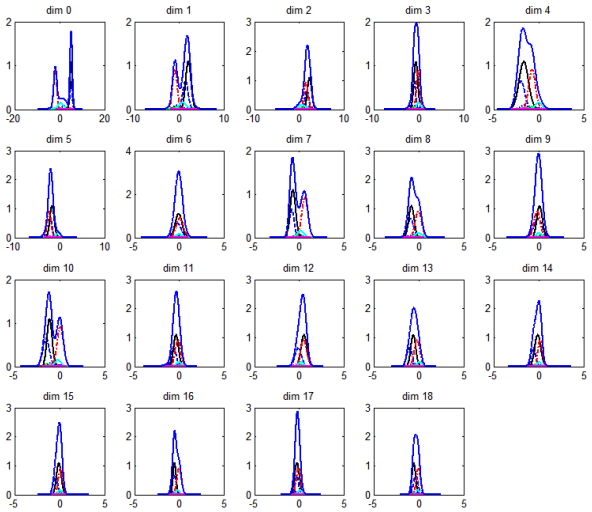


Figure 1. A 32-mixture GMM of 19-dimensional MFCCs. The topmost solid blue line shows the joint marginal density.

While the first several coefficients show a multimodal or skewed distribution, many of the later dimensions look more Gaussian-like. Different feature types like linear prediction-based cepstrum coefficients (LPCC) show a similar characteristic. Others, like line spectral pairs (LSP) or filterbank energies are more Gaussian-like in any of their dimensions, while Pitch's single dimension is quite non-Gaussian. In combination, the marginal densities of most practical feature sets exhibit a similar structure as shown in Figure 1.

This leads us to the following reasoning: different coefficients obviously have different distributions, so different (often small) numbers of 1D Gaussian mixtures are necessary to approximate their true marginal density. In contrast, a standard GMM uses a certain (high) number of multivariate mixtures, giving equal modeling power to each dimension, also to those with very simple marginal densities. Visual inspection suggests: several parameters could be saved by modeling the dimensions independently, i.e. decoupling the number of mixtures for one dimension from the number of mixtures for any other dimension.

Accomplished in a straightforward fashion, to decouple the dimensions means to fit a one-dimensional GMM to each dimension of the feature vectors instead of training a single multimodal mixture model; the complete model would then be the ordered set of univariate GMMs, renouncing to model any interrelation of the marginals. In fact, practical GMMs use diagonal covariance matrices, assuming that the features are uncorrelated (as is the case with MFCCs) or that this information is unimportant or can be modeled via more mixture components. The only correlation information still possibly present in such a multivariate model is introduced by the training procedure: a complete (multivariate) mixture is always trained based on complete (multivariate) feature vectors. Thus, the togetherness of values for different dimensions in one mixture component allows inferring the co-occurrence of these values in the training data. However, this information is currently not used for speaker recognition and might only play a role in speech synthesis.

3. The Dimension-Decoupled GMM

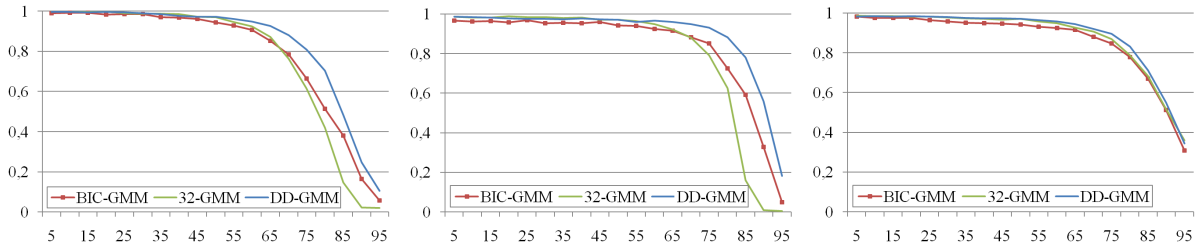
The Dimension-Decoupled GMM (DD-GMM) λ_{DD} can be formalized as follows:

$$\lambda_{DD} = \{(M_d, \lambda_d) \,|1 \leq d \leq D\} \cup \{\Omega\} \quad (1)$$

$$\lambda_d = \{(w_m, \mu_m, \sigma_m) \,|1 \leq m \leq M_d\} \quad (2)$$

The DD-GMM is essentially a set of tuples, one for each dimension d within the dimensionality D of the feature vectors. Each tuple contains an univariate GMM λ_d and the number of mixtures M_d in this dimension. Ω is the matrix of eigenvectors of the covariance matrix of the training data, used to perform an orthogonal transformation on the (training and evaluation) data prior to modeling/recognition in order to further decorrelate the features and thus to justify the decoupled modeling, as suggested by Liu and He [5]. After transforming the training set this way, each λ_d is trained on only the d^{th} dimension of the training vectors using the standard expectation maximization-based maximum likelihood training procedure. The optimal number of mixtures M_d for each model is estimated via the Bayesian

[1] Produced with *PlotGMM*, see `http://www.informatik.uni-marburg.de/~stadelmann/eidetic.html`.



(a) %training/evaluation data removed. (b) %training data removed (eval. data fix). (c) %eval. data removed (training data fix).

Figure 2. Speaker identification accuracy (vertical) vs. changing data availability conditions.

information criterion (BIC) by training all different candidate models with $1 \leq m \leq 32$ mixtures, penalizing the likelihood of the training data with their number of parameters and choosing the candidate that maximizes the BIC score [1]. The model is evaluated, then, on the Ω-transformed evaluation set of feature vectors by calculating the likelihood l according to Equation (3):

$$l = \prod_{d=1}^{D} \prod_{t=1}^{T} \sum_{m=1}^{M_d} w_m \cdot \mathcal{N}\left(x(t)_d, \mu_m, \sigma_m\right) \qquad (3)$$

Here, $x(t)_d$ is the d^{th} dimension of the t^{th} feature vector from overall T vectors, $\mathcal{N}(..)$ is the univariate normal distribution, and w_m, μ_m and σ_m are the weight- mean- and standard deviation-parameters of the m^{th} mixture in GMM λ_d, respectively.

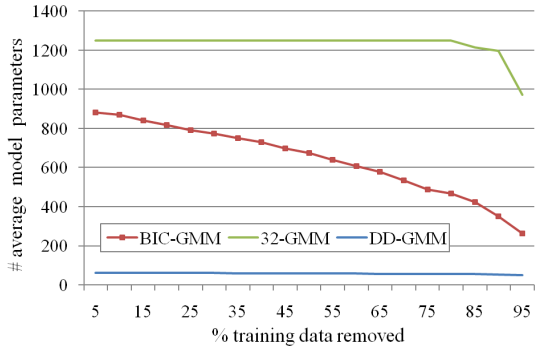
We have implemented the DD-GMM within our C++ class library `sclib` as a mere wrapper around the existing GMM class; using Liu and He's code [5] for the orthogonal transform, the essential parts constitute less than 80 lines of code. On the one hand, this leaves room for speed optimizations (e.g. by integrating the DD-GMM with the GMM); on the other hand, this shows that the approach can be integrated with any existing GMM implementation.

4. Experimental Results

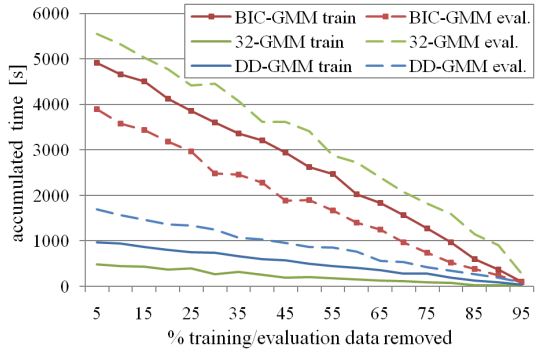
We have conducted several experiments to validate that the DD-GMM improves speaker recognition performance while saving free parameters and reducing computational cost. Reynolds' experimental speaker identification scenario is used as our basic setting [7]: The 630 speakers of the TIMIT database are split into a training- and a separate test set, leading to an average of 21.67/5.09 seconds of training/evaluation utterance length. The minimum and maximum length are 14.57/2.93 and 33.54/8.18 seconds, respectively, leading to a standard deviation of 2.82/0.90 seconds for the two phases. The utterances are transformed to MFCC feature vectors (20ms frames with 50% overlap, coefficients 1–19 discarding the 0^{th}). For the 630 training utterances, models are built a priori, then an identification experiment is run for the 630 test utterances. As models we use the standard 32-mixtures GMM from Reynolds (32-GMM in the figures) and a BIC-tuned multivariate GMM with $1 \leq m \leq 32$ mixtures (BIC-GMM) as baselines to compare with our DD-GMM. To simulate various short utterance conditions, we reduce the training- and/or evaluation data lengths in steps of 5% from 95% of their original length to 5% and observe the corresponding models' behavior.

First, Figure 2(a) shows speaker identification accuracy for all three models as training and evaluation utterance length drops simultaneously. While until 45% reduction the models' identification performance is about the same (with the 32-GMM having a small advantage), the DD-GMM then outperforms the other two competitors clearly. With $\geq$ 50% reduction, the DD-GMM performs on the average 7.56% better then the best competitor using the same amount of data (vertical distance), while it achieves similar recognition scores as the best competitor with an average of 4.17% less data (horizontal distance) in this general short utterance case. This effect increases in the case of merely reducing training data (with evaluation data fixed at reasonable 50%), as depicted in Figure 2(b), while it is smaller, yet still visible, when only the evaluation data rate drops, as in Figure 2(c). This validates the dimension-decoupled modeling scheme at least for MFCC features.

Second, Figure 3(a) shows the evolution of the parameter count in the three model types as the utterances get shorter. The drop in 32-GMM's parameters towards the end is due to the fact that here the amount of data is too small to find even enough cluster centers for mixture candidates. Thus, the mixture count is reduced in this case until a valid model can be trained. Besides this anomaly, the figure shows the efficiency of the DD-GMM in reducing the number of free parameters in the model, even more so in comparison with the standard parameter optimization via the BIC: the saving here still



(a) . . . number of model parameters.



(b) . . . computing time and computational costs.

Figure 3. Effect of short utterances on. . .

constitutes 90.98% on the average. For comparison, Liu and He achieved a parameter saving of about 75% using their orthogonal GMM without additionally enabling short utterance support or boosting accuracy [5].

Finally, runtime plots given in Figure 3(b) show the computational efficiency of our approach: due to the BIC parameter search (training essentially 32 times as many models as for the 32-GMM), the DD-GMM's training time is on the average 2.3 times longer than for the 32-GMM, but 5.1 times faster than with the BIC-GMM and still 13.5 times faster than real-time. In the evaluation phase that occurs more often in practice, the DD-GMM is the fastest, outperforming the BIC-GMM and 32-GMM by a factor of 2.1 and 3.6, respectively, taking only 54.5% of real-time.

5. Conclusion

We have presented the dimension-decoupled GMM as a novel approach to cope with short (training and evaluation) utterances in speaker recognition tasks. In the case of lacking data, the DD-GMM gives more reliable results (i.e. higher accuracy) than the baselines, while it is computationally more efficient even in the case of having plenty of data, where it also gives competitive accuracy. The DD-GMM allows to recognize speakers in regions where baseline GMM approaches are not usable anymore (i.e. more than 80% recognition accuracy with less than 5.5 seconds of training- and 1.3 seconds of evaluation data). At the same time, our approach can easily be integrated into other short utterance schemes, allowing for synergetic effects, and can straightforwardly be implemented in any GMM environment. Areas for future work are: testing the DD-GMM with other feature types, evaluating its performance using further data sets, and applying it in other domains than speaker recognition.

Acknowledgments

This work is funded by the Deutsche Forschungsgemeinschaft (SFB/FK615, Project MT).

References

[1] S. S. Chen and P. Gopalakrishnan. Clustering via the Bayesian Information Criterion with Applications in Speech Recognition. In *Proc. of ICASSP'98*, volume 2, pages 645–648, Seattle, WA, USA, May 1998.

[2] S. Furui. 40 Years of Progress in Automatic Speaker Recognition. In *Proc. of the* 3^{rd} *Int. Conf. on Advances in Biometrics*, pages 1050–1059, 2009.

[3] T. Kinnunen and H. Li. An Overview of Text-Independent Speaker Recognition: from Features to Supervectors. *Speech Communication*, 52:12–40, 2010.

[4] A. Larcher, J.-F. Bonastre, and J. Mason. Short Utterance-based Video Aided Speaker Recognition. In *Proc. of IEEE* 10^{th} *Workshop on Multimedia Signal Processing*, pages 897–901, 2008.

[5] L. Liu and J. He. On the Use of Orthogonal GMM in Speaker Recognition. In *Proc. of ICASSP'99*, volume 2, pages 845–848, Phoenix, AZ, USA, 1999.

[6] T. Merlin, J.-F. Bonastre, and C. Fredouille. Non Directly Acoustic Process for Costless Speaker Recognition and Indexation. In *Int. Workshop on Intelligent Communication Technologies and Applications*, 1999.

[7] D. A. Reynolds. Speaker Identification and Verification using Gaussian Mixture Speaker Models. *Speech Communication*, 17:91–108, 1995.

[8] R. Vogt, C. J. Lustri, and S. Sridharan. Factor Analysis Modelling for Speaker Verification with Short Utterances. In *Proc. of Odyssey 2008: The Speaker and Language Recognition Workshop*, Stellenbosch, South Africa, January 2008.

[9] R. Vogt and S. Sridharan. Minimising Speaker Verification Utterance Length through Confidence Based Early Verification Decisions. *Lecture Notes in Computer Science*, 5558/2009:454–463, 2009.

[10] R. Vogt, S. Sridharan, and M. Mason. Making Confident Speaker Verification Decisions with Minimal Speech. In *Proc. of Interspeech*, pages 1405–1408, Brisbane, Australia, September 2008.