Constrained MLLR 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) are two widely-used techniques for speaker adaptation in large-vocabulary speech recognition systems. Recently, using MLLR transforms as features for speaker recognition tasks has been proposed, achieving performance comparable to that obtained with cepstral features. This paper describes a new feature extraction technique for speaker recognition based on CMLLR speaker adaptation which avoids the use of transcripts. Modeling is carried out through Support Vector Machines (SVM). Results on the NIST Speaker Recognition Evaluation 2005 dataset are provided as well as in combination with two cepstral approaches such as MFCC-GMM and MFCC-SVM, for which system performance is improved by a 10% in Equal Error Rate relative terms.

Index Terms— speaker verification, CMLLR, GMM, SVM

1. INTRODUCTION

The success of current state-of-the-art speaker recognition systems is still mainly based on the use of short-term acoustic cepstral features. Modeling is typically accomplished by means of Gaussian Mixture Models (GMM) or discriminative approaches such as Support Vector Machines (SVM), obtaining similar performance. Since these acoustic features only cover a few tens of milliseconds of context, longer time-span interaction and higher-level linguistic cues are ignored. Some approaches have focused on characterizing speaker speaking style to overcome this limitation. These typically go further up the acoustic level and make use of phonetic-level or word-level units. However, they rely on Automatic Speech Recognition (ASR) system output accuracy.

Cepstral features result from the interaction among several sources of information such as message, acoustic context, channel or speaker, being the latter one of the factors exhibiting the lowest variability [1]. Therefore, compensating channel effects and minimizing text-dependency become an important concern. Several channel and session compensation techniques, such as Feature Mapping [2], Factor Analysis [3] or Nuisance Attribute Projection (NAP) [4] have been successfully applied in GMM or SVM modeling. Score normalization [5] can also alleviate some of these harmful effects. As for message normalization, phone-conditioned and word-specific cepstral models have been proposed.

In a different direction, [6] proposes modeling speakers using Maximum-Likelihood Linear Regression (MLLR) speaker adaptation transforms estimated using a large-vocabulary speech recognition system. The primary goal here is to capture the speaker-independent to speaker-dependent difference (in the form of an affine transformation), with the hope of normalizing text-dependency out. Nonetheless, it is again an approach which is dependent on ASR output accuracy and it is furthermore language-dependent.

Inspired by this latter approach, we present a technique aiming at getting the benefits of MLLR speaker modeling while avoiding the need for transcripts. The key idea is to consider the property in Constrained MLLR (CMLLR) that allows to apply the estimated transform in the feature domain. When a single-class CMLLR transform is estimated, Speaker Adaptive Training (SAT) [7] can be applied in order to get a more speaker-dependent feature data set.

This paper is organized as follows: Section 2 introduces MLLR and CMLLR and details the CMLLR feature extraction procedure used in our approach. Section 3 describes all the components of the speaker verification system as well as the evaluation task. Individual results for the CMLLR-SVM system as well as fusion results on the NIST 2005 Speaker Recognition Evaluation are provided and discussed in Section 4. Conclusions are presented in Section 5.

2. CMLLR IN SPEAKER RECOGNITION

CMLLR Background

Maximum-Likelihood Linear Regression [8, 9] is an adaptation technique by which means, and optionally covariance matrices, of a HMM model are transformed by an affine transformation aiming at maximizing the likelihood function given new adaptation data and the model, as

$$\hat{\mu} = \mathbf{A}\mu + \mathbf{b} \tag{1}$$

$$\hat{\mathbf{\Sigma}} = \mathbf{H}\mathbf{\Sigma}\mathbf{H}^T \tag{2}$$

where μ is a mean vector in the model, $\mathbf{\Sigma}$, its covariance matrix, $\hat{\mu}$ and $\hat{\mathbf{\Sigma}}$ the adapted mean and covariance matrix, respectively. ($\mathbf{A}$,$\mathbf{b}$) is the affine transformation for mean adaptation, and $\mathbf{H}$, the transformation matrix for covariance adaptation. To find the optimal parameters, Expectation Maximization (EM) is typically used [9] in two steps, by estimating the covariance transformation $\mathbf{H}$ after the mean transform given by $\mathbf{A}$ and $\mathbf{b}$.

A variant of this approach, Constrained MLLR (CMLLR) [10], forces the transformation to be the same for both $\hat{\mu}$ and $\hat{\mathbf{\Sigma}}$ as

$$\hat{\mu} = \mathbf{A_c}\mu - \mathbf{b_c} \tag{3}$$

$$\hat{\mathbf{\Sigma}} = \mathbf{A_c}\mathbf{\Sigma}\mathbf{A_c}^T \tag{4}$$

The estimation of the transform here is also achieved by means of iterative optimization, typically EM.

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

CMLLR allows the transform to be applied at the feature level as

$$\hat{\mathbf{o}}_\mathbf{t} = {\mathbf{A}_\mathbf{c}}^{-1}\mathbf{o}_\mathbf{t} + {\mathbf{A}_\mathbf{c}}^{-1}\mathbf{b}_\mathbf{c} \quad (5)$$

where $\mathbf{o_t}$ is the observation vector at time t. This can be particularly useful in SAT which is used in the feature extraction technique presented in the next section.

CMLLR Feature Extraction

(C)MLLR can be used in speaker recognition systems to extract features that are more specifically focused on speaker-related characteristics than standard spectral envelope features. [6] proposes using a large-vocabulary speech recognition system to derive several class-dependent MLLR transforms the coefficients of which are later stacked vector-wise and their concatenation used as a feature vector. Depending on the number of phonetic classes over which the MLLR transforms are estimated on a HMM model, finer or coarser text normalization can be achieved.

We propose a slightly different approach which consists of two stages. As a first step, a GMM/UBM model is built upon background speaker cepstral features. Next, CMLLR transforms are estimated for each speaker of interest by using this UBM and are eventually rearranged as vectors to be modeled later.

To build the UBM model an iterative approach is followed. In order not to be dependent on a speech recognition system we use a GMM, instead of an HMM, which is trained on background speaker cepstral features. This greatly simplifies the training procedure. Only one CMLLR transform is estimated per speaker. At this point, the transforms can be applied onto the same background speaker features that were used to estimate the transforms, which results in an improved feature set regarding speaker-dependency. The UBM model can then be re-estimated by using these new features. This process can be iterated several times until a convergence criterion is accomplished. Fig. 1 illustrates the GMM/UBM estimation process.

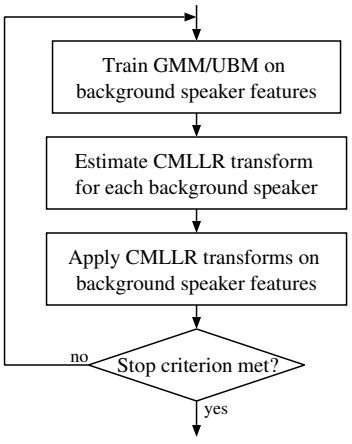

Figure 1: Block diagram for GMM/UBM re-estimation.

As a second step, last iteration's UBM model is used to estimate CMLLR transforms for each speaker for which we want to perform feature extraction. Each of the transform matrices is then stacked either column-wise or row-wise, optionally adding the offset vector $\mathbf{b}$. This results in one high-dimensional feature per speaker which is specially well-suited for SVM modeling.

The main advantage of this technique over [6] is that the training procedure is not transcript-dependent or language-dependent while still capturing differences between speaker-independent and speaker-dependent acoustic features. On the other side, since a GMM is used to estimate the transform the resulting transform is less precise and probably more dependent on the message.

A similar procedure, which shares certain similarities with the GMM Supervector (GSV) approach [11], can be also derived for Maximum A Posteriori (MAP) adaptation [12]. Here, a GMM/UBM model and a MAP-adapted model for the speaker of interest are considered. Speaker-related features are obtained by taking the difference of mean supervectors of the two models instead of just the MAP-adapted mean supervector. This procedure is analogous to the CMLLR approach although correlation between feature components is not exploited.

3. EXPERIMENTAL SETUP

Task and evaluation data

Speaker verification experiments are conducted on conversational telephone speech. The data is provided for the one-conversation two-channel condition task of the NIST 2005 evaluation[1]. Given a speech segment in a around 5-minute long conversation, the goal is to decide whether this segment is spoken by the target speaker or not. For each target model (274 male / 372 female), a speech segment in an around 5-minute long conversation is available for model training. Overall, 2429 test segments (1074 male and 1355 female) need to be scored against roughly 10 gender-matching impostors and against the true speaker. The gender of each target speaker is known.

Cepstral Feature Extraction

All systems share the same cepstral feature extraction setup. Features are extracted from the speech signal every 10ms using a 30ms window, and estimated on a 0-3.8kHz bandwidth. Feature vectors consist of 15 MEL-PLP cepstrum coefficients, 15 Δ coefficients plus Δ energy, and 15 $\Delta\Delta$ coefficients plus $\Delta\Delta$ energy (47-D features). Speech Activity Detection (SAD) is performed using the voicing level extracted using the Snack Sound Toolkit[2]. Speech frames with invalid voicing values are dropped. Channel compensation for GSM, CDMA, TDMA, landline-carbon and landline-electret data is performed for both genders using feature mapping [2]. Speech segments from test speakers from NIST SRE 1997 to 2002 evaluations (24769 speech segments) are chosen for model training. Around 6-hour speech data is used to train each gender-dependent channel model. After feature mapping, feature warping [13] is performed over a sliding window of about 3 seconds to reshape the cepstral histogram into a Gaussian distribution.

MFCC-GMM system

The MFCC-GMM system [14] is based on Gaussian Mixture Models (GMM) with diagonal covariance matrices trained using MAP adaptation [12]. For speaker modeling, GMMs are trained by 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 is a 1536-mixture

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

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

GMM, built by merging three GMMs, each with 512 Gaussians trained on cellular, landline-electret and landline-carbon data. Around 60-hour speech data is used to train each gender-dependent channel-specific 512-mixture GMM. The training data is chosen from target speakers in NIST SRE 97,98,99,00,01 and 03 evaluations and test speakers in NIST SRE03 evaluation (for a total of 9041 speech segments). Score normalization is performed using T-norm [5] on 500 speech segments (250 males and 250 females) from the Fisher corpus[3]. Speech segments in this corpus are 10-minute long telephone conversation excerpts. The first 5 minutes of each segment are extracted for score normalization.

MFCC-SVM system

The MFCC-SVM system is based on SVM modeling and several steps of feature extraction that expand the discriminative power of the base cepstral features. Polynomial feature extraction expands the MEL-PLP features into high-dimensional feature vectors through a third order monomial expansion for each frame. The resulting features are variance normalized and averaged over the whole segment to obtain a single 20824-D vector. The dimension of this speaker-specific vector is reduced via Kernel Principal Component Analysis (KPCA) [15] using a 2nd order cumulative homogeneous polynomial kernel, resulting in one 3197-D feature vector per speaker. An affine transform maps each feature component into the range [-1/3197,1/3197] so that only normalized dot products are processed by the SVM. The minimum and maximum values are taken from the impostor speaker set, which is chosen from target speakers in NIST SRE 99,00,01,02 and 04 evaluations (3198 speakers). SVM training is performed with a linear kernel using SVMTorch [4] from IDIAP.

CMLLR-SVM system

The CMLLR-SVM system uses the feature extraction scheme described in Section 2. The impostor speakers are used as background speaker set, and split into male and female genders to train two gender-dependent GMM/UBM models. The number of iterations for GMM/UBM estimation was fixed to 2, which exhibited the best performance on the evaluation. The CMLLR transforms result in 2256-D (47x47 + 47, including $\mathbf{b}$) feature vectors, after stacking their coefficients. These are min-max normalized in the range [-1/2256,1/2256] and modeled exactly in the same way as in the MFCC-SVM system.

Score-level fusion

Each of the MFCC-GMM, the MFCC-SVM and CMLLR-SVM systems consists of a forward and a backward sub-systems [16]. The forward sub-systems use a conventional approach in which test speech is compared to the speaker models trained on training speech. In the backward sub-systems, training speech is compared to the speaker models trained on test speech. Therefore, 6 scores are fused for the all-combination system.

Three-fold cross-validation is adopted in our performance evaluation. Here, the evaluation data is split into three independent balanced sets. Each set of scores is zero-mean and unit-variance normalized based on the statistics of the other two sets. Finally, an uniform weighting average score is computed for each trial.

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

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

Performance Measure

A speaker detection system is subject to two kinds of errors, i.e. missed detections and false alarms. The primary performance measure for the NIST speaker detection task is the Detection Cost Function (DCF) defined as a weighted sum of both error probabilities, the normalized cost taking the following form $C_{Norm} = P_{Miss} + 9.9 \times P_{FalseAlarm}$. For all results, we report the Minimal DCF (MDC) value obtained a posteriori for the best possible detection threshold. However, this operating point favors false alarms, so the Equal Error Rate (EER) is also provided as an alternative performance measure. We use Detection Error Tradeoff (DET) curves as well to evaluate system behaviour in the full range of operating points.

4. RESULTS

The behaviour of the CMLLR-SVM system is first assessed as a function of the number of iterations in the GMM/UBM model re-estimation. Table 1 shows MDC and EER results for the first five iterations. MDC exhibits a decreasing trend as the number of iterations grows. EER follows the same overall behaviour despite some fluctuation on the score. It is clear that most of the performance improvement is achieved at the second iteration (5% in MDC and 8% in EER in relative terms). Further iterations obtain just a marginal gain in performance at the expense of much higher computational cost. This makes CMLLR-SVM 2 it. the candidate system to be fused with the other individual systems.

System	MDC (x100)	EER (%)
CMLLR-SVM 1 it.	0.395	8.97
CMLLR-SVM 2 it.	0.372	8.19
CMLLR-SVM 3 it.	0.372	8.31
CMLLR-SVM 4 it.	0.365	8.23
CMLLR-SVM 5 it.	0.369	8.10

Table 1: MDC and EER for several iterations of UBM re-estimation in the forward CMLLR-SVM system.

Table 2 shows forward+backward fusion results for each individual system, MFCC-GMM (a), MFCC-SVM (b) and CMLLR-SVM (c), and the baseline system (a+b) and all-combination system (a+b+c). CMLLR-SVM is competitive with the other individual systems (a and b) in terms of EER. Both MFCC-GMM and MFCC-SVM significantly outperform CMLLR-SVM in MDC, though. This trend is confirmed after fusion of all individual systems. Including CMLLR-SVM in the fusion brings about a 10% relative improvement over the baseline in EER, but leaves MDC at the same level. Fig. 2(a) shows DET curves for the individual systems. MFCC-SVM outperforms the two other systems, and CMLLR-SVM and MFCC-GMM complement each other, depending on the DET curve region. Fig. 2(b) shows DET curves for the baseline and all-combination systems. The all-combination system consistently outperforms the baseline system along the whole curve. This improvement is small at the MDC operating point and gets larger for lower miss probability values, for instance, a 10% relative improvement in EER.

5. CONCLUSIONS

We presented a new feature extraction approach for speaker recognition based on the CMLLR speaker adaptation technique that doesn't depend on ASR transcripts. We showed that a competitive system can be built by using CMLLR transforms as features and

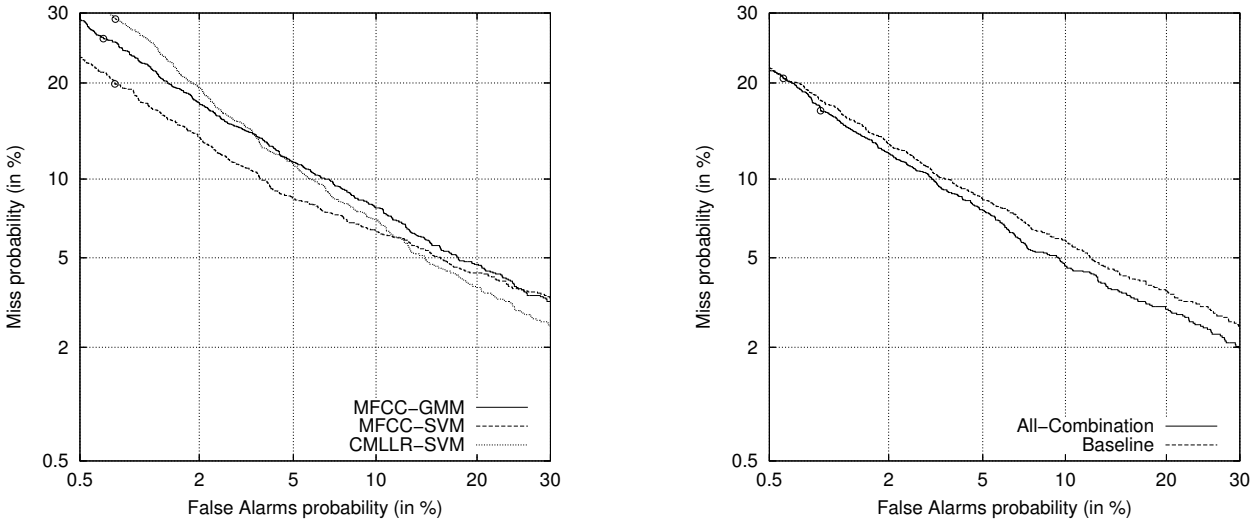

Figure 2: DET curve for the individual systems (left, a): MFCC-GMM, MFCC-SVM and CMLLR-SVM and for the baseline and all-combination systems (right, b). MDC operating points are shown as dots.

System	MDC (x100)	EER (%)
MFCC-GMM (a)	0.330	8.61
MFCC-SVM (b)	0.277	7.41
CMLLR-SVM (c)	0.370	8.15
Baseline (a+b)	0.266	7.11
All-combination (a+b+c)	0.260	6.40

Table 2: MDC and EER for the individual, baseline and all-combination systems.

SVM modeling. When this approach is combined with a MFCC-GMM and a MFCC-SVM systems performance is significantly improved. Despite the simplicity of the fusion method, a 10% relative improvement in EER was achieved.

REFERENCES

[1] S. Kajarekar, N. Malayath, and H. Hermansky, "Analysis of Speaker and Channel Variability in Speech," *Proceedings of the IEEE Workshop on Speech Recognition and Understanding*, December 1999.

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

[3] P. Kenny and P. Dumouchel, "Experiments in Speaker Verification using Factor Analysis Likelihood Ratios," *Proceedings of the IEEE Speaker Odyssey*, June 2004.

[4] A. Solomonoff, C. Quillen, and W. M. Campbell, "Channel Compensation for SVM Speaker Recognition," *Proceedings of the IEEE Speaker Odyssey*, pp. 57–62, 2004.

[5] 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.

[6] 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.

[7] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proc. ICSLP '96*, Philadelphia, PA, 1996, vol. 2, pp. 1137–1140.

[8] 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.

[9] 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.

[10] V. V. Digalakis, D. Rtischev, and L. G. Neumeyer, "Speaker Adaptation Using Constrained Estimation of Gaussian Mixtures," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 357–366, September 1995.

[11] 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.

[12] 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.

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

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

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