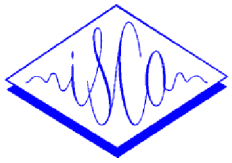
MLLR Techniques for Speaker Recognition

Marc Ferràs, Cheung Chi Leung, Claude Barras, Jean-Luc Gauvain



Odyssey 2008: The Speaker and Language Recognition Workshop
Stellenbosch, South Africa
January 21-24, 2008

MLLR Techniques for Speaker Recognition

Marc Ferràs[1], *Cheung Chi Leung*[1], *Claude Barras*[1,2] *and Jean-Luc Gauvain*[1]

[1]LIMSI-CNRS, BP 133, 91403, Orsay, France
[2]Univ Paris-Sud, F-91405, Orsay, France
{ferras,ccleung,barras,gauvain}@limsi.fr

Abstract

Maximum-Likelihood Linear Regression (MLLR) and Constrained MLLR (CMLLR) have been recently used for feature extraction in speaker recognition. These systems use (C)MLLR transforms as features that are modeled with Support Vector Machines (SVM). This paper evaluates and compares several of these approaches for the NIST Speaker Recognition task. Single CMLLR and up to 4-phonetic-class MLLR transforms are explored using Gaussian Mixture Models (GMM) and large-vocabulary speech recognition Hidden Markov Models (HMM), using both speaker recognition and speech recognition cepstral front-ends and normalizations. Results for the individual systems as well as in combination with two standard cepstral systems are provided. Relative gains of 3% and 12% were obtained when combining the best performing CMLLR-based and MLLR-based systems with two standard cepstral systems, respectively.

1. Introduction

Recently, several novel feature extraction approaches for speaker recognition have been proposed. Together with already well-known modeling techniques, such as Gaussian Mixture Models (GMM) or Support Vector Machines (SVM), these systems obtain excellent performance, comparable to that of the well-known MFCC-GMM paradigm. Gaussian Supervectors (GSV-SVM) [1] is one of such approaches, which successfully combines both GMM and SVM modeling together in a simple and easy-to-develop framework. In a different direction, modeling Maximum-Likelihood Linear Regression (MLLR) transforms by means of SVM has also become successful. GSV-SVM and MLLR-SVM become more and more present in state-of-the-art text-independent speaker recognition systems [2, 3].

Using MLLR transforms as features for speaker recognition was introduced in [4]. One or several MLLR transforms are estimated using a large-vocabulary HMM-based speech recognition system along with the automatic transcription of the speech data. These linear transforms represent the difference between a speaker-independent and speaker-dependent model and they are used as feature vectors to be classified by a SVM. This approach typically requires acoustic models of an Automatic Speech Recognition (ASR) system, as well as a pronunciation lexicon. A slightly different approach is presented in [5]. There, Constrained MLLR (CMLLR) transforms are computed on a GMM/UBM the speaker-independence of which is improved by means of Speaker Adaptive Training (SAT) [6]. Since an overall transform is computed for the speaker training data, there is no need for phonetic-class segmentation and, thus, no need for transcripts either.

In this paper we focus on the comparison of several CMLLR and MLLR approaches which can be found in current speaker recognition systems. We explore both approaches with either GMM or ASR phonemic HMM. We also investigate the role of the cepstral features, since they rarely follow the same normalization steps in speech recognition and speaker recognition systems. The paper is organized as follows: Section 2 introduces MLLR and CMLLR and describes the way they are used for feature extraction purposes in speaker recognition. Section 3 describes all the components of the speaker verification system as well as the evaluation task. Results on the NIST Speaker Recognition Evaluation 2005 are given for the described systems in Section 4. These include individual results as well as fusion results with other acoustic-level systems. Conclusions are given in Section 5.

2. MLLR and CMLLR in Speaker Recognition

Maximum-Likelihood Linear Regression (MLLR) [7, 8] and its variant Constrained MLLR (CMLLR) [9] are two adaptation techniques typically used for speaker adaptation purposes in speech recognition systems. The parameters of an HMM are adapted via an affine transform. This results in a significant reduction of the amount of parameters to be estimated compared to a direct adaptation approach [10]. In the general MLLR framework, both mean and variance parameters are transformed, as

$$\hat{\mu} = \mathbf{A}\mu + \mathbf{b} \quad (1)$$

$$\hat{\mathbf{\Sigma}} = \mathbf{H}\mathbf{\Sigma}\mathbf{H}^T \quad (2)$$

where μ is a mean vector in the model, $\mathbf{\Sigma}$, its corresponding covariance matrix and $\hat{\mu}$ and $\hat{\mathbf{\Sigma}}$ are the adapted mean and covariance matrix, respectively. The likelihood function of the adaptation data given the model is to be maximized with respect to the transform parameters $(\mathbf{A}, \mathbf{b}, \mathbf{H})$. This is typically done using Expectation Maximization (EM) in two steps [8], by first estimating the mean transform given by $(\mathbf{A}, \mathbf{b})$ and,

This work was partially funded by the European Commission under the FP6 Integrated Project IP 506909 CHIL.

next, the covariance transform $\mathbf{H}$. To further reduce the number of parameters a diagonal transformation matrix is typically assumed or only mean adaptation is performed.

Constrained MLLR (CMLLR) [9] forces the transform to be the same for both mean and variance parameters, $(\hat{\mu}, \hat{\mathbf{\Sigma}})$, as

$$\hat{\mu} = \mathbf{A_c}\mu - \mathbf{b_c} \quad (3)$$

$$\hat{\mathbf{\Sigma}} = \mathbf{A_c}\mathbf{\Sigma}\mathbf{A_c}^T \quad (4)$$

This constraint allows the covariance matrix to be transformed as well without increasing the amount of parameters to be estimated. Furthermore, it makes possible to apply the transform at the feature level as

$$\hat{\mathbf{o_t}} = \mathbf{A_c}^{-1}\mathbf{o_t} + \mathbf{A_c}^{-1}\mathbf{b_c} \quad (5)$$

where $\mathbf{o_t}$ is the observed feature vector at time t.

2.1. MLLR Feature extraction

The first approach to MLLR feature extraction for speaker recognition was proposed in [4]. There, MLLR transforms are estimated for each speaker of interest using a large-vocabulary HMM-based ASR system. One or more matrices can be computed depending on the amount of speech data available for adaptation and the desired number of phonetic classes. Using many classes results in a finely represented phonetic space but less speech data is available for each class-dependent transform. Once computed, the transform parameters are stacked together to be used as a feature vector that is suited for SVM modeling. We will refer to this approach as MLLR/HMM from now on.

Since a large-vocabulary ASR system needs huge amounts of data and resources to be trained, a simple and cost-effective alternative to this approach is to replace the HMM by a GMM. Feature vectors are now to be aligned against a single state with a global gaussian mixture and, therefore, phonetic-class alignment is not possible anymore (multiple transforms could still be computed for a given subset of gaussians). Because of its simplicity, it is also feasible to perform training on any type of feature vector. In the context of speaker recognition, this translates into being able to use any type of front-end setup or normalization of the cepstral features, for instance, any number of cepstral coefficients or channel compensation technique. All other steps in the feature extraction are kept as in the system described above. We will refer to this simplified approach as MLLR/GMM.

2.2. CMLLR Feature Extraction

The approach to CMLLR feature extraction proposed in [5] is a natural extension of MLLR/GMM to CMLLR transforms. In addition, though, an iterative approach is adopted such that the features used for GMM training are CMLLR-transformed in a per-speaker basis to obtain a more speaker-independent GMM. Fig. 1 illustrates this process. A GMM/UBM is first trained using cepstral features from a set of background speakers. Next, assuming only one speaker per segment, a CMLLR transform is computed for each of the speakers. Finally, each of the CMLLR transforms is applied onto the corresponding segment and the GMM is trained again using the new features. This SAT-like scheme leads to a more speaker-independent model at each iteration. Once the GMM is ready, CMLLR transforms can be computed on it by following a procedure analogous to that in the MLLR/GMM approach. We will refer to this approach as CMLLR/GMM.

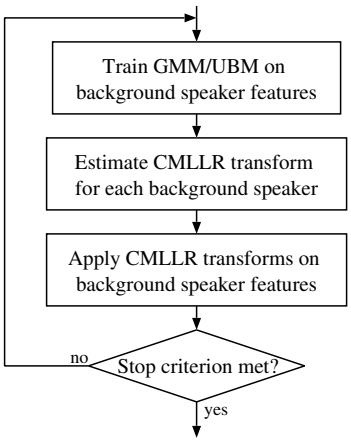


Figure 1: Block diagram for iterative CMLLR GMM/UBM re-estimation in the CMLLR/GMM approach.

For a large vocabulary ASR HMM, a CMLLR/HMM approach can be easily setup by extracting CMLLR transforms using ASR phonemic HMM, which may be trained using SAT as well. We will refer to this approach as CMLLR/HMM.

3. Experimental Setup

3.1. Task and evaluation data

All the systems were evaluated on speaker verification experiments conducted on conversational telephone speech. The system is asked to decide whether a given target speaker spoke in a particular speech segment. We used the primary condition task evaluation data of the NIST Speaker Recognition Evaluation 2005[1], containing 5-minute-long speech segments in English language, both for training and test. A total of 646 (274 male, 372 female) segments were available for target model training. Overall, 2429 test segments (1074 male, 1355 female) were scored against roughly 10 gender-matching impostors and against the true speaker.

3.2. Cepstral Feature Extraction

3.2.1. Speaker Recognition (PLP15N)

All the non-ASR-based systems explored in this paper shared the same cepstral front-end. Cepstral feature vectors were extracted every 10ms using a 30ms window on the 0-3.8kHz bandwidth. They consisted of 15 MEL-PLP cepstrum coefficients, 15 Δ coefficients plus Δ energy, and 15 $\Delta\Delta$ coefficients plus $\Delta\Delta$ energy (47 features). The frames selected by the

[1]The NIST year 2005 speaker recognition evaluation plan, http://www.nist.gov/speech/tests/spk/2005/

Snack Sound Toolkit[2] for pitch extraction were considered only, and unvoiced speech frames were dropped. Channel compensation for GSM, CDMA, TDMA, landline-carbon and landline-electret data was performed using per-gender feature mapping [11]. Speech segments from test speakers from the NIST SRE 1997 to 2002 evaluations (24769 speech segments) were chosen for model training. Around 6 hours of speech data were used to train each gender-dependent channel model. Finally, feature warping [12] was performed over a sliding window of about 3 seconds to reshape the histogram of the cepstral coefficients into a Gaussian distribution.

3.2.2. Speech Recognition (PLP12)

We used the LIMSI RT04 English CTS speech recognition system [13] for alignment and computation of MLLR transforms. Acoustic models are gender-independent. Computing (C)MLLR transforms for the speaker recognition task required all training and test cepstral features to be of the same type as used in the ASR system. The front-end used 39-dimensional feature vectors made out of 12 MEL-PLP cepstrum coefficients plus log-energy along with their corresponding Δ and $\Delta\Delta$ coefficients. Mean and variance normalization was next applied to each segment of interest.

3.3. MFCC-GMM system

The MFCC-GMM system [14] is based on GMM with diagonal covariance matrices trained using MAP adaptation [15]. For speaker modeling, GMMs were trained using MAP adaptation of the Gaussian means of the corresponding gender-dependent UBM using 3 iterations of the EM algorithm. Each of the two gender-dependent UBMs was a 1536-gaussian mixture model, built by merging three GMMs, each with 512 Gaussians trained on cellular, landline-electret and landline-carbon data. Around 60-hours of speech data was used to train each gender-dependent channel-specific 512-mixture GMM. The training data was chosen from target speakers in NIST SRE 97-01 and 03 evaluations and test speakers in NIST SRE03 evaluation (for a total of 9041 speech segments). Score normalization was performed using T-norm [16] on 500 speech segments (250 males and 250 females) from the Fisher corpus[3]. The first 5 minutes of each segment in this corpus were extracted for score normalization.

3.4. MFCC-SVM system

The MFCC-SVM system is based on several feature extraction steps that expand the discriminative power of the base cepstral features and SVM modeling [17]. We used a polynomial feature extraction scheme to transform the MEL-PLP features into high-dimensional feature vectors by means of a third order monomial expansion. The resulting features were variance normalized and averaged over the whole segment to obtain a single 20824-dimensional vector. The dimension of this speaker-specific vector was reduced via Kernel Principal Component Analysis (KPCA) [18] using a 2nd order cumulative homogeneous polynomial kernel, resulting in one 3200-dimensional feature vector per speaker. An affine transform mapped each feature component into the range [-1/$\sqrt{D}$,1/$\sqrt{D}$], D being the dimension of the feature vector, so that only normalized dot products were processed by the SVM.

The impostor speaker set consisted of 2243 speech segments from the NIST SRE04 training data plus 4854 speech segments from the Switchboard I corpus. All of them were in English language and had a minimum duration of 10 seconds of speech (after forced-alignment). This configuration allowed all SVM-based systems to share the same impostor data, as transcripts[4] were available for all of the 7097 segments.

Kernel PCA used a subset of the impostor speakers as training data. Statistics for feature normalization were also taken from the impostor speaker set. A linear kernel was chosen for SVM modeling using SVMTorch [5] from IDIAP. No T-norm score normalization was applied in this system.

3.5. CMLLR-SVM systems

CMLLR-SVM systems use one of the feature extraction schemes described in Section 2.2, i.e., using either the iterative GMM training approach plus CMLLR computation or CMLLR computation using the LIMSI RT04 English CTS ASR system. In the former approach, the two gender-dependent GMMs used 2 iterations of GMM/UBM re-estimation. The impostor speakers were used as the background speaker set. For the PLP15N cepstral front-end, the CMLLR transforms resulted in 2256-dimensional (47x47 + 47, including $\mathbf{b}$) feature vectors, after stacking their coefficients. For PLP12 features, the CMLLR transforms were 1560-dimensional (39x39 + 39). All of these were min-max normalized and modeled exactly in the same way as in the MFCC-SVM system.

3.6. MLLR-SVM systems

MLLR-SVM systems use one of the feature extraction schemes described in Section 2.1, i.e., computing MLLR transforms on either a GMM or an ASR HMM. In any case, only mean adaptation was performed. For MLLR/GMM, the MLLR transforms were computed on two gender-dependent GMMs directly trained on the background-speaker cepstral features. For MLLR/HMM, an analogous procedure is followed in the LIMSI RT04 English CTS ASR system, which uses speaker-independent (SI) acoustic models. We experimented with one to four MLLR transforms. The classes (non-speech/speech, non-speech/vowel/consonant, non-speech/vowel1/vowel2/consonant) were derived manually using acoustic and phonetic-level rules. MLLR tranforms for the non-speech class were not used as they were assumed not to carry any relevant speaker information. Feature vector dimensions were the same as in the CMLLR-SVM systems

[2]The Snack Sound Toolkit, http://www.speech.kth.se/snack/ .

[3]Fisher Corpus, LDC Catalog, http://www.ldc.upenn.edu/Catalog

[4]Human and ASR transcripts were available for SRE04 and Switchboard I corpora.

[5]SVMTorch: a Support Vector Machine for Large-Scale Regression and Classification Problems - http://www.idiap.ch/learning/SVMTorch.html

for one-class transforms, depending on the type of cepstral features, either PLP15N or PLP12. 1560, 3120 and 4680 features were used for two-class, three-class and four-class MLLR transforms, respectively. Feature normalization as well as modeling were kept the same as in CMLLR-SVM systems.

3.7. Score-level fusion

Each of the systems consisted of a forward sub-system that scored test speaker speech against train speaker models, and a backward sub-system that scored train speaker speech against test speaker models [19]. Therefore, 2 scores were obtained per system and trial.

A three-fold cross-validation scheme was used for evaluation purposes. The NIST SRE05 evaluation data was split into three independent datasets, the scores of which were zero-mean and unit-variance normalized based on the statistics of the two other sets.

As for system fusion, averaging was used by weighting each of the sub-system scores uniformly.

3.8. Performance Measure

As described in the NIST SRE05 evaluation plan, we used the Detection Cost Function (DCF) as the primary performance measure in our experiments. This function weights missed detections and false alarms as $DC = P_{Miss} + 9.9 \times P_{FalseAlarm}$. All results were reported as Minimal DCF (MDC) value, obtained a posteriori for the optimal decision threshold. Therefore, they do not include calibration mismatch. Since DC and MDC favor false alarm errors, Equal Error Rate (EER) is also provided as an alternative performance measure. Detection Error Trade-off (DET) curves are provided to assess system behaviour in the full range of operating points.

4. Results

We conducted experiments to explore the behaviour of several CMLLR-SVM and MLLR-SVM systems, by varying the type of model used to compute the MLLR transforms, and the cepstral features. In order to simplify system naming, Table 1 shows the naming convention used for all the tested systems.

System	Xform type	Model	Feature setup	SAT	#Classes	#Xforms
CG15	CMLLR	GMM	PLP15N	no	1	1
CG15/S	CMLLR	GMM	PLP15N	yes	1	1
CG12/S	CMLLR	GMM	PLP12	yes	1	1
CH12	CMLLR	HMM	PLP12	no	1	1
MG15	MLLR	GMM	PLP15N	no	1	1
MH12/1c	MLLR	HMM	PLP12	no	1	1
MH12/1t	MLLR	HMM	PLP12	no	2	1
MH12/2t	MLLR	HMM	PLP12	no	3	2
MH12/3t	MLLR	HMM	PLP12	no	4	3

Table 1: System naming convention for CMLLR-SVM and MLLR-SVM systems.

A two-axis set of experiments was first conducted for CMLLR-SVM systems. The first axis aimed at evaluating the impact of the model choice on performance. The second axis focused on the choice of cepstral features, i.e., PLP15N or PLP12 cepstral front-ends. Table 2 shows results for the CG15, CG15/S (using PLP15N features and none and one SAT iterations), CG12/S (PLP12 features and one SAT iteration) and CH12 (PLP12 features and ASR HMM modeling) systems. A gain of about 9% in MDC and 11% in EER was found when switching from GMM to HMM modeling (CG12/S vs. CH12). This may be explained by the derivation of more precise MLLR transforms when using the phone HMMs. However, when the GMM-based system takes advantage of speaker recognition feature normalizations (PLP15N vs PLP12), the reported improvement is more than that provided by HMM modeling. In that sense, a gain of 7% in MDC and 10% in EER relative terms is obtained (CG15/S vs. CH12). If the SAT-like procedure for GMM re-estimation is not used (CG15), performance decreases but the system still outperforms CH12. This stresses the importance of normalization techniques, particularly channel compensation, in speaker recognition. By taking advantage of these normalizations (CG15/S vs. CG12/S) up to a 20% gain in EER was found.

System	minDCF	EER (%)
CG15	.0397	9.77
CG15/S	.0393	8.90
CG12/S	.0468	11.23
CH12	.0423	9.94

Table 2: MDC and EER SRE 2005 results for several CMLLR-SVM systems. Systems are described by MLLRTransformType/ModelType/FeatureSetUp/NumberOfSATIterations parameters in the CMLLR computation (See Table 1 for system naming convention).

As for MLLR-SVM systems, we assessed performance as a function of model choice and, for HMM-based systems, the number of computed MLLR transforms, i.e., from one to four broad phonetic classes. Table 3 show results for the MG15 (using PLP15N features and GMM modeling) and MH12/1c, MH12/1t, MH12/2t and MH12/3t (using PLP12 features, ASR HMM modeling and 1, 2, 3 or 4 phonetic classes, i.e. 1, 1, 2 and 3 transforms). The non-speech MLLR transform was dropped. The use of HMM modeling compares favorably to the use of a GMM even if PLP15N features were used for the GMM-based system. A relative gain of 8% in MDC and 15% in EER is obtained when switching from GMM to HMM (MG15 vs. MH12/1c). Using more classes results in a significant improvement in MDC but not in EER, which suggests a rotation of the DET curve. In any case, using 3 phonetic classes (MH12/2t) is found to be optimal in terms of EER and it obtains a relative gain of 8% MDC versus the one-class system (MH12/1c) at the same time.

Regarding CMLLR vs MLLR techniques, it seems advantageous to use CMLLR in GMM-based systems, obtaining an improvement of almost 4% both in EER and MDC terms (CG15 vs. MG15). Besides, multiple iterations of SAT

System	minDCF	EER (%)
MG15	.0413	10.15
MH12/1c	.0377	8.61
MH12/1t	.0362	8.86
MH12/2t	.0344	8.40
MH12/3t	.0334	9.56

Table 3: MDC and EER SRE 2005 results for several MLLR-SVM systems. Systems are described by MLLRTransform-Type/ModelType/FeatureSetUp/NumberOfPhoneticClasses parameters in the MLLR computation (See Table 1 for system naming convention).

training in the CMLLR system can further increase this gain. Interestingly, it seems that MLLR is better suited for HMM systems. Relative gains of 10% MDC and 13% EER are found for one-class MLLR (CH12 vs. MH12/1c), going up to 18% MDC when 3 phonetic classes (2 transforms) are used. All MLLR/HMM systems outperform the best performing CMLLR system (CG15/S).

We next assessed performance of CMLLR and MLLR systems in combination with two standard cepstral systems, MFCC-GMM (a) and MFCC-SVM (b). CG15/S-SVM (c) and MH12/2t-SVM (c') were chosen as the best performing CMLLR and MLLR system candidates. Figure 2(a) shows DET curves for all individual systems. The MLLR approach clearly outperforms CMLLR as well as MFCC-GMM, especially near the MDC. We set the baseline to be the combination of MFCC systems, (a+b). Table 4 shows results for the individual systems, the baseline combination, the three-way combination systems (a+b+c), (a+b+c') and an all-combination system, (a+b+c+c'). When combined with the baseline, the CMLLR system (c) brings an improvement of 3% MDC and 5% EER, whereas the MLLR system (d) obtains a gain of 12% and 14% respectively. Figure 2(b) shows DET curves for the combined systems. Systems behave consistently all along the DET curves. The all-combination system (a+b+c+c'), globally outperforms all other systems, although this is not clear enough on the MDC and EER operating points.

System	minDCF	EER (%)
MFCC-GMM (a)	.0330	8.65
MFCC-SVM (b)	.0279	7.23
CG15/S-SVM (c)	.0373	8.62
MH12/2t-SVM (c')	.0292	8.11
Baseline (a+b)	.0264	6.35
(a+b+c)	.0255	6.03
(a+b+c')	.0232	5.41
(a+b+c+c')	.0231	5.41

Table 4: MDC and EER SRE 2005 results for forward+backward individual, baseline and other combined systems (See Table 1 for system naming convention).

5. Conclusions

We compared several CMLLR and MLLR approches for feature extraction in speaker recognition systems and we evaluated them on the NIST SRE 2005 evaluation task. The iterative training CMLLR/GMM approach system outperformed all other CMLLR-based systems, including computing CMLLR on an ASR HMM system. CMLLR/GMM is able to use any kind of cepstral features and normalization whereas in CMLLR/HMM, the ASR system is constrained to using a certain type of features that might be less suited for speaker recognition tasks. A relative gain of 10% EER was obtained for CMLLR/GMM vs CMLLR/HMM. MLLR outperformed all CMLLR approaches specially when more than one phonetic class were used. A gain of 10% was obtained by switching from CMLLR to MLLR, both using transforms computed on an ASR HMM using speech recognition cepstral features. The best performing CMLLR and MLLR systems were combined with two standard cepstral systems, obtaining relative gains of 3% and 12% over the baseline, respectively.

6. References

[1] W. M. Campbell, D.E. Sturim, and D. A. Reynolds, "Support Vector Machines using GMM Supervectors for Speaker Verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, 2006.

[2] A. Stolcke, L. Ferrer, and S. Kajarekar, "Improvements in MLLR-Transform-Based Speaker Recognition," *Proceedings of the IEEE Speaker Odyssey*, June 2006.

[3] P. Matejka, L. Burget, P. Schwarz, O. Glembek, M. Karafiat, F. Grezl, J. Cernocky, D. A. van Leeuwen, N. Brummer, and A. Strasheim, "STBU System for the NIST 2006 Speaker recognition Evaluation," *Proceedings of IEEE Conference on Audio Speech and Signal Processing*, vol. 4, pp. 221–224, April 2007.

[4] A. Stolcke, L. Ferrer, S. Kajarekar, E. Shriberg, and A. Venkataraman, "MLLR Transforms as Features in Speaker Recognition," *Proceedings of Eurospeech*, pp. 2425–2428, September 2005.

[5] M. Ferràs, C. C. Leung, C. Barras, and J-L Gauvain, "Constrained MLLR for Speaker Recognition," *Proceedings of IEEE Conference on Audio Speech and Signal Processing*, April 2007.

[6] H. Jin, S. Matsoukas, R. Schwartz, and F. Kubala, "Fast Robust Inverse Transform SAT and Multi-stage Adaptation," *Proceedings DARPA Broadcast News Transcription and Understanding Workshop*, pp. 105–109, February 1998.

[7] C. J. Leggetter and P. C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[8] M. J. F. Gales and P. C. Woodland, "Mean and Variance Adaptation within the MLLR Framework," *Computer Speech and Language*, vol. 10(4), pp. 249–264, October 1996.

[9] V. V. Digalakis, D. Rtischev, and L. G. Neumeyer, "Speaker Adaptation Using Constrained Estimation of

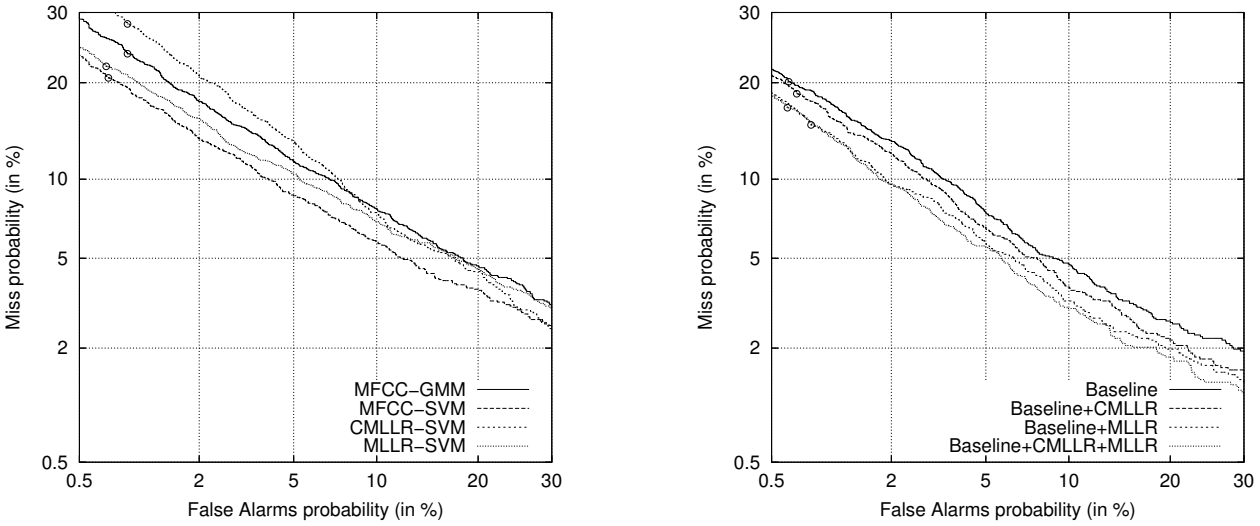


Figure 2: DET curve for the forward individual systems (left, a): MFCC-GMM, MFCC-SVM, CG15/S-SVM and MH12/2t-SVM, and for the baseline and combination systems (right, b). MDC operating points are shown as dots.

Gaussian Mixtures," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 357–366, September 1995.

[10] 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.

[11] D. A. Reynolds, "Channel Robust Speaker Verification via Feature Mapping," *Proceedings of the IEEE ICASSP*, pp. 53–56, 2003.

[12] J. Pelecanos and S. Sridharan, "Feature Warping for Speaker Verification," *Proceedings of IEEE Speaker Odyssey*, 2001.

[13] R. Prasad, S. Matsoukas, C.-L. Kao, J. Ma, D.-X. Xu, T. Colthrust, O. Kimball, R. Schwartz, J.-L. Gauvain, L. Lamel, H. Schwenk, G. Adda, and F. Lefevre, "The 2004 BBN/LIMSI 20xRT English Conversational Telephone Speech Recognition System," *Proceedings of Interspeech*, 2005.

[14] C. Barras and J. L. Gauvain, "Feature and score normalization for speaker verification of cellular data," in *ICASSP*, Hong Kong, April 2003, pp. II–49–52.

[15] C. H. Lee and J. L. Gauvain, "Speaker Adaptation Based on MAP Estimation of HMM Parameters," *Proceedings of IEEE Conference on Audio Speech and Signal Processing*, vol. 2, pp. 558–561, 1993.

[16] P. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score Normalization for Text-Independent Speaker Verification Systems," *Digital Signal Processing*, vol. 10, pp. 42–54, 2000.

[17] W. M. Campbell, "Generalized Linear Discriminant Sequence Kernels for Speaker Recognition," *Proceedings of IEEE Conference on Audio Speech and Signal Processing*, 2002.

[18] B. Scholkopf, A. Smola, and K. R. Muller, "Kernel Principal Component Analysis," *Advances in Kernel Methods-Support Vector Learning*, 1999.

[19] N. Brummer, "The Spescom DataVoice and University of Stellenbosch NIST SRE 2005 System," *NIST Speaker Recognition Workshop*, June 2005.