Optimization of Gabor Features for Text-Independent Speaker Identification

Volker Mildner, Stefan Goetze and Karl-Dirk-Kammeyer
University of Bremen,
Dept. of Communications Engineering,
D-28334 Bremen, Germany
Email: goetze@ant.uni-bremen.de

Alfred Mertins
Carl von Ossietzky-University, Oldenburg,
Signal Processing Group,
D-26111 Oldenburg, Germany
Email: alfred.mertins@uni-oldenburg.de

***Abstract*— For text-independent speaker identification a prominent combination is to use Gaussian Mixture Models (GMM) for classification while relying on Mel-Frequency Cepstral Coefficients (MFCC) as features. To take temporal information into account the time difference of features of adjacent speech frames are appended to the initial features. In this paper we investigate the applicability of spectro-temporal features obtained from Gabor-Filters and present an algorithm for optimizing the possible parameters. Simulation results on a database show that spectro-temporal features achieve higher recognition rates than purely temporal features for clean speech as well as for disturbed speech.**

I. Introduction

To realize text-independent speaker identification a general approach is to classify speech sequences via Gaussian Mixture Models (GMM) as introduced by Reynolds [1], [2]. Based on features extracted from training sequences of a single speaker the model parameters of the GMM are estimated via the EM algorithm [3]. For closed-set classification features extracted from a test-sequence are compared to the GMMs of all speakers and the model yielding the highest likelihood is assumed to indicate the speaker of the test sequence. Common features in speech processing are the well known Mel Frequency Cepstral Coefficients (MFCC). As shown previously in [1], [2] they are an appropriate feature for the purpose of speaker identification. To include temporal information the difference of the MFCCs of adjacent frames are computed ('Delta-features') and appended to the initial features.

In this contribution we investigate to which extent not only temporal but spectro-temporal information is of relevance for speaker identification. We do so by filtering a log-compressed Mel-Spectrum with Gabor-Filters [4]. The obtained Gabor-Features are appended to MFCCs and compared to the performance when appending Delta-MFCCs. We evaluate the different features by simulation on a database and show that spectro-temporal features are able to outperform purely temporal features. The remainder of this paper is organized as follows: In Section II we review the concept of text-independent speaker identification via GMMs with MFCCs as features. Gabor-Filters and parameter adaptation are explained in Section III. The combination of MFCCs with Gabor-features is also outlined. The performance of the proposed features is evaluated by simulation results in Section IV followed by conclusions in Section V.

II. Text-Independent Speaker Identification

A. Feature Extraction

The signal of a speech sequence $s(k)$ is sampled at f_s=8kHz. It is segmented into frames of length K=256 with an overlap between adjacent frames of 80 samples, with a frame index τ =1..T. A feature vector $\mathbf{f}_\tau = (f_{\tau,1}...f_{\tau,D})^{\mathrm{T}}$ of D dimensions is extracted from each frame, leading to a set of T feature vectors $\mathbf{F} = (\mathbf{f}_1, .., \mathbf{f}_T)^{\mathrm{T}}$.

B. Gaussian Mixture Models

For feature-vector $\mathbf{f}_\tau$ the probability density function given by a single gaussian mixture with index i is

$$b_i(\mathbf{f}_\tau) = \frac{1}{(2\pi)^{D/2} \cdot |\mathbf{C}_{ff}|^{1/2}} \cdot \exp\left[-\frac{1}{2}(\mathbf{f}_\tau - \boldsymbol{\mu}_i)^{\mathrm{T}} \mathbf{C}_{ff,i}^{-1} (\mathbf{f}_\tau - \boldsymbol{\mu}_i)\right] \quad (1)$$

The D-dimensional mean-vector is denoted by $\boldsymbol{\mu}_i$ and the $D \times D$ covariance matrix by $\mathbf{C}_{ff,i}$. To reduce computational complexity $\mathbf{C}_{ff,i}$ is restricted to contain diagonal elements only. Each mixture is weighted by a factor p_i satisfying $\sum^{O} p_i = 1$, with $i = 1..O$ and O as the total number of mixtures also called model order. The parameters of all O mixtures are summarized as the parameter set $\lambda_q = \{p_i, \boldsymbol{\mu}_i, \mathbf{C}_{ff,i}\}$ of speaker q. The probability of an observed feature vector given by the model λ is

$$p(\mathbf{f}_\tau|\lambda) = \sum_{i=1}^{O} p_i \cdot b_i(\mathbf{f}_\tau)\,. \quad (2)$$

Under the assumption of independence of the observations the overall probability of a set $\mathbf{F}$ given a model λ_q is

$$p(\mathbf{F}|\lambda_q) = \prod_{\tau=1}^{T} p(\mathbf{f}_\tau|\lambda_q)\,. \quad (3)$$

After extraction of the feature vectors from a training sequence $\mathbf{F}_{Train}$ it is the aim of model estimation to find those parameters λ_q which maximize the probability $p(\mathbf{F}_{Train}|\lambda_q)$. This is

accomplished by application of the Expectation-Maximization Algorithm (EM) [3], [2]. From an initial set of parameters λ a new set $\bar{\lambda}$ is estimated for which $\log p\left(\mathbf{F}|\bar{\lambda}\right) \geq \log p\left(\mathbf{F}|\lambda\right)$ is guaranteed. We terminate the algorithm if from iteration step w to step $w+1$ the criterion

$$\Theta = \left|\log p\left(\mathbf{F}|\bar{\lambda}^{w+1}\right) / \log p\left(\mathbf{F}|\bar{\lambda}^{w}\right) - 1\right| \tag{4}$$

falls below the threshold $\Theta < 1e-6$.

The O mean vectors $\boldsymbol{\mu}_i$ are initialized by the O centers found via vector-quantization [5]. The covariance matrices are initialized as identity matrices scaled by the maximum of all variance values $\sigma^2_{1..D}$ of the training data while the mixture weights are set to $p_i = 1/O$.

The closed set-classification of feature vectors extracted from a test sequence $\mathbf{F}_{Test} = (\mathbf{f}_1, .., \mathbf{f}_T)^{\mathrm{T}}$is carried out by choosing that model $\hat{q}$ as the speaker model which yields the highest probability

$$\hat{q} = \arg \max_{1 \leq q \leq Q} \sum_{\tau=1}^{T} \log p\left(\mathbf{f}_\tau^{Test}|\lambda_q\right) . \tag{5}$$

C. Mel-Frequency Cepstral Coefficients

Let us shortly review the extraction of MFCCs from a discrete-time signal [2]. After framewise segmentation and multiplication by a Hann-window

$$\hat{s}_\tau(k) = s_\tau(k) \cdot w_{hann}(k) \tag{6}$$

the signal is transformed into the frequency domain by the Discrete Fourier Transform (DFT)

$$S_\tau(r) = \mathrm{DFT}\{\hat{s}_\tau(k)\} = \sum_{k=0}^{K-1} \hat{s}_\tau(k) \cdot e^{-j2\pi r \frac{k}{K}} . \tag{7}$$

Triangular Mel-Filters as depicted in Fig. 1 are applied to the spectral values $S_\tau(r)$ leading to the log-compressed Mel-Spectrum

$$S_\tau^{Mel}(m) = 10 \log \left[\sum_{r_{low}(m)}^{r_{high}(m)} |S_\tau(r)| \cdot D_{Mel}(m, r) \right] \tag{8}$$

with $m = 1..M$ as the filter index and M =19 filters at f_s=8kHz. Denoted by $r_{low}(m)$ is the lower discrete frequency bin of the m-th Mel-Filter and by $r_{high}(m)$ the higher discrete frequency bin respectively. The weighting value of the m-th filter at frequency bin r is denoted by $D_{Mel}(m, r)$. The cepstral coefficients $C_\tau(n)$ called MFCCs are yielded by the discrete cosine transform (DCT) of the log-compressed Mel-Spectrum

$$C_\tau(n) = 2 \cdot \sum_{m=1}^{M} S_\tau^{Mel}(m) \cdot \cos\left(\frac{\pi}{M}\, n \left(M - \frac{1}{2}\right)\right) \tag{9}$$

where $n = 1..N$ is the index of the cepstral coefficients. Here, we restrict the number of coefficients to $N = 12$. A feature vector for frame τ is thus formed by the concatenation of the MFCCs as $\mathbf{C}_\tau = (C_\tau(1)..C_\tau(N))^{\mathrm{T}}$. The Delta-Vectors $\Delta\mathbf{C}_\tau = \mathbf{C}_\tau - \mathbf{C}_{\tau-1}$ as the difference between vectors of adjacent frames can be concatenated to the initial vectors in order to take temporal information into account.

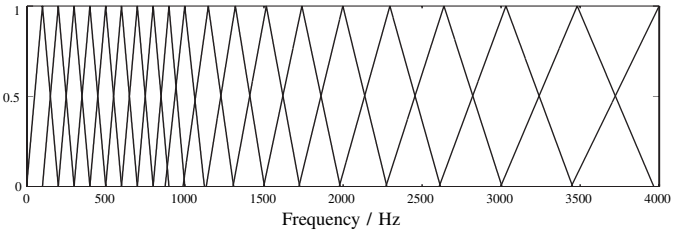

Fig. 1. Triangular Filters of the Mel-Filterbank

III. Gabor Features

Gabor-Filters as spectro-temporal filters have been applied by Kleinschmidt for feature generation in the field of speech recognition [4]. The time- and frequency- continuous two-dimensional Gabor function $g(f,t)$, with f as the frequency and t as the time, is the product of a complex Euler Function $e(f,t)$ and a Gaussian Function $n(f,t)$

$$\begin{aligned} g(f,t) &= e(f,t) \cdot n(f,t) \\ e(f,t) &= \exp\left(i\omega_f(f-f_0) + i\omega_t(t-t_0)\right) \\ n(f,t) &= \frac{1}{2\pi\sigma_f\sigma_t} \cdot \exp\left[\frac{-(f-f_0)^2}{2\sigma_f^2} + \frac{-(t-t_0)^2}{2\sigma_t^2}\right] \end{aligned} \tag{10}$$

with the parameters

$$\begin{aligned} t_0, f_0 &= \text{central time, central frequency} \\ \sigma_t, \sigma_f &= \text{standard deviations of time and frequency} \\ \omega_t, \omega_f &= \text{circular frequencies of the Euler Function} \end{aligned}$$

For a discrete-time and discrete-frequency implementation the Gabor-Filter can be expressed by the frame index τ and the Mel-Filter index m as $g(m,\tau)$. As an example the real part of a Gabor-filter is depicted in Fig. 2. The log-compressed Mel-Spectrum $S_\tau^{Mel}(m)$ in Fig. 3a) is convolved in time direction with the real part of the Gabor-Filter $\Re\{g(m,\tau)\}$ leading to the result in Fig. 3b). The feature values

$$G(\tau) = \sum_{m=1}^{M} \sum_{\acute{\tau}=-\tau_{max}}^{\acute{\tau}=\tau_{max}} \Re\{g(m,\tau)\} \cdot S_\tau^{Mel}(m, \tau+\acute{\tau}) \tag{11}$$

are obtained by summation over frequency and can be seen in Fig. 3c). The value of τ_{max} is here set to 50 frames, the time width of the Gabor-Filter. The final feature vector $\mathbf{G}_\tau$ is obtained by applying L different Gabor-Filters to the Mel-Spectrogram and concatenating the feature values from each filter as $\mathbf{G}_\tau = \left(G^1(\tau)...G^L(\tau)\right)^{\mathrm{T}}$.

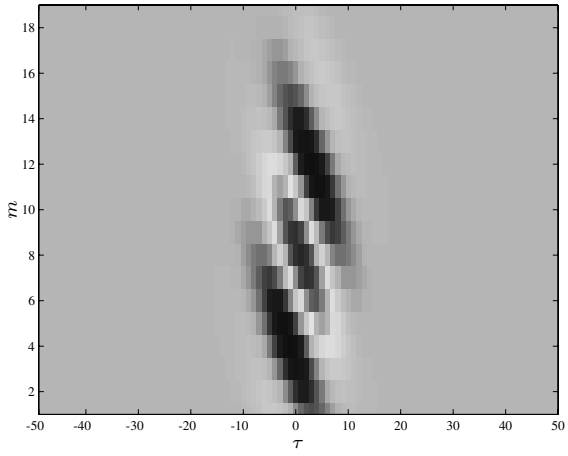

Fig. 2. Amplitude of the real part of a Gabor-Filter; dark values indicate high amplitude values

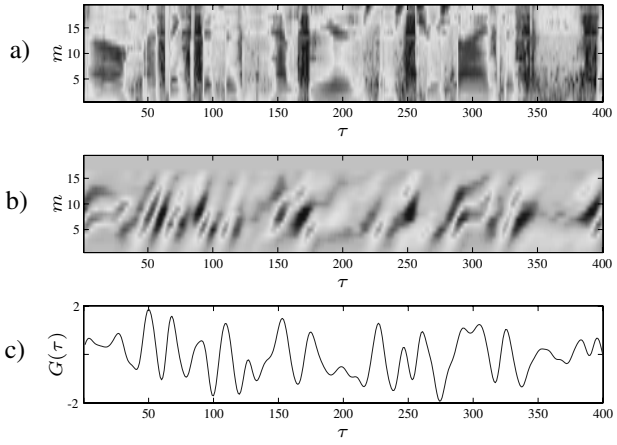

Fig. 3. a) Mel-Spectrogram b) Result from Gabor-Filter c) Feature values after summation over frequency

A. Optimization of Parameters

Finding Gabor-Filters with optimal parameters for the task of speaker identification is achieved by examining the F-ratio [6] as the measure for the separability of the different speaker classes. A set of L filters is initialized and for each speaker the features are extracted from a training sequence. The Gabor-Features $\mathbf{G}_\tau$ are concatenated with the MFCC features $\mathbf{C}_\tau$. For each class (or speaker) q, we compute the mean vector of the features $\boldsymbol{\mu}_q$ and the sample covariance matrix $\mathbf{W}_q$. The overall mean vector of all classes shall be $\boldsymbol{\mu}_0 = (1/Q)\sum^{Q}\boldsymbol{\mu}_q$. The F-ratio [6] can be determined by the within-class scatter matrix

$$\mathbf{S}_w = \frac{1}{Q}\sum_{q=1}^{Q}\mathbf{W}_q \tag{12}$$

and the between-class scatter matrix

$$\mathbf{S}_b = \frac{1}{Q}\sum_{q=1}^{Q}(\boldsymbol{\mu}_q - \boldsymbol{\mu}_0)(\boldsymbol{\mu}_q - \boldsymbol{\mu}_0)^{\mathrm{T}} \tag{13}$$

yielding the separability criterion

$$J = \mathrm{tr}\left({\mathbf{S}_w}^{-1}\mathbf{S}_b\right) . \tag{14}$$

In order to find a set of filters with a higher separability the following algorithm is carried out:

1) Initialize a set of Gabor-Filters at random.
2) Measure the relevance of each Gabor-Filter by computing the separability J_l of the classes when the features of the filter l are not taken into account.
3) The measure J_l with the highest value indicates the Gabor-Filter l which contributes least to the separability of the classes, thus filter l is discarded.
4) Draw a new filter at random to replace filter l.
5) Return to step 2.

The resulting values of J for three sets of different sizes L =12, 20 and 30 are plotted in Fig. 4. The algorithm was termintated after 1000 iterations. Although the separability obviously increases as we increase the number of filters we have to keep in mind that this only means that we achieve a higher separability of the features from the *training* sequences. It still remains as an issue whether this will result in a higher recognition rate of features from the *test* sequences. Having generated a set with L Gabor-Filters we extract the

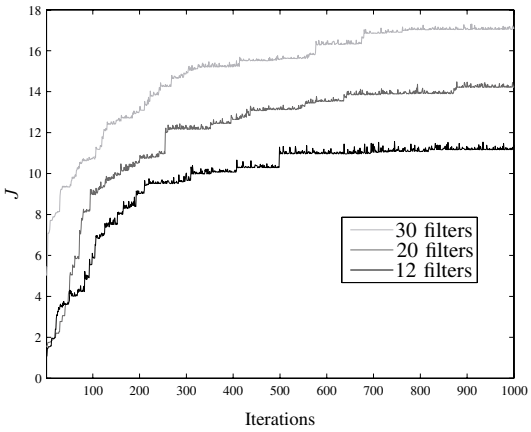

Fig. 4. Separability J for 12, 20 and 30 Gabor-Filters

features from the training sequence of each speaker obtaining the feature set $\mathbf{F}_q \epsilon \mathbb{R}^{T\times L}$. We then perform a singular value decomposition (SVD) of the set of training features

$$\mathbf{F}_q = \mathbf{U}\cdot\mathbf{S}\cdot\mathbf{V}^{\mathrm{T}} \tag{15}$$

and use the reduced base $\mathbf{V}_\mathbf{q}^\mathbf{red}\epsilon\,\mathbb{R}^{T\times 12}$ for projection of the features via

$$\hat{\mathbf{F}}_q = \mathbf{F}_q\cdot\mathbf{V}_\mathbf{q}^\mathbf{red} \tag{16}$$

to obtain 12 Gabor-features. The feature set of a test sequence has to be projected also onto the bases $\mathbf{V}_\mathbf{q}^\mathbf{red}$ before testing against speaker model λ_q.

IV. Results

The features discussed above were tested by simulations on the KING database [7]. The number of speakers was 26, training sequences had a length of 90 seconds and test sequence 10 seconds with a total number of 400 test sequences. Prior to feature extraction speech pauses were removed via voice activity detection [8]. At first, the performance of 12 MFCCs alone were compared to the combination of 12 MFCCs + 12 Delta-MFCCs as well as 12 MFCCs + 12 Gabor-Features. The results for a varying model order O are depicted in Fig. 5. Although the consideration of Delta-MFCCs increased the recognition rate compared to using MFCCs only, the combination with Gabor-Features achieved better results for any model order. The combination of Gabor-Features together with Delta-MFCCs resulted in slight improvement. The influence of the size of the Gabor set L =12, 20 and 30 was examined and the results for different feature combinations are plotted in Fig. 6. Although the separability J of the classes increased for greater sets, as it was shown in Section 2, the recognition rates decrease. A classification of the training sequences gave a recognition rate of 100%, which indicates overadaption on the training data. Obviously, an optimal set size remains to be determined. As a last experiment the speech sequences were disturbed by ICRA noise type 1 at a noise level with a segmental SNR [5] of 10dB. The results are depicted in Fig. 7. While Delta-MFCCs yielded no improvement of the recognition rate, a combination with Gabor-Features achieves this even in noisy conditions. Further work should focus on the optimization of Gabor-Features for different noise-conditions and investigate the effects of noise reduction techniques.

V. Conclusions

We investigated spectro-temporal features obtained from Gabor-Filters for the task of text-independent speaker identification. An algorithm for parameter optimization was presented and the features were tested via simulations on a database. In comparison to MFCCs and Delta-MFCCs as purely temporal features the spectro-temporal features yielded the highest recognition rates for clean speech and also when moderate noise was added.

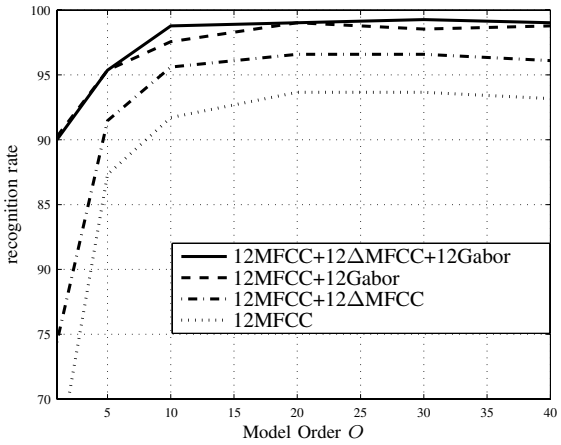

Fig. 5. Recognition rates for different feature types

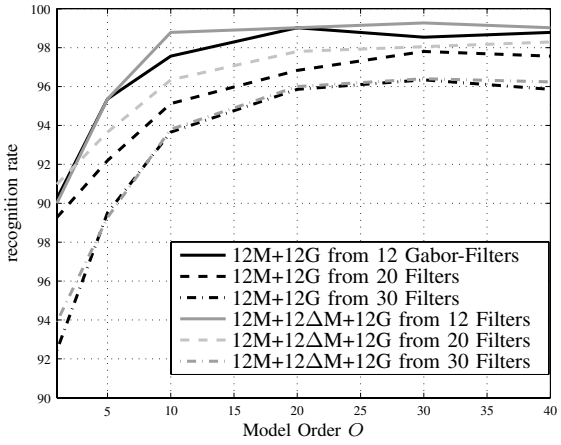

Fig. 6. Recognition rates for sets of different sizes

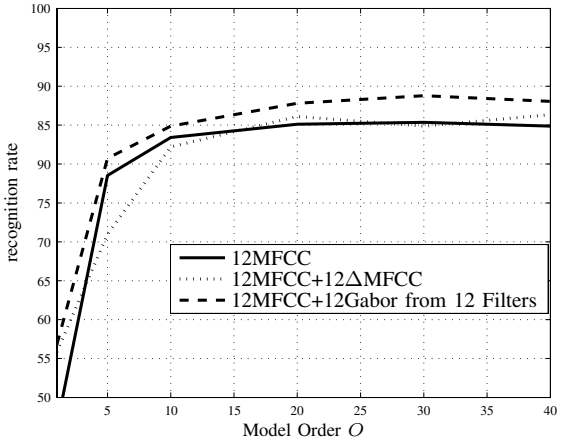

Fig. 7. Recognition rates at noise level of 10dB

References

[1] D. A. Reynolds, "An Overview of Automatic Speaker Recognition Technology," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Orlando, Florida, May 2002.

[2] D. A. Reynolds and R. C. Rose, "Robust Text-Independend Speaker Identification Using Gaussian Mixture Speaker Models," *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, Jan. 1995.

[3] A. Dempster, N. Laird, and D. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. Royal Stat. Soc.*, vol. 39, pp. 1–38, 1977.

[4] M. Kleinschmidt, *Robust Speech Recognition Based on Spectro-Temporal Processing*. PhD thesis at University of Oldenburg, 2002.

[5] P. Vary and R. Martin, *Digital Speech Transmission - Enhancement, Coding, and Error Concealment*, 1st ed. Wiley & Sons, 2006.

[6] S. Umesh and L. Cohen, "Scale Transform in Speech Analysis," in *IEEE Transactions on Speech and Audio Processing, vol.7*, January 1999.

[7] J. Godfrey and D. Graff, "Public Databases for Speaker Recognition and Verification," *ECSA Workshop Automat. Speaker Recognition*, vol. 10, pp. 39–42, March 1994.

[8] Q. Li, J. Zheng, A. Tsai, and Q. Zhou, "Robust Endpoint Detection and Energy Normalization for Real-Time Speech and Speaker Recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 3, pp. 146–157, March 2002.