Analysis of feature extraction and channel compensation in a GMM speaker recognition system

Lukáš Burget*, *Member, IEEE,* Pavel Matějka, *Member, IEEE,* Petr Schwarz, *Member, IEEE,* Ondřej Glembek, *Student Member, IEEE,* and Jan "Honza" Černocký, *Member, IEEE*

***Abstract*—In this paper, several feature extraction and channel compensation techniques found in state-of-the-art speaker verification systems are analyzed and discussed. For the NIST SRE 2006 submission, Cepstral Mean Subtraction, Feature Warping, RASTA filtering, HLDA, Feature Mapping and Eigenchannel Adaptation were incrementally added to minimize the system's error rate. The paper deals with Eigenchannel Adaptation in more detail, and includes its theoretical background and implementation issues. The key part of the paper is however the post-evaluation analysis, undermining a common myth that "the more boxes in the scheme, the better the system". All results are presented on NIST SRE 2005 and 2006 data.**

***Index Terms*—Speaker recognition, GMM, Feature Warping, RASTA, HLDA, Feature Mapping, Eigenchannel Adaptation.**

EDICS Category: SPE-SPKR

I. Introduction

In the NIST 2006 Speaker Recognition Evaluation [1], the Brno University of Technology (BUT) participated with its own submission and also contributed to systems developed by the STBU[1] consortium. Both the BUT and STBU primary systems were fusions of several individual subsystems, namely: systems based on Gaussian Mixture Modeling (GMM) [2], and systems based on sequence kernel Support Vector Machines (SVM) classifying either GMM mean supervectors [3] or vectors constructed from Maximum Likelihood Linear Regression (MLLR) transformations [4], which are transformations commonly used in speech recognition for speaker adaptation. In this paper, we provide an analysis of the BUT GMM system that took part in both the BUT and STBU primary systems, and which was also submitted as a BUT stand-alone secondary system. The overall description of the BUT and STBU systems can be found in [5], [6].

The BUT GMM system is based on a standard Universal Background Model-Gaussian Mixture Modeling (UBM-GMM) paradigm [2] and employs a number of techniques that have previously proven to improve GMM modeling capability and help fight against the main problem in speaker verification - diversity in channel and acoustic conditions. These techniques are: Cepstral Mean Subtraction, Feature Warping [7], RelAtive SpecTrAl (RASTA) filtering [8], Heteroscedastic Linear Discriminant Analysis (HLDA) [9], Feature Mapping [10] and Eigenchannel Adaptation [11]. The aim of this paper is to analyze the importance of the individual techniques in terms of their contribution to overall system performance.

The paper is organized as follows: A detailed description of the BUT GMM speaker recognition system is provided in section II. Section III documents building the system and reports the improvements in performance obtained by adding individual techniques. Section IV presents our post-evaluation activity and analyzes the importance of the individual techniques in the full system. The result obtained by fusing the GMM system with the SVM-based systems are presented in section V. We conclude the paper in section VI.

II. System Description

A. Features

The features used in the system are Mel-frequency cepstral coefficients (13 MFCC coefficients including C0, 20 ms window, 10 ms shift, 23 bands in a Mel filter bank). To compensate for channel mismatch in different conversations, three simple feature processing techniques were successively applied: the cepstral mean over the whole conversation is subtracted from the features, Feature Warping [7] (3 sec window, warping into a normal distribution) is applied and finally temporal trajectories of individual feature vector coefficients are filtered using a standard RASTA filter [8][2]. After this processing, each feature vector is augmented with its first, second and third order derivatives. This results in 52 dimensional feature vectors containing information about the context of 13 frames.

B. Segmentation

At this stage, non-speech frames are discarded and only speech frames are considered in the following stages of training models and verification. Speech/non-speech segmentation is performed by our Hungarian phoneme recognizer [12],

All authors are with Speech@FIT, Faculty of Information Technology, Brno University of Technology, Czech Republic. Lukáš Burget is the corresponding author.

This work was partly supported by by European projects AMIDA (IST-033812) and Caretaker (FP6-027231), by Grant Agency of the Czech Republic under projects No. 102/05/0278 and GP102/06/383 and by the Czech Ministry of Education under project No. MSM0021630528. The hardware used in this work was partially provided by CESNET under projects No. 119/2004, No. 162/2005 and No. 201/2006.

We are grateful to the anonymous reviewers for their valuable comments on the first version of this paper. Many thanks to Ken Froehling from the Faculty of Electrical Engineering and Communication of BUT for help with the English of the paper.

[1]BUT, TNO Human Factors (The Netherlands), Spescom DataVoice (South Africa) and the University of Stellenbosch (South Africa).

[2]Cepstral Mean Subtraction has no effect after the application of Feature Warping and RASTA filtering as both techniques also ensure the mean removal. However, it will be interesting to see the effectiveness of these techniques compared to Cepstral Mean Subtraction alone.

where all phoneme classes are linked to speech classes. A postprocessing with two rules based on the short time energy of the signal is applied: 1) If the average energy in a speech segment is 30dB less than the maximum energy in the conversation side, then the segment is labeled as silence. 2) If the energy in the opposite conversation side[3] is bigger than the maximum energy minus 3dB in the processed side, the segment is also labeled as silence.

C. HLDA

As the next step, we have employed Heteroscedastic Linear Discriminant Analysis (HLDA), which is also in common use in speech recognition systems. HLDA provides a linear transformation that can de-correlate the features and reduce the dimensionality while preserving the discriminative power of features. The theory of HLDA is described in detail in [9], [13]. HLDA needs classes to estimate its class-covariance statistics (which are then used to estimate the transformation matrix). For this purpose, GMM with 2048 Gaussian components is trained on test data from SRE2004 and the feature frames aligned with individual GMM mixture components are considered as classes. HLDA transformation reducing the dimensionality from 52 to 39 is estimated. GMM is then updated in the new HLDA space (by projecting collected class-covariance and mean statistics through HLDA transformation). Features are also projected into HLDA space and GMM is re-estimated (still only on SRE2004 test data) by few additional Expectation-Maximization (EM) iterations to obtain the Universal Background Model (UBM).

D. Feature Mapping

To further compensate for channel mismatch, Feature Mapping [10] was applied to all enrollment and test conversations. Feature Mapping requires a set of models, each adapted from UBM using data of particular acoustic condition (channel). We have used 14 such models: 6 models were adapted for 3 channels (cell,cord,stnd) and 2 genders given the labels from 2004 test data. The remaining 8 models were initially adapted for 4 channels (cdma, cord, elec, gsmc) and 2 genders using the TNO Feature Mapping labels used in SRE-2005. However, these 8 models were then iteratively used to re-cluster the training data in an unsupervised fashion and again adapted using the new clustering (20 iterations lead to stable clustering) [14].

E. Training speaker model and verification

Each speaker model is obtained by a traditional *relevance Maximum A-Posteriori (MAP)* adaptation [15] of UBM using enrollment conversation. Only means are adapted with a relevance factor $\tau = 19$.

In the verification phase, standard Top-N Expected Log Likelihood Ratio (ELLR) scoring [15] is used to obtain a verification score, where $N = 10$ in our system. However, for each trial, both the speaker model and UBM are adapted to the channel of test conversation using simple Eigenchannel Adaptation [11] prior to computing the log likelihood ratio score. Note, that when T-norm [16] is used to normalize the score, each T-norm model is also adapted to the channel of relevant tested conversation.

F. Eigenchannel subspace estimation

We adopted the term 'eigenchannel' as used in speaker recognition from Kenny [17]. It was introduced to the NIST SRE by SDV in 2004 [11], revisited by Kenny and Vogt [18] in SRE 2005, and again by several sites in various forms in SRE 2006.

Let *supervector* be a MD dimensional vector constructed by concatenating all GMM mean vectors and *normalized by corresponding standard deviations*. M is the number if Gaussian mixture components in GMM and D is dimensionality of features. Before Eigenchannel Adaptation can be applied, we must identify directions in which the *supervector* is mostly affected by a changing channel. These directions, which we will refer to as eigenchannels, are defined by columns of $MD \times R$ matrix $\mathbf{V}$, where R is the chosen number of eigenchannels ($R = 30$ in our system). The matrix $\mathbf{V}$ is given by R eigenvectors of average within class covariance matrix, where each class is represented by supervectors estimated on different segments spoken by the same speaker.

More precisely, we have selected all (310) speakers from NIST SRE2004 data for which at least two conversations are available. For each speaker, i, and all his conversations, $j = 1, \ldots, J_i$, UBM is adapted to obtain a supervector, $\mathbf{s}_{ij}$. The corresponding speaker average supervector given by $\overline{\mathbf{s}}_i = \sum_{j=1}^{J_i} \mathbf{s}_{ij}/J_i$ is subtracted from each supervector, $\mathbf{s}_{ij}$, and resulting vectors form columns of $MD \times J$ matrix $\mathbf{S}$, where J is the number of all conversations from all selected speakers ($J = 2961$ in our case). Eigenchannels (columns of matrix $\mathbf{V}$) are given by R eigenvectors of $MD \times MD$ average within speaker covariance matrix[4] $\frac{1}{J}\mathbf{S}\mathbf{S}^T$ corresponding to R largest eigenvalues. Unfortunately, for our system, where $MD = 2048 \times 39 = 79872$, direct computation of these eigenvectors is unfeasible. A possible solution is to compute eigenvectors, $\mathbf{V}'$, of $J \times J$ matrix $\frac{1}{J}\mathbf{S}^T\mathbf{S}$; eigenchannels are then given by $\mathbf{V} = \mathbf{S}\mathbf{V}'$. In case the maximum a-posteriori (MAP) criterion is used for Eigenchannel adaptation (see below), the length of each eigenchannel must be also normalized to the average within speaker standard deviation of supervectors along the direction of the eigenchannel (i.e. each eigenvector obtained in the previous step must be multiplied by the square root of the corresponding eigenvalue). This normalization is irrelevant in the case of maximum likelihood (ML) criterion.

G. Eigenchannel Adaptation

Once the eigenchannels are identified, a speaker model (or UBM) can be adapted to the channel of a test conversation by shifting its supervector in the directions given by eigenchannels to better fit the test conversation data. Mathematically,

[3]In NIST SRE2006 evaluations, our system participated only in the primary condition, where two separate recordings for the two sides of each phone conversation are available.

[4]Note that matrix $\frac{1}{J}\mathbf{S}\mathbf{S}^T$ is a true covariance matrix as the zero mean over columns of S is guaranteed by the subtraction of the speaker average supervectors described above.

this can be expressed as finding the *channel factors*, $\mathbf{x}$, that maximize the following MAP criterion:

$$p(\mathbf{O}|\mathbf{s}+\mathbf{V}\mathbf{x})\mathcal{N}(\mathbf{x};\mathbf{0},\mathbf{I}), \quad (1)$$

where $\mathbf{s}$ is a supervector representing the model to be adapted[5], $p(\mathbf{O}|\mathbf{s}+\mathbf{V}\mathbf{x})$ is the likelihood of the test conversation given the adapted supervector (model) and $\mathcal{N}(\cdot;\mathbf{0},\mathbf{I})$ denotes a normally distributed vector. Assuming a fixed occupation of the Gaussian mixture components by test conversation frames, $\mathbf{o}_t, t = 1,\ldots,T$, it can be shown [11] that $\mathbf{x}$ maximizing criterion (1) is given by:

$$\mathbf{x} = \mathbf{A}^{-1}\sum_{m=1}^{M}\mathbf{V}_m^T\sum_{t=1}^{T}\gamma_m(t)\frac{\mathbf{o}_t-\boldsymbol{\mu}_m}{\boldsymbol{\sigma}_m}, \quad (2)$$

where $\mathbf{V}_m$ is $M\times R$ part of matrix $\mathbf{V}$ corresponding to the m^{th} mixture component, $\gamma_m(t)$ is the probability of occupation mixture component m at time t, $\boldsymbol{\mu}_m$ and $\boldsymbol{\sigma}_m$ are the mixture component's mean and standard deviation vectors and

$$\mathbf{A} = \mathbf{I} + \sum_{m=1}^{M}\mathbf{V}_m^T\mathbf{V}_m\sum_{t=1}^{T}\gamma_i(t). \quad (3)$$

In our implementation, occupation probabilities, $\gamma_m(t)$, are computed using UBM and assumed to be fixed for given test conversation. This allows us to pre-compute matrix $\mathbf{A}^{-1}$ only once for each test conversation. For each frame, only Top-N occupation probabilities are assumed not to be zero. In the following ELLR scoring, only the same top-N mixture components are also considered. All these facts ensure that adapting and scoring different speaker or T-norm models on a test conversation can be performed very efficiently.

Eigenchannel Adaptation can be also performed by maximizing ML criterion instead of MAP criterion. This corresponds to dropping the prior term, $\mathcal{N}(\mathbf{x};\mathbf{0},\mathbf{I})$, in criterion (1) and term $\mathbf{I}$ in equation 3. In our experiments, there is always enough adaptation data (test conversations contain approximately 2.5 minutes of speech) making the prior term in MAP criterion negligible. Therefore, we have not found any differences in performance when using the two criteria.

Our system uses a very simple scheme of modeling channel variability that affects only the verification phase. However, more sophisticated schemes can be considered. In [19] the verification phase is equivalent to that described here, however, modeling channel variability is considered also in training speaker models. This may become important especially when speaker models are trained using more than one enrollment conversation.

A very elaborate scheme can be found in [17], where modeling channel variability is considered in all phases: training background model, training speaker models and verification. Instead of finding eigenvectors, channel subspace $\mathbf{V}$ is obtained also by maximizing MAP criterion similar to (1). For enrollment data, instead of finding MAP point estimates of model parameters, posterior probabilities of model parameters are considered and integrated over to obtain the likelihood score for a test conversation.

[5]Note again that by our definition, a supervector is a mean supervector normalized by the corresponding standard deviations.

III. Building the system

In the following experiments, results will be presented for "1-side training, 1-side test, all trials" condition from SRE2005 NIST evaluation, which we have used for system development, and for primary condition (1-side training, 1-side test, English only trials) from SRE2006 NIST evaluation. In the tables, results are presented in terms of EER (Equal Error Rate) and $C_{\text{Det}}^{\min}$ as defined by SRE2006 NIST evaluation rules [1]. For SRE2006 primary condition, performances are also presented in the form of DET (Detection Error Tradeoff) curves.

Table I and figure 1 document the process of building our system. It shows line-by-line the improvements in performance obtained by successively adding different techniques. Our starting point was GMM system with 2048 Gaussian mixture components, features were 13 MFCC coefficients augmented with their deltas and processed by cepstral mean subtraction. The error rate of this system is very high and is almost halved by simply adding RASTA filtering. Replacing RASTA with Feature Warping improved the performance; however, a further small gain was obtained from the combination of both techniques. The application of RASTA filtering on top of Feature Warping appeared to be slightly more advantageous than doing it in the opposite order. In the next two steps, features were also augmented with double-delta and triple-delta coefficients. While adding double-deltas is clearly beneficial for both SRE2005 and SRE2006 evaluation sets, the advantage of adding triple-deltas, which we have seen during development on SRE2005 data, was not confirmed on SRE2006.

The following three steps, each significantly improving the system performance, were: projection of 52 dimensional features into 39 dimensional HLDA space, application of our 14 classes Feature Mapping and Eigenchannel Adaptation.

So far, all the presented results were obtained without normalizing the verification scores by any standard technique, such as T-normalization or Z-normalization (Z-norm/T-norm) [16]. As can be seen in Table I, T-norm was not effective in improving the performance of our full system. We have also experimented with Z-norm and ZT-norm, nevertheless, results obtained with all normalization techniques were mixed and unconvincing. This contradicted the conclusions drawn in [17], [18], [19], where Z-norm or ZT-norm was found necessary for making channel variability modeling techniques really effective.

Most of GMM based speaker verification systems, for which the results are published by various sites, use less than 2048 Gaussian components. The last line of table I show results for a system with the usual number of only 512 Gaussian components, which is otherwise identical to our full system. It can be seen that the performance of a 2048 component system is superior to this smaller one.

IV. Post-evaluation analysis

In the previous section, we have shown how adding individual techniques improves system performance. However, it will be even more interesting to see whether and how the individual techniques are important in the full system.

System	SRE2005		SRE2006	
	EER	$C_{\text{Det}}^{\min}$	EER	$C_{\text{Det}}^{\min}$
MFCC+Δ, CMS, 2048 G.	26.6%	.089	23.8%	.088
+ RASTA	14.3%	.055	11.8%	.059
+ Feature Warping	12.4%	.052	10.0%	.051
+ $\Delta\Delta$	11.2%	.047	9.1%	.049
+ $\Delta\Delta\Delta$	10.6%	.047	9.3%	.048
+ HLDA (52→39)	9.7%	.042	8.2%	.041
+ Feature Mapping	7.3%	.033	6.2%	.032
+ Eigenchannel Adapt.	4.6%	.020	4.0%	.020
+ T-norm	4.6%	.020	4.0%	.018
Full system, 512 Gauss.	4.9%	.026	4.7%	.024

TABLE I
THE IMPROVEMENTS IN PERFORMANCE OBTAINED BY SUCCESSIVELY ADDING DIFFERENT TECHNIQUES.

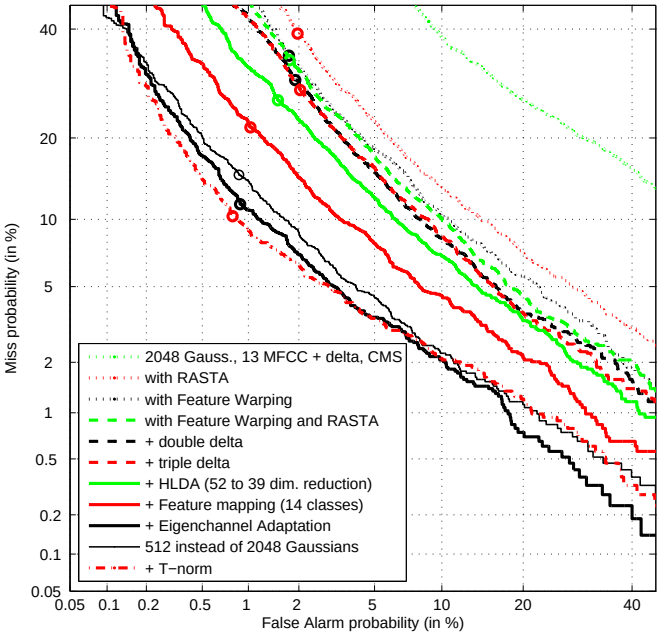

Fig. 1. DET curves showing improvement in successive adding different techniques.

A. The importance of RASTA and Feature Warping

Table II and figure 2 present results obtained with the baseline full system[6] and two of its modifications leaving out either RASTA filtering or Feature Warping. While Feature Warping turns out to be an important part of the system, leaving out RASTA filtering even slightly improves the system performance. This may support the conclusions in [8], where RASTA was found to discard important speaker information lying under its cut-off frequency and a filter more appropriate for speaker verification was designed.

B. Analyzing the effect of HLDA

The left half of Table III shows the effect of HLDA for systems without the following Feature Mapping and Eigenchannel Adaptation. The first two results (already presented in Table I) demonstrate the effectiveness of HLDA at this

[6]System with 2 Gender Feature Mapping (see below) is used as a baseline system in this experiment for efficiency reasons.

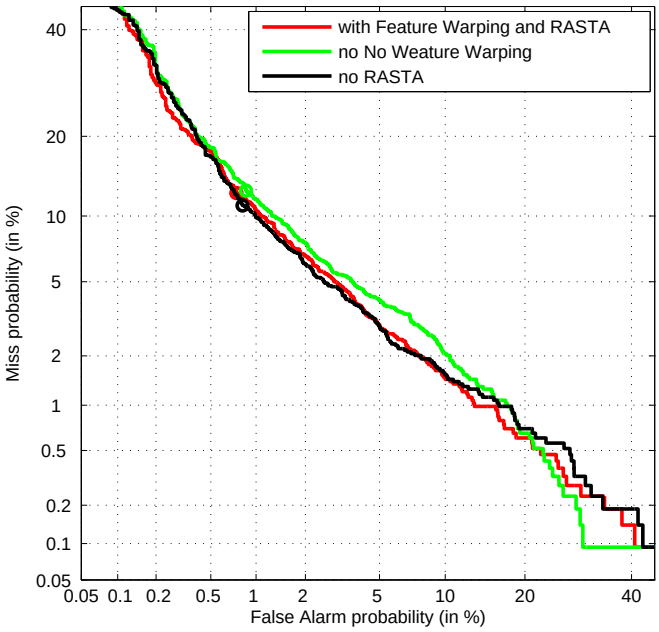

Fig. 2. The importance of RASTA filtering and Feature Warping.

System	SRE2005		SRE2006	
	EER	$C_{\text{Det}}^{\min}$	EER	$C_{\text{Det}}^{\min}$
Full system	4.5%	.019	3.8%	.020
No RASTA	4.4%	.019	3.8%	.019
No Feature Warping	5.1%	.020	4.3%	.021

TABLE II
THE IMPORTANCE OF RASTA AND FEATURE WARPING.

stage. The dimensionality reduction from 52 to 39 was chosen as exactly the same scheme had already been proven to be effective for speech recognition [20]. Since it was not clear whether this scheme is optimal for our speaker verification system, reductions to various dimensionalities were examined and the best results were obtained without any dimensionality reduction[7] (last line of Table III).

The situation is different in the right half of Table III, where Feature Mapping and Eigenchannel Adaptation are used. Performances of systems using HLDA are still superior to the one that leaves HLDA out; however, the system with dimensionality reduction outperforms the one without reduction. The possible explanation is that the significant increase in GMM (and supervector) size makes it impossible to robustly estimate eigenchannels given the limited number of supervectors available for their estimation. The summary of HLDA and MLLT results can be also found in figure 3.

C. Eigenchannels vs. Feature Mapping

The left half of Table IV and dotted DET curves in figure 4 show the effect of Feature Mapping for systems without the following Eigenchannel Adaptation. The first two results (already presented in Table I) demonstrate the effectiveness

[7]HLDA without dimensionality reduction is often referred to as a Maximum Likelihood Linear Transform (MLLT) [21]

System	Without channel comp.				With channel comp.			
	SRE2005		SRE2006		SRE2005		SRE2006	
	EER	$C_{\text{Det}}^{\text{min}}$	EER	$C_{\text{Det}}^{\text{min}}$	EER	$C_{\text{Det}}^{\text{min}}$	EER	$C_{\text{Det}}^{\text{min}}$
No HLDA	10.6%	.047	9.3%	.048	5.1%	.024	5.0%	.025
HLDA 52 $\rightarrow$ 39	9.7%	.042	8.2%	.041	4.5%	.019	3.8%	.020
HLDA 52 $\rightarrow$ 52	8.7%	.038	7.5%	.037	4.6%	.023	4.2%	.021

TABLE III
THE EFFECT OF HLDA ON SYSTEM PERFORMANCE.

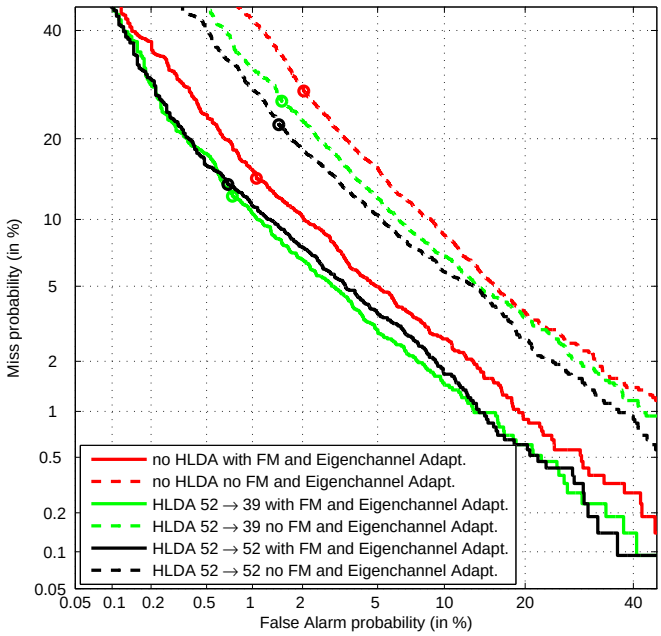

Fig. 3. The effect of HLDA and MLLT (HLDA without dimensionality reduction) on system performance.

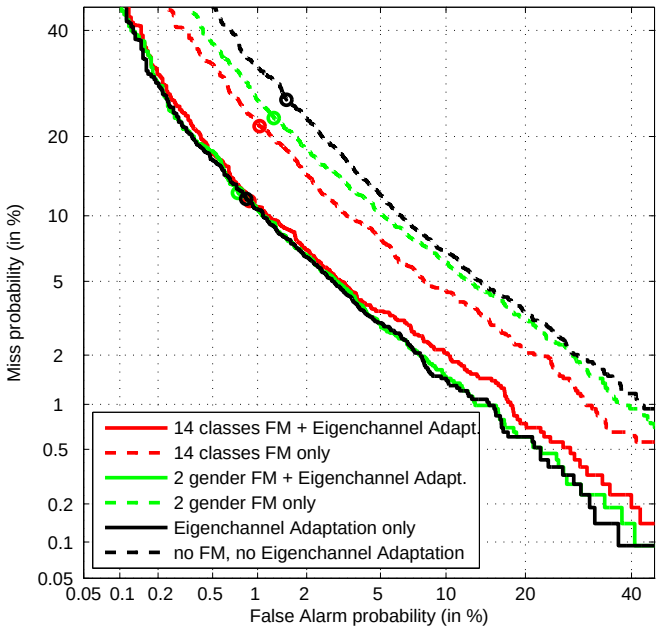

Fig. 4. The importance of Feature Mapping and Eigenchannel Adaptation.

of Feature Mapping at this stage. In the third line, the performance of a system using Feature Mapping based on only two models adapted on male and female specific data is shown. This allows us to compensate for the fact that our system uses only a single UBM instead of the usual approach where two genders are handled separately using two UBMs. Although such 2-gender Feature Mapping significantly outperforms the system leaving Feature Mapping out, it still reaches only about half of the gain in performance compared to 14 classes Feature Mapping used in our final system.

The right half of Table IV and solid DET curves in figure 4 show similar results for systems applying also Eigenchannel Adaptation. We can see that without Feature Mapping, Eigenchannel Adaptation causes an impressive improvement in system performance (more than 50% relative in both EER and $C_{\text{Det}}^{\text{min}}$ points). There is *no advantage in using Feature Mapping after the Eigenchannel Adaptation is applied*, which allows us to simplify the verification system considerably by leaving Feature Mapping out. In fact, the use of our 14 classes Feature Mapping causes even slight degradation in the performance. It was surprising for us that even 2-gender Feature Mapping did not turn out to be effective, as eigenchannels are not trained to model the directions of differences between male and female specific models.

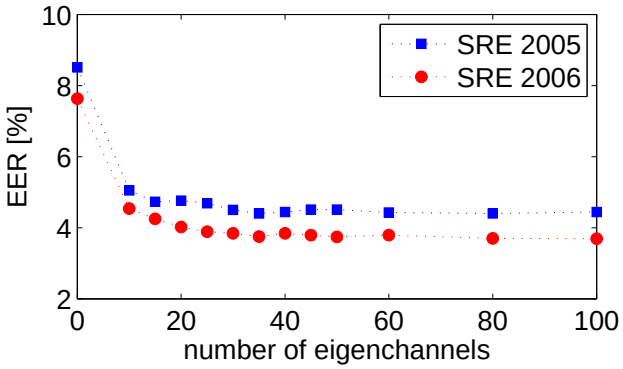

Fig. 5. The dependency of EER on the number of eigenchannels used for adaptation.

D. Number of eigenchannels

The number of eigenchannels was chosen to be $R = 30$ for our system submitted to SRE2006 NIST evaluations. Figure 5 shows the dependency of EER on the number of eigenchannels used for adaptation. A similar trend has also been observed for $C_{\text{Det}}^{\text{min}}$ values. It can be seen that our system is not very sensitive to the exact selection of the number of eigenchannels.

System	Without Eigenchannel Adapt.				With Eigenchannel Adapt.			
	SRE2005		SRE2006		SRE2005		SRE2006	
	EER	$C_{\text{Det}}^{\text{min}}$	EER	$C_{\text{Det}}^{\text{min}}$	EER	$C_{\text{Det}}^{\text{min}}$	EER	$C_{\text{Det}}^{\text{min}}$
No Feature Mapping	9.7%	.042	8.2%	.041	4.6%	.019	3.8%	.020
14 classes Feature Mapping	7.3%	.033	6.2%	.032	4.6%	.020	4.0%	.020
2-gender Feature Mapping	8.5%	.037	7.6%	.036	4.5%	.019	3.8%	.020

TABLE IV
THE IMPORTANCE OF FEATURE MAPPING AND EIGENCHANNEL ADAPTATION.

System	SRE2005		SRE2006	
	EER	$C_{\text{Det}}^{\text{min}}$	EER	$C_{\text{Det}}^{\text{min}}$
no T-n., no RASTA, no FM, 50 EA	4.4%	.017	3.6%	.018

TABLE V
RESULTS OF THE FINAL TUNED AND SIMPLIFIED SYSTEM.

V. FUSING WITH SVM BASED SYSTEMS

The performance of the GMM system was also tested in combination with speaker recognition systems based on a different classification paradigm – Support Vector Machines (SVM). Figure 6 contains a summary of results for SRE2006 primary condition. Results are presented for BUT stand-alone systems as well as for fused systems that were BUT and STBU submissions into the SRE2006 NIST evaluations.

These systems (from the worst to the best) are:

- SVM-MLLR, where MLLR and constrained MLLR (CMLLR) speaker adaptation matrices from a speech recognition system are classified by SVM. Two variants are shown: with and without T-norm
- SVM-GMM, where GMM supervectors are classified by SVMs. Two variants are shown: with and without T-norm
- **GMM** is the full system described in this paper. Two variants (already presented in Table I) are shown: with and without T-norm
- BUT02 is a fusion of 3 systems: GMM, SVM-GMM and SVM-MLLR, all with T-norm applied
- BUT01 (BUT primary system) is a fusion of 6 systems: GMM, SVM-GMM and SVM-MLLR, each in two variants: with and without T-norm
- STBU1-N is fusion of 10 systems from the partners in the STBU consortium.
- STBU1-U (STBU primary system) is fusion of the same 10 systems, plus one more SVM-GMM system implementing unsupervised adaptation to test data according to SRE2006 NIST evaluation rules [1].

A detailed description of different systems can be found in [5], [6]. The fusion was performed using linear logistic regression implemented in the FoCal toolkit[8] and it is also described and commented on in [6].

VI. CONCLUSION

BUT GMM system contains nothing more than techniques that were already published – its main contribution is in a thorough analysis and discussion of these techniques in a full speaker recognition system. Starting in the feature extraction,

[8]www.dsp.sun.ac.za/~nbrummer/focal/

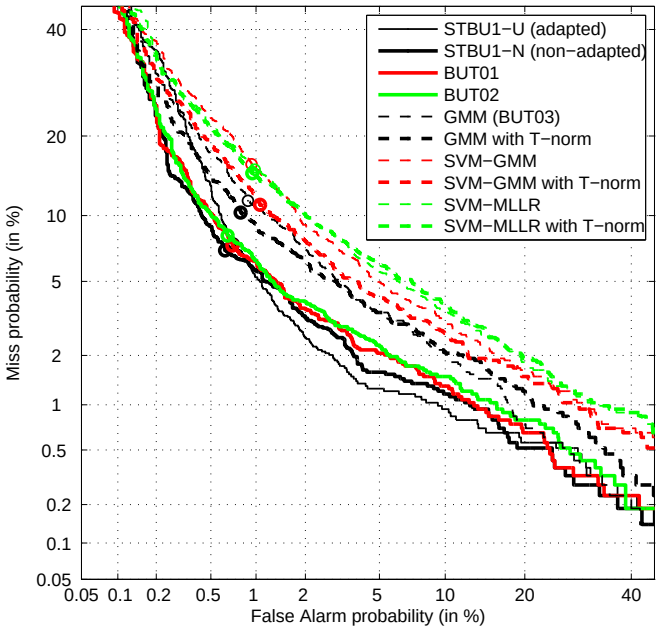

Fig. 6. Fusion of GMM system with SVM based systems.

the main conclusion is that RASTA did not help in the full system. On the other hand, HLDA significantly improved its performances, although we know that there is still work to be done (different dimensionality reductions examined with the full system, not using triple-deltas, etc.). In fighting the channel variability, even the simple Eigenchannel Adaptation turned out to be very effective, erasing the advantages of Feature Mapping, which is actually not important when applied together with Eigenchannel Adaptation. Table V presents the results of the final tuned and simplified system, containing 50 eigenchannels, no T-norm, no RASTA and no Feature Mapping. All the conclusions may, however, not hold for other than 1-side training, 1-side test condition examined in this work. Our current and future work aims at these conditions as well as at using the described GMM system as an excellent baseline for further experiments.

REFERENCES

[1] "The NIST year 2006 speaker recognition evaluation plan," 2006, Available from: http://www.nist.gov/speech/tests/spk/2006/sre-06_evalplan-v9.pdf.
[2] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1–3, pp. 19–41, 2000.
[3] A. Solomonoff, W. Campbell, and I. BoardmanCampbell, "Advances in channel compensation for SVM speaker recognition," in *Proc. ICASSP*, vol. I, Philadelphia, PA, USA, Mar. 2005, pp. 629–632.

[4] A. Stolcke, L. Ferrer, S. Kajarekar, E. Shriberg, and A. Venkataraman, "MLLR transforms as features in speaker recognition," in *Proc. Eurospeech*, Lisbon, Portugal, Sep. 2005, pp. 2425–2428.
[5] P. Matějka, L. Burget, P. Schwarz, O. Glembek, M. Karafiát, F. Grézl, J. Černocký, D. A. van Leeuwen, N. Brümmer, and A. Strasheim, "STBU system for the NIST 2006 speaker recognition evaluation," in *Proc. ICASSP*, Honolulu, Hawaii, USA, Apr. 2007.
[6] N. Brümmer, L. Burget, J. Černocký, O. Glembek, F. Grézl, M. Karafiát, D. A. van Leeuwen, P. Matějka, P. Schwarz, and A. Strasheim, "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006," *IEEE Trans. on Audio, Speech and Language Processing*, 2007, submitted.
[7] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. A Speaker Odyssey*, Crete, Grece, 2001, pp. 213–218.
[8] S. van Vuuren and H. Hermansky, "On the importance of components of the modulation spectrum for speaker verification," in *Proc. ICSLP*, vol. 7, Sydney, Australia, May 1998, pp. 3205–3208.
[9] N. Kumar, "Investigation of silicon-auditory models and generalization of linear discriminant analysis for improved speech recognition," Ph.D. dissertation, John Hopkins University, Baltimore, 1997.
[10] D. A. Reynolds, "Channel robust speaker verification via feature mapping," in *Proc. ICASSP*, vol. II, Apr. 2003, pp. 53–56.
[11] N. Brümmer, "Spescom DataVoice NIST 2004 system description," in *Proc. NIST Speaker Recognition Evaluation 2004*, Toledo, Spain, Jun. 2004.
[12] P. Schwarz, P. Matějka, and J. Černocký, "Hierarchical structures of neural networks for phoneme recognition," in *Proc. ICASSP*, France, May 2006, pp. 325–328.
[13] M. J. F. Gales, "Semi-tied covariance matrices for hidden markov models," *IEEE Transactions Speech and Audio Processing*, vol. 7, pp. 272–281, 1999.
[14] M. Mason, R. Vogt, B. Baker, and S. Sridharan, "Data-driven clustering for blind feature mapping in speaker verification," in *Proc. Eurospeech*, Lisbon, Portugal, Sep. 2005, pp. 3109–3112.
[15] D. A. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," in *Proc. Eurospeech*, Rhodes, Greece, Sep. 1997, pp. 963–966.
[16] R. Auckenthaler, M. Carey, and H. Lloyd-Tomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, pp. 42–54, 2000.
[17] P. Kenny and P. Dumouchel, "Disentangling speaker and channel effects in speaker verification," in *Proc. ICASSP*, vol. 1, Montreal, Canada, May 2004, pp. 47–40.
[18] R. Vogt, B. Baker, and S. Sridharan, "Modelling session variability in text-independent speaker verication," in *Proc. Eurospeech*, Lisbon, Portugal, Sep. 2005, pp. 3117–3120.
[19] R. Vogt and S. Sridharan, "Experiments in session variability modelling for speaker verification," in *Proc. ICASSP*, vol. 1, Toulouse, France, May 2006, pp. 897–900.
[20] T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, I. McCowan, D. Moore, V. Wan, R. Ordelman, and S. Renals, "The 2005 AMI system for the transcription of speech in meetings," in *Proc. NIST Rich Transcription 2005 Spring Meeting Recognition Evaluation*, Edinburgh,UK, Jul. 2005.
[21] R. Gopinath, "Maximum likelihood modeling with Gaussian distributions for classification," in *Proc. ICASSP*, vol. II, Seattle, Washington, USA, May 1998, pp. 661–664.

Lukáš Burget (Ing. [MS]. Brno University of Technology, 1999, Ph.D. Brno University of Technology, 2004) is employed as an assistant professor at the Faculty of Information Technology, University of Technology, Brno, Czech Republic. The topic of his PhD dissertation that he successfully defended in November 2004 was: "Complementarity of Speech Recognition Systems and System Combination". From 2000 to 2002, he was a visiting researcher at OGI Portland, USA under the supervision of Prof. Hynek Hermansky. He is a member of IEEE and ISCA. His scientific interests are in the field of speech processing, namely acoustic modeling for speech recognition.

Pavel Matějka (Ing. [MS]. Brno University of Technology, 2001) is a PhD student at the Institute of Radio-electronics, Faculty of Electrical Engineering and Communication and Department of Computer Graphics and Multimedia, FIT, BUT. He is planning to submit his doctoral thesis "Language identification based on phonetic cues" in summer 2007. He has been with the Anthropic speech processing group at the Oregon Graduate Institute of Science and Technology, USA. He is a member of IEEE and ISCA. His research interests include speaker recognition, language identification, speech recognition, namely phoneme recognition based on novel feature extractions (temporal patterns) and neural networks. He has been active also in keyword-spotting and on-line implementation of speech processing algorithms. He was a finalist in the Student paper contest at ICASSP2006 in Toulouse.

Petr Schwarz (Ing. [MS]. Brno University of Technology, 2001) is a PhD student of Speech processing group at the Faculty of Information Technology (FIT), BUT since September 2001 and is planning to submit his doctoral thesis "Robust phoneme recognition" in 2007. He has been with the Anthropic speech processing group of the Oregon Graduate Institute of Science and Technology, USA. He is a member of IEEE and ISCA. His research interests include speech recognition, namely phoneme recognition based on novel feature extractions (temporal patterns) and neural networks. He has been active also in keyword-spotting and on-line implementation of speech processing algorithms.

Ondřej Glembek (Ing. [MS]. Brno University of Technology, 2005) was student at the Brno University of Technology, Faculty of Electrical Engineering and Computer, later the Faculty of Information Technology from 1999. From September till December 2003, he was at the University of Joensuu, Finland as a participant of the Socrates/Erasmus program. From October till November 2004, he was working on a project concerning wavelet transforms at Izhevsk State Technical University, Izhevsk, Russia. From 2005, he is a PhD student in Speech@FIT - he is concentrating on acoustic modeling for speech recognition, recognition of Czech and STK toolkit development.

Jan "Honza" Černocký (Ing. [MS] 1993 Brno University of Technology (BUT); Dr. [PhD] 1998 Universite Paris XI and BUT) was with the Institute of Radio-electronics, BUT (Faculty of Electrical Engineering and Computer Science) as an assistant professor from 1997. Since February 2002, he is with the Faculty of Information Technology (FIT), BUT as an Associate Professor (Doc.) and Deputy Head of the Institute of Computer Graphics and Multimedia. With Prof. Hynek Hermansky he is leading the Speech@FIT group at FIT BUT. He supervises several PhD students, and coordinates Speech@FIT activities in several European and national projects. His research interests include signal processing, speech processing (very low bit rate coding, verification, recognition), segmental methods, data-driven determination of speech units and speech corpora. He is a member of IEEE and ISCA and serves on the board of the Czechoslovak section of IEEE.