STBU SYSTEM FOR THE NIST 2006 SPEAKER RECOGNITION EVALUATION

P. Matějka, L. Burget, P. Schwarz, O. Glembek,
M. Karafiát, F. Grézl and J. Černocký [*]
Brno University of Technology
Speech@FIT, Faculty of Information Technology
Brno, Czech Republic

D. A. van Leeuwen
TNO Human Factors
Postbus 23, 3769 ZG Soesterberg
The Netherlands

N. Brümmer
Spescom DataVoice
Stellenbosch, South Africa

A. Strasheim
University of Stellenbosch
Department of Electrical and Electronic Engineering
Stellenbosch, South Africa

ABSTRACT

This paper describes STBU 2006 speaker recognition system, which performed well in the NIST 2006 speaker recognition evaluation. STBU is consortium of 4 partners: Spescom DataVoice (South Africa), TNO (Netherlands), BUT (Czech Republic) and University of Stellenbosch (South Africa). The primary system is a combination of three main kinds of systems: (1) GMM, with short-time MFCC or PLP features, (2) GMM-SVM, using GMM mean supervectors as input and (3) MLLR-SVM, using MLLR speaker adaptation coefficients derived from English LVCSR system. In this paper, we describe these sub-systems and present results for each system alone and in combination on the NIST Speaker Recognition Evaluation (SRE) 2006 development and evaluation data sets.

Index Terms— Speaker recognition, GMM, SVM, eigen-channel, NAP.

1. INTRODUCTION

The 2006 NIST Speaker Recognition (SRE-06) evaluation was part of an ongoing series of yearly evaluations conducted by NIST. These evaluations provide an important contribution to the direction of research efforts and the calibration of technical capabilities. They are intended to be of interest to all researchers working on the general problem of text independent speaker recognition. For more information see: *http://www.nist.gov/speech/tests/spk/2006*.

STBU consortium was created to encourage collaboration between four institutes and to learn and share the technologies and knowhow available between:

- Spescom DataVoice (SDV), South Africa
- TNO, Netherlands
- Brno University of Technology (BUT), Czech Republic
- University of Stellenbosch (SUN), South Africa

In this paper, we first describe sub-systems developed at all four partners. We present the results of separate sub-systems and the final submitted systems on both development and evaluation data sets.

2. SYSTEM DESCRIPTION

We used three main kinds of systems:

- GMM, with short-time MFCC or PLP features. (SDV, BUT).
- GMM-SVM, using GMM mean supervectors as input. (SDV, TNO, BUT, SUN).
- MLLR-SVM, using MLLR speaker adaptation coefficients derived from a speech recognizer (BUT, SUN).

All systems used linear supervector-space channel compensation techniques. In the GMM case (BUT), this technique is referred to as eigen-channel MAP-adaptation. In the SVM case, it is referred to as NAP (nuisance attribute projection). In all cases, we used SRE 2004 and 2005 data to derive these adaptation coefficients or projection matrices.

2.1. Eigen-channel GMM (BUT)

At first, speech/silence segmentation is performed by our Hungarian phoneme recognizer [1], where all phoneme classes are linked to 'speech' class.

12 MFCC coefficients plus C0 are computed and cepstral mean subtraction and short time gaussianization over 300 frames are applied. RASTA filtering of the features follows. First, second and third order derivatives computed over 5 frames are appended to each feature vector, which results in dimensionality 52. HLDA [2] (where individual UBM mixture components are considered as classes) is used to decorrelate the features and reduce the dimensionality from 52 to 39.

[*] This work was partly supported by European projects AMI (FP6-506811) and Caretaker (FP6-027231), and by Grant Agency of Czech Republic under project No. 102/05/0278. The hardware used in this work was partially provided by CESNET under project No. 119/2004, No. 162/2005 and No. 201/2006. Lukáš Burget was supported by Grant Agency of Czech Republic under project No. GP102/06/383.

Feature mapping [3] with 14 models adapted from UBM for different conditions is used: 6 models were adapted for 3 channels (cell, cord, stnd) and 2 genders given the labels from 2004 test data. Remaining 8 models were initially adapted for 4 channels (cdma, cord, elec, gsmc) and 2 genders on labels obtained from TNO's channel recognition output (for 2005 SRE). However, these 8 models were then iteratively used to re-cluster the training data in unsupervised fashion and again adapted using the new clustering (20 iteration lead to stable clustering) [4].

GMM models adapted from UBM (2048 Gaussians) by MAP-adaptation [5] were used to model the target speakers (only means were adapted). Relevance factor 19 was used for the MAP adaptation. For each trial, the target model and UBM are adapted to the channel of test segment using eigen-channel adaptation [6], where mean super-vector m is adapted to $m_a = m + Vx$. Here, V is eigen-channel space matrix estimated the same way as for SVM-GMM system (see description below). The weight vector x, which is considered to be normally distributed, is obtained by maximizing probability $p(\mathrm{data}|m + Vx)p(x)$ in one iteration of EM algorithm.

The final score for each trial is given by log likelihood ratio: $\log\ p(\mathrm{data}|m_a) - \log\ p(\mathrm{data}|\mathrm{ubm}_a)$, where data is the test segment, m_a is channel adapted target speaker model and ubm_a is channel adapted UBM model.

Note, that for T-normalized version of GMM system, each T-norm model is also adapted to the channel of tested segment. For more details and analysis of this eigen-channel GMM system see [7].

2.2. GMM-SVM T-norm (BUT)

This system was based on GMMs but their means were classified by support-vector machines (SVM) [8].The feature extraction and UBM training was done in the same way as above in BUT eigen-channel GMM, but the UBM had only 512 Gaussian components and eigen-channel adaptation was not applied. Each training, test and background segment is represented by means of its Gaussian components: each mean is normalized by the corresponding standard deviation. All normalized mean vectors of all GMM mixture components are then concatenated to form one super-vector.

Nuisance attribute projection (NAP) [8] is used to remove unwanted channel variability. NAP is based on eigen-channel spaces given by eigenvectors of average within class covariance matrix, where each class is represented by super-vectors estimated on different segments spoken by the same speaker (estimated on SRE 2004 data). We used first 40 NAP eigenvectors. Also, rank normalization was used (the target feature distribution was trained on impostor speakers data).

The SVM used to classify mean super-vectors uses linear kernel. It is trained on one positive example from the target speaker. The negative examples are taken from 2002 SRE data (260 speakers) and from Fisher1 (2606 speakers). In the testing, the trial is scored by the respective SVM. The SVM training and scoring software was built with LibSVM library [9].

2.3. MLLR-SVM (BUT)

In this system, the coefficients from constrained maximum likelihood linear regression (CMLLR) and maximum likelihood linear regression (MLLR) transforms, estimated in an ASR system, are classified by SVMs [10].

The core of AMI system submitted to NIST RT 2005 [11] was used in MLLR/CMMLR work. We did not generate our own ASR transcriptions, but used the ASR output provided by NIST. Since NIST did not provide pronunciation dictionary, we used the AMI dictionary and we generated the missing pronunciations automatically. With this, we were able to generate the triphone alignment and to apply VTLN.

CMLLR and MLLR transforms are trained for each speaker. At first, CMLLR is trained with two classes (speech + silence). On the top of it, MLLR with three classes (2 speech classes obtained by automatic clustering on the ASR training data + silence) is estimated. Both CMLLR and MLLR transform matrices were estimated as block-diagonal in 13-coefficient wide streams (reminiscence from originally used MFCCs).

The transform matrices from CMLLR speech classes and MLLR are concatenated to one vector with $3 \times 3 \times 13 \times 13 + 3 \times 39 = 1638$ features. Fifteen NAP eigenvectors are estimated on all usable SRE 2004 data (1-side, 3-side, 8-side and 16-side) and rank normalization is trained on impostor speakers.

The same SVM classification as above is used. The impostor data (310 speakers) was taken from NIST 2004, as this also contains the ASR transcripts provided by NIST.

2.4. GMM-SVM T-norm (SUN)

This system used the GMM means from the BUT 512-mixture, 39-dimensional feature GMM system, which resulted in 19968-dimensional supervectors. Each supervector dimension was normalized by dividing by corresponding standard deviation of the GMM UBM. LibSVM [9] C-classification with a linear kernel was used for all the SUN GMM-SVM and MLLR-SVM systems. 2606 speakers from the Fisher database were used as background speakers for the SVM. 300 from this set are also used to train leave-one-out T-norm models.

Experiments on the 2005 data indicated that using 40 NAP eigenvectors provided the best EER. 4433 segments from the NIST 2004 Extended data, spoken by 301 speakers were used to obtain a 19968×40 adaptation matrix. All supervectors were adapted using this matrix.

2.5. MLLR-SVM (SUN)

Two variations of this system were implemented. In both cases, CMLLR and MLLR transforms from the BUT system were used: (1) CMLLR + 1 MLLR transform, (2) CMLLR + 2 MLLR transforms. The silence transform was discarded in all cases. Each transform was made up of a block-diagonal matrix containing three 13×13 matrices and a 39-dimensional bias vectors, yielding 546 components per transform, or 1092- and 1638-dimensional supervectors. Rank normalization was applied to the supervectors, with normalization parameters estimated from 4266 segments from the NIST 2004 Extended data, spoken by about 310 speakers. The same data was used to train the SVM models.

T-norm was found to reduce performance, so it was not utilized in this system.

For the MLLR system, experiments on the 2005 data indicated that 15 NAP eigenvectors provided the best EER. 4159 segments from the NIST 2004 Extended data, spoken by 301 speakers. This yielded a 1092×15 or 1638×15 adaptation matrix. All supervectors were adapted using this matrix.

2.6. GMM-SVM T-norm, w/o unsupervised adapt. (TNO)

This is a system similar to the one described in section 2.2. Using feature-warped normalized PLP features a 512-Gaussian UBM was trained on 1640 channel-balanced speaker sides from Switchboard and Fisher databases. The first channel compensation using feature-mapping was performed using 16 channels defined on the same databases. Training and test segments were used to MAP adapt the UBM using means-only, and these means were used as supervectors. The second channel-compensation was applied using NAP, projecting out 40 dimensions of channel variability found in same-speaker conversation sides of NIST SRE-2004 1c4w trials. NAPped supervectors were used to train a linear-kernel SVM for each training segment (using a background of the same 1640 UBM-speakers), and test scores were obtained by calculating inner product of the test supervector with the folded SVM model. SVM scores were T-normalized using all SRE-2004 training speakers.

2.7. GMM-SVM T-norm, with unsupervised adapt. (TNO)

The system described in section 2.6 was also run in unsupervised adaptation mode. Given a model speaker, all test segments were processed in order. If a T-normed score exceeded a threshold a, the ‘current model’ GMM means supervector was MAP adapted towards the current test segment using relevance factor r, and a new SVM model was trained. The parameters $a = 4$, $r = 36$ were chosen to optimize C_{det} for NIST SRE-2005.

2.8. GMM-SVM forward and reverse, T-norm (SDV)

In the fusion of systems, we found it advantageous to include in each fusion several very similar (but not identical) GMM-SVM systems. These systems were different because each was built by a different team, using different front-ends, different development databases and somewhat different flavors of the NAP channel compensation technique described above. The GMM-SVM system put together by SDV had the following distinguishing features: (i) It used a very aggressive frame-selection procedure, retaining only about 25% of the total segment duration, using only speech frames from presumably strongly voiced syllable nuclei, detected by finding local maxima of an appropriately filtered energy contour. Moreover, frames where strong cross-talk between the two telephone channels was detected (by comparing channel energies) were also rejected. (ii) This system was diversified into two variants, namely (a) a *forward* system where SVM models were trained on train segments and then evaluated against test segments; and (b) a *reverse* system, where the roles of train and test segments were reversed.

3. SUBMISSION

We submitted three fusions of several sub-systems contributed by partners. All submissions are run only on the primary evaluation task, both with and without unsupervised adaptation.

3.1. Fusion and Calibration

All systems were fused with linear logistic regression [1]. We had the complication that not all sub-systems were able to contribute a score for each trial, because of failure to detect speech in training or test segment, or lack of ASR transcription. This necessitated a two step fusion strategy:

First, each system on its own was subjected to an affine calibration transformation, also trained via logistic regression. We used a logistic regression prior-weighting of 0.5 here. The training data for this calibration were all trials that the system could contribute out of the SRE 2005 (1c4w-1c4w) trials.

Next, scores for missing trials of each system were inserted as log-likelihood ratio (LLR) = 0. Now, all systems had valid scores for all trials and could be fused, with linear logistic regression, but this time using a prior-weighting of 0.0917 to best serve the NIST-CDET operating point.

Two of our systems: STBU-1 and STBU-2, relied purely on the affine calibration afforded by the fusion step. For these two systems, we did (somewhat arbitrarily) clip the LLR magnitude to ± 15. All that remained, was to threshold decisions at an LLR threshold of $\log 9.9$ and then to exponentiate LLR scores to submit LR scores. STBU-3 followed the fusion with a soft saturating non-linearity, called S-Cal, which is also trained with logistic regression. For details on fusion and calibration, see the Focal toolkit [1].

3.2. Submission systems

STBU-1 - The unsupervised adaptation mode (u-mode) of this system is our primary system. This is an 11-fold fusion of: GMM-SVM forward, T-normed (SDV), GMM-SVM reverse, T-normed (SDV), Eigen-channel GMM (BUT), Eigen-channel GMM T-normed (BUT), GMM-SVM T-normed (BUT), MLLR-SVM (BUT), GMM-SVM T-normed (SUN), MLLR-SVM v1 and v2 (SUN), GMM-SVM T-normed, without and with unsupervised adaptation (TNO). For the non-adaptive (n-mode) variant of this system, we simply omitted the last sub-system.

[1] Tools for fusion and calibration of automatic speaker detection systems, *http://www.dsp.sun.ac.za/∼nbrummer/focal/*.

system	SRE 2005 data		SRE 2006 data	
	DCF	EER	DCF	EER
GMM (BUT)	.0174	3.88%	.0178	3.44%
GMM-SVM (SUN)	.0153	4.19%	.0171	3.61%
GMM-SVM (BUT)	.0158	4.66%	.0185	3.71%
GMM-SVM-U (TNO)	.0116	3.72%	.0185	3.81%
GMM-SVM (TNO)	.0178	5.17%	.0190	4.10%
GMM-SVM-TF (SDV)	.0221	6.05%	.0227	4.91%
GMM-SVM-TR (SDV)	.0220	6.10%	.0238	5.18%
MLLR3-SVM (SUN)	.0212	6.05%	.0218	4.49%
MLLR3-SVM (BUT)	.0196	6.17%	.0220	4.78%
MLLR2-SVM (SUN)	.0264	7.50%	.0270	5.56%
STBU-1U	.0070	2.98%	.0132	2.26%
STBU-1	.0096	3.21%	.0126	2.32%
STBU-2U	.0073	3.17%	.0132	2.26%
STBU-2	.0099	3.59%	.0129	2.51%
STBU-3U			.0132	2.26%
STBU-3			.0126	2.32%

Table 1. Results of the sub-systems and the submitted one on primary condition: English trials.

system	SRE 2005 data		SRE 2006 data	
	DCF	EER	DCF	EER
GMM (BUT)	.0201	4.83%	.0283	5.40%
GMM-SVM (TNO)	.0192	5.77%	.0285	6.04%
MLLR-SVM (BUT)	.0224	7.15%	.0327	7.57%
STBU-1	.0114	3.97%	.0214	3.83%
STBU-1U	.0085	3.50%	.0208	3.30%

Table 2. The best performing sub-systems from each category and the submitted results on all trials.

STBU-2 - This is the same as STBU-1 in all respects, except that the eigen-channel GMM systems were omitted. This makes this STBU-2 a pure fusion of SVM systems.
STBU-3 - This system is the same as STBU-1, except that the above-mentioned S-Cal non-linearity was added as a further calibration aid. We found very similar, but not identical, S-Cal coefficients for the n and u modes. On our development data, APE-curve analysis showed that this brought a significant improvement in quality of calibration.

4. RESULTS

Table 1 describes results on primary condition for development data (SRE 2005) and for evaluation data (SRE 2006). Results are reported for all sub-systems together with fused results which are with and without unsupervised adaptation.

Table 2 describes results on all trials from development and evaluation data. Only results of the best sub-system from each category is presented.

5. CONCLUSION

Over a period of about 10 weeks, we developed completely new systems and added new features to old ones. Over 600 emails were sent and finally we fused a selection of all our sub-systems. Although we shared (wiki, email, sms) papers, advices, ideas, formulas, code, supervectors, scores, there was still a competition inside the STBU partners which paid off in final results. From the experiments conducted during the evaluation it comes out that the choice of data for UBM, SVM-background, NAP/eigen-channel, T-norm is important. We found out, that independently developed systems (even with the same structure) tend to fuse well, especially thanks to the excellent Focal toolkit.

6. REFERENCES

[1] P. Schwarz, P. Matějka, and J. Černocký, "Hierarchical structures of neural networks for phoneme recognition," in *Proc. ICASSP*, France, May 2006, pp. 325–328.

[2] N. Kumar, *Investigation of Silicon-Auditory Models and Generalization of Linear Discriminant Analysis for Improved Speech Recognition*, Ph.D. thesis, John Hopkins University, Baltimore, 1997.

[3] D. A. Reynolds, "Channel robust speaker verification via feature mapping," in *Proc. ICASSP*, Hong Kong, China, Apr. 2003, vol. II, pp. 53–56.

[4] M. Mason, R. Vogt, B. Baker, and S. Sridharan, "Data-driven clustering for blind feature mapping in speaker verification," in *Proc. Eurospeech*, Lisbon, Portugal, Sept. 2005, pp. 3109–3112.

[5] D. A. Reynolds, *A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification*, Ph.D. thesis, Georgia Institute of Technology, sep 1992.

[6] Niko Brummer, "Spescom DataVoice NIST 2004 system description," in *Proc. NIST Speaker Recognition Evaluation 2004*, Toledo, Spain, June 2004.

[7] L. Burget, P. Matějka, O. Glembek and J. Černocký, "Analysis of feature extraction and channel compensation in GMM speaker recognition system," *submitted to ICASSP*, 2007.

[8] A. Solomonoff, W. Campbell, and I. BoardmanCampbell, "Advances in channel compensation for SVM speaker recognition," in *Proc. ICASSP*, Philadelphia, PA, USA, Mar. 2005, vol. I, pp. 629–632.

[9] Chih-Chung Chang and Chih-Jen Lin, "Libsvm: a library for support vector machines," http://www.csie.ntu.edu.tw/∼cjlin/libsvm, 2001.

[10] A. Stolcke, L. Ferrer, S. Kajarekar, E. Shriberg, and A. Venkataraman, "MLLR transforms as features in speaker recognition," in *Proc. Eurospeech*, Lisbon, Portugal, Sept. 2005, pp. 2425–2428.

[11] Hain T. et al., "The 2005 AMI system for the transcription of speech in meetings," in *Proc. NIST Rich Transcription 2005 Spring Meeting Recognition Evaluation*, Edinburgh,UK, July 2005.