TEXT-INDEPENDENT SPEAKER IDENTIFICATION USING TEMPORAL PATTERNS

Tobias Bocklet and Andreas Maier and Elmar Nöth

University of Erlangen Nuremberg, Chair for Pattern Recognition,
Martenstr.3, 91058 Erlangen, Germany
Andreas.Maier@informatik.uni-erlangen.de

Abstract. In this work we present an approach for text-independent speaker recognition. As features we used Mel Frequency Cepstrum Coefficients (MFCCs) and Temporal Patterns (TRAPs). For each speaker we trained Gaussian Mixture Models (GMMs) with different numbers of densities. The used database was a 36 speakers database with very noisy close-talking recordings. For the training a Universal Background Model (UBM) is built by the EM-Algorithm and all available training data. This UBM is then used to create speaker-dependent models for each speaker. This can be done in two ways: Taking the UBM as an initial model for EM-Training or Maximum-A-Posteriori (MAP) adaptation. For the 36 speaker database the use of TRAPs instead of MFCCs leads to a frame-wise recognition improvement of 12.0 %. The adaptation with MAP enhanced the recognition rate by another 14.2 %.

1 Introduction

The extraction of speaker-dependent information out of the voice of the user, so that a person can be identified or additional speaker specific information is obtained, is an important task these days. Speaker-dependent information is the identity of a speaker, the language, the age, the gender, or the channel he or she is calling from.
These pieces of information about the identity of the speaker or specific characteristics of the person are helpful for several applications. Identification of a person can be used to allow or restrict a person the use of certain services or the access to certain places. In these cases the user does not need to have a password, an account, a personal identification number (PIN), or a door-key anymore. The access is granted or denied only by the person's voice. It is also possible to perform the identification process in a secure way over the telephone.

In our approach each speaker is modeled by a *Gaussian Mixture Model* (GMM). To train the system first of all a Universal-Background-Model (UBM) is created comprising the complete amount of training data. This is achieved by the EM-Algorithm. The UBM is then used to create a speaker model in two ways: Either EM-Training is performed and the UBM is needed as an initial

© Springer-Verlag http://www.springer.de/comp/lncs/index.html
published in: Text, Speech and Dialogue (TSD 2007), Proceedings, LNAI 4629, Springer, Berlin, Heidelberg, 2007, pp. 318-325

speaker model or Maximum-A-Posteriori (MAP) adaptation is applied, where the UBM is combined with the speaker-dependent training data. We used two different features in this work: *Mel Frequency-Cepstrum-Coefficients* (MFCCs), which extract the features over a short temporal context (16 ms) and *TempoRAl Patterns* (TRAPs). TRAPs calculate the features over a very long temporal context (150 ms).

For training and evaluation we employed a database provided by the company MEDAV (`www.medav.com`). The database is called `SET-M`. The Verbmobil [1] database was used to generate transformation matrices for the dimension reduction of the TRAPs by *Linear Discriminant Analysis* (LDA). These two databases are presented in the following.

2 Databases

2.1 `SET-M`

The `SET-M`-corpus contains speech recordings of 36 persons, each of them reading two newspaper articles. The texts are semantically different. One text is a newspaper article dealing with computer viruses, the other article is about children who have attention deficit disorder (ADD). The data was recorded by a close-talking microphone. In order to simulate telephone quality, the data was μ-law coded. Additionally it was artificially corrupted by convolution with Gaussian noise. In total 84 min of speech was available, recorded with a sample rate of 22kHz and re-sampled to 16kHz. The computer virus text was used for training, the other one for testing. The total amount of the training set was 45 min and the length of the test set was 39 min respectively.

2.2 Verbmobil

The Verbmobil (VM) database (see [1]) is a widely used speech collection. We used a German subset of the whole corpus which was already investigated in [2]. The scenario of the corpus is human-human communication with the topic of business appointment scheduling. It contains in total 27.7 hours of continuous speech by 578 speakers of which 304 were male and 274 were female. The size of the vocabulary is 6825 words. On average each of the 12,030 utterances contains 22 words. The data of this corpus was transliterated and a forced alignment was performed. This produced phonetic labels for each speech frame. These labels are then utilized to train the transformation matrix of the Linear Discriminant Analysis which is used to reduce the dimension of our TRAPs from 556 to 24.

3 Applied Methods

3.1 Features

As features the commonly used *Mel Frequency Cepstrum Coefficients* (MFCCS) and *TempoRAl Patterns* (TRAPs) are employed. MFCCs calculate the features

on a short temporal context but they take the complete frequency domain into consideration. TRAPs examine each frequency band of the recordings separately over a very long temporal context.

Mel Frequency Cepstrum Coefficients The 24 dimensional MFCCs consist of 12 static and 12 dynamic components. The 12 static features are composed by the spectral energy and 11 cepstral features. Furthermore the 12 dynamic features are calculated as an approximation of the first derivative of the static features using a regression line over 5 time frames. The time frames are computed for a period of 16 ms with a shift of 10 ms.

Temporal Patterns The TRAPs we used in this work are quite similar to the original approach of Hermansky ([3]). The main difference of our approach are the time trajectories and their processing. Fig. 1 shows the complete extraction method. The time trajectories consider a long temporal context (150 ms) of 18 mel-bands. These mel-bands are generated by a convolution of the spectrum with triangular filter-banks. Each trajectory is smoothed by a Hamming window and transformed by application of the discrete Fast Fourier Transformation afterwards. These magnitudes in the frequency domain are then filtered by canceling all frequencies except the interval from 1 to 16Hz. A detailed explanation can be found in [4]. The fusion of the trajectories combined with a dimension reduction is not performed by neural networks, as in the original paper, but by concatenation of the high-dimensional features and application of either *Linear Discriminant Analysis* (LDA) or *Principal Component Analysis* (PCA) afterwards. The result of this dimension reduction were 24-dimensional features, as in case of MFCCs.

To train the transformation matrices of the LDA transform, labeled data was needed. We decided to use the Verbmobil database, because the data of this corpus was already transliterated and forced aligned. This produced labels in form of 47 German phonetic classes.

3.2 Classifier specifications and test phase

In this work the speakers are modeled by *Gaussian Mixture Models* (GMMs) as described in [5]. Each speaker λ is modeled by M unimodal weighted Gaussian Distributions:

$$p(\boldsymbol{x} \mid \lambda) = \sum_{i=1}^{M} w_i p_i(\boldsymbol{x}). \tag{1}$$

with

$$p_i(\boldsymbol{x}) = \frac{1}{(2\pi)^{D/2} \mid \boldsymbol{K_i} \mid^{1/2}} e^{-(1/2)(\boldsymbol{x}-\mu_i)^T \boldsymbol{K_i}^{-1}(\boldsymbol{x}-\mu_i)} \tag{2}$$

where μ_i denotes the mean vector and $\boldsymbol{K_i}$ the covariance matrix of the Gaussians. Unlike [5] we used full covariance matrices in our work, because preliminary comparisons showed a slight advantage of fully occupied matrices. The number

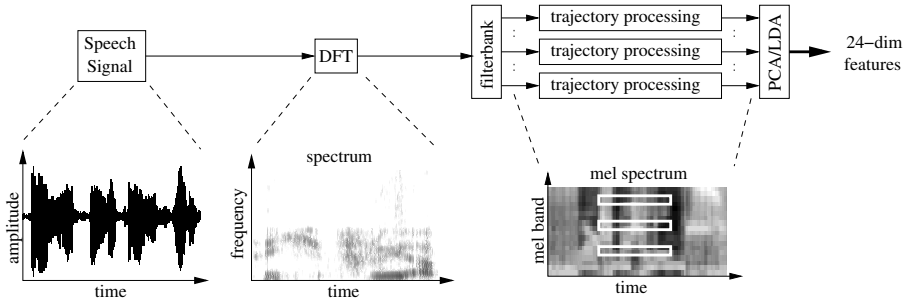

Fig. 1. Feature extraction for Temporal Patterns

of densities is varied from 16 to 2048 in 2^x steps. For classification a standard Gaussian Classifier is used. The classifier calculates for each feature vector of a specific speaker an allocation probability for each speaker model. This is done for all speech frames of one utterance. Then the probabilities of each model are accumulated. The model which achieved the highest value is expected to be the correct one.

3.3 Training

In Fig. 2 the general procedure of the training phase is shown. After feature extraction a *Universal Background Model* (UBM) is generated. Therefore, we comprised all the available training data. Then either a standard EM-Training or MAP adaptation [6, 7] was applied to derive speaker-dependent models.

The EM-algorithm consist of the E-step (Eq. 3) where the A Posteriori probabilities of a feature vector $\boldsymbol{x_t}$ for every mixture i is calculated.

$$p(i \mid \boldsymbol{x_t}) = \frac{\omega_i p_i(\boldsymbol{x_t})}{\sum_{j=1}^{M} \omega_j p_j(\boldsymbol{x_t})}. \tag{3}$$

$p(i \mid \boldsymbol{x_t})$ is then used in the M-Step to reestimate the components of the new speaker model λ':

$$\text{Mixture weights: } w_i' = \frac{1}{T}\sum_{t=1}^{T} p(i \mid \boldsymbol{x_t}) \tag{4}$$

$$\text{Mean values: } \boldsymbol{\mu_i'} = \frac{\sum_{t=1}^{T} p(i \mid \boldsymbol{x_t})\boldsymbol{x_t}}{\sum_{t=1}^{T} p(i \mid \boldsymbol{x_t})} \tag{5}$$

$$\text{Covariance matrices: } \boldsymbol{K_i'} = \frac{\sum_{t=1}^{T} p(i \mid \boldsymbol{x_t})}{\sum_{t=1}^{T} p(i \mid \boldsymbol{x_t})}(\boldsymbol{x_t} - \boldsymbol{\mu_i'})(\boldsymbol{x_t} - \boldsymbol{\mu_i'})^{\boldsymbol{T}} \tag{6}$$

$(\boldsymbol{x_t} - \boldsymbol{\mu_i'})^{\boldsymbol{T}}$ in Eq. 6 describes the transposed mean subtracted feature vector. After the M-step the model λ is replaced by the new estimated model λ'.

The MAP-adaptation also uses (Eq. 3) to estimate $p(i \mid \boldsymbol{x_t})$ out of the UBM parameters and the speaker-dependent feature vectors $\boldsymbol{x_t}$. The weight ($\tilde{\omega_i}$), mean ($\tilde{\boldsymbol{\mu_i}}$) and variance ($\tilde{\boldsymbol{K}}_{\boldsymbol{i}}$) parameters of each mixture i are computed by:

$$\tilde{\omega_i} = \sum_{t=1}^{T} p(i \mid \boldsymbol{x_t}) \tag{7}$$

$$\tilde{\boldsymbol{\mu_i}} = \sum_{t=1}^{T} p(i \mid \boldsymbol{x_t}) \boldsymbol{x_t} \tag{8}$$

$$\tilde{\boldsymbol{K}}_{\boldsymbol{i}} = \sum_{t=1}^{T} p(i \mid \boldsymbol{x_t}) \boldsymbol{x_t} \boldsymbol{x_t}^T \tag{9}$$

Finally these newly calculated statistics are combined with the UBM statistics to create the parameters for the adapted density i: $\hat{\omega_i}, \hat{\mu_i}, \hat{K_i}$ (see [6, 7]):

$$\hat{\omega_i} = [\alpha_i \tilde{\omega_i}/T + (1 - \alpha_i)\omega_i]\gamma \tag{10}$$

$$\hat{\boldsymbol{\mu_i}} = \alpha_i \tilde{\boldsymbol{\mu_i}} + (1 - \alpha_i)\boldsymbol{\mu_i} \tag{11}$$

$$\hat{\boldsymbol{K}}_{\boldsymbol{i}} = \alpha_i \tilde{\boldsymbol{K}}_{\boldsymbol{i}} + (1 - \alpha_i)(\boldsymbol{K_i} + \boldsymbol{\mu_i}\boldsymbol{\mu_i^T}) - \boldsymbol{\mu_i}\boldsymbol{\mu_i^T} \tag{12}$$

The adaptation coefficient α_i is defined as:

$$\alpha_i = \frac{n_i}{n_i + \tau}, \tag{13}$$

where τ has to be selected by the user. In preliminary experiments we distinguished the best value to be 50 for our database. (Eq. 10) contains the scale factor γ, which normalizes the sum of all new estimated a priori probabilities $\hat{\omega_i}, i \in 1, ..M$ to 1.

Both algorithms take the UBM as an initial model and for each single speaker one speaker-distinguishing model is created. The difference between EM-Training and MAP adaptation is, that MAP adaptation calculates the parameters of the speaker-dependent Gaussian mixtures in only one iteration step and combines them with the UBM-parameters.

4 Experiments and Results

In preliminary experiments we investigated the best TRAPs parameters. For the preliminary experiments we used speaker models with 64 Gaussian densities and standard EM-Training. The parameters we varied were the context (15 or 30), the use of filtered and normal TRAPs and the application of PCA or LDA alternatively. For the `SET-M` database we used a context of 15 and filtered TRAPs. The feature reduction was performed by LDA, because it outperformed the PCA approach.

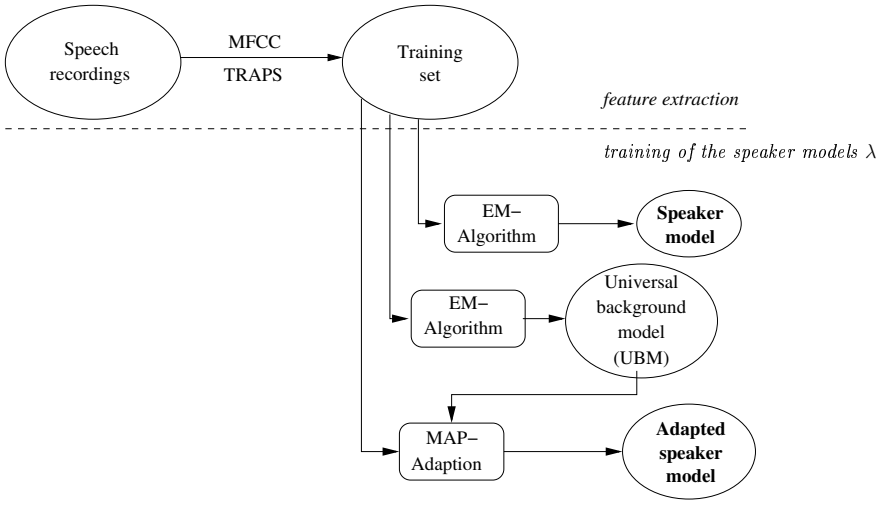

Fig. 2. General proceeding of the training phase

Table 1 shows the recognition results for the EM-Training and the MAP adaptation. It contains the results for both features: MFCC and TRAPs. *fw* denotes the recognition result, reached when deciding for each frame separately (frame-wise) and *sp* denotes the recognition results of the classification of all vectors of one speaker (speaker-level). The columns named *100f* and *500f* contain the recognition results after classification with 100 and 500 frames each (no overlap).

In the case of EM-Training we observed a maximal frame-wise recognition rate of 24.16 % with TRAPs features and 32 Gaussian mixtures. The maximal recognition rate for the *speaker* decision was 91.67 %. For the `SET-M` corpus the MAP adaptation outperforms the EM-Training. In the case of MAP adaptation the highest frame-wise recognition result (27.58 %) was achieved by 32-dimensional speaker models and TRAPs. The maximal value in case of *speaker* decision (100 %) was accomplished with 512-dimensional models and MFCCs.

In Fig. 3 we plotted the recognition results dependent on the amount of test feature vectors. Therefore, we classified all data of the test speakers after a given number of frames (no overlap). One can see, that the slope of the recognition results is almost zero when more than 500 frames are used.

5 Discussion

Using TRAPs in case of text-independent speaker recognition can improve the recognition results, especially if the recordings are very noisy, like the database of this paper. So the recognition could be improved from 21.57 % to 24.84 %

	EM-Training								MAP							
	MFCC				TRAPs				MFCC				TRAPs			
Density	fw	100f	500f	sp	fw	100f	500f	sp	fw	100f	500f	sp	fw	100f	500f	sp
32	21.6	76.4	79.7	92	24.2	70.2	86.7	92	24.8	80.7	91.7	92	27.6	76.3	90.6	97
64	19.2	67.5	79.7	83	23.2	68.8	85.1	92	24.4	81.4	92.6	92	26.9	77.0	90.2	97
128	16.7	65.6	77.0	86	22.4	66.5	84.0	92	23.4	81.2	92.1	97	24.9	73.9	86.7	92
256	14.5	61.7	74.6	83	21.5	60.5	83.4	92	22.3	80.4	91.7	94	22.4	73.3	85.8	92
512	12.5	36.4	39.6	33	21.2	52.6	63.2	67	19.9	78.5	91.9	100	20.2	72.6	86.7	92
1024	12.0	48.2	59.5	64	16.1	34.0	36.7	36	16.8	73.7	90.4	97	16.9	67.4	86.9	94
2048	9.5	20.7	19.5	14	15.9	31.0	34.4	31	12.4	64.5	82.3	97	15.1	64.4	88.4	94

Table 1. frame-wise (fw) and speaker-level (sp) recognition results achieved on the `SET-M` corpus

(12.0 %) in case of frame-wise recognition. Due to the fact, that the amount of training data was very low in this database, the use of more Gaussian mixtures even decreased the frame-wise recognition result. The *speaker* recognition reaches its maximum (91.67 %) at speaker models with 32 densities.

The training with MAP-adaptation improves the frame-wise recognition rate by another 14.2 % from 24.84 % to 27.58 %. But an increase of the number of Gaussian densities does not achieve an improvement of the recognition rate of the TRAPs. On speaker-level MFCCs obtain better results than TRAPs, if a larger number of densities is chosen.

Therefore, we conclude that TRAPs have a better recognition rate on frame-level due to the larger context and that the properties of a speaker can be modeled with TRAPs using fewer Gaussian densities than MFCCs. We will examine this aspect further in future experiments.

6 Summary

In this paper we evaluated a system for speaker-independent speaker recognition. We used 2 different kinds of features: MFCCs and TRAPs. Both analyze the spectrum of a given recording. MFCCs examine the complete frequency domain on a short temporal context and TRAPs calculate features by analyzing different frequency bands over a longer time period. For the training we created a UBM by standard EM-Training on all the available training data. To build one model for every speaker we took this UBM as an initial model and applied EM-Training or MAP adaptation respectively. For this step we only used the speaker-dependent training data. For the evaluation of our system we performed experiments on the `SET-M` database.

We improved the frame-wise recognition result by 12.0 % when using TRAPs instead of MFCCS. The application of MAP adaptation improved the frame-wise recognition results by additional 14.2 %. The *speaker* recognition result also was increased and for 512 Gaussian densities 100 % were reached.

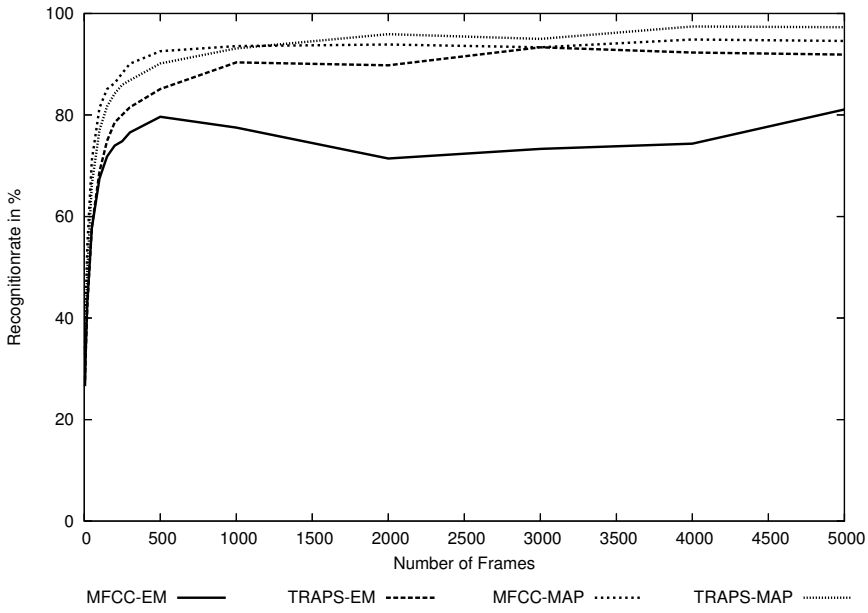

Fig. 3. Speaker recognition results dependent on the amount of test feature vectors

References

1. W. Wahlster, *Verbmobil: Foundations of Speech-to-Speech Translation*, Springer, New York, Berlin, 2000.
2. G. Stemmer, *Modeling Variability in Speech Recognition*, Ph.D. thesis, Chair for Pattern Recognition, University of Erlangen-Nuremberg, Germany, 2005.
3. H. Hermansky and S. Sharma, "TRAPS – classifiers of temporal patterns," in *Proc. International Conference on Spoken Language Processing (ICSLP)*, Sydney, Australia, 1998.
4. A. Maier, C. Hacker, S. Steidl, E. Nöth, and H. Niemann, "Robust Parallel Speech Recognition in Multiple Energy Bands," in *Pattern Recognition, 27th DAGM Symposium, August 30 - September 2005, Vienna, Austria, Proceedings*, Berlin, Heidelberg, 2005, Lecture Notes in Computer Science, pp. 133–140, Springer.
5. Douglas A. Reynolds and Richard C. Rose, "Robust Test-Independent Speaker Identification Using Gaussian Mixture Speaker Models," *IEEE Transaction on Speech and Audio Processing*, vol. 3, pp. 72–83, 1995.
6. J.L. Gauvain and C.H. Lee, "Maximum A-Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 291–298, 1994.
7. Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, pp. 19–41, 2000.