

Incorporating MAP Estimation and Covariance Transform for SVM based Speaker Recognition

Cheung-Chi Leung¹, Donglai Zhu¹, Kong-Aik Lee¹, Bin Ma¹ and Haizhou Li^{1,2}

¹ Human Language Technology Department,

Institute for Infocomm Research, A*STAR, Singapore 138632

² Department of Computer Science and Statistics, University of Eastern Finland, Finland

{ccleung, dzhu, kalee, mabin, hli}@i2r.a-star.edu.sg

Abstract

In this paper, we apply Constrained Maximum *a Posteriori* Linear Regression (CMAPLR) transformation on Universal Background Model (UBM) when characterizing each speaker with a supervector. We incorporate the covariance transformation parameters into the supervector in addition to the mean transformation parameters. Maximum Likelihood Linear Regression (MLLR) covariance transformation is adopted. The auxiliary function maximization involved in Maximum Likelihood (ML) and Maximum *a Posteriori* (MAP) estimation is also presented. Our experiment on the 2006 NIST Speaker Recognition Evaluation (SRE) corpus shows that the two proposed techniques provide substantial performance improvement.

Index Terms: MLLR, MAPLR, speaker adaptation, speaker recognition

1. Introduction

Recently, the use of Gaussian supervector (GSV) [1] and Maximum Likelihood Linear Regression (MLLR) supervector [2, 3, 4] have become well-known techniques in Support Vector Machine (SVM) based speaker recognition. The latter uses the concatenation of transformation parameters from MLLR, which are estimated using either Universal Background Model (UBM) [4] or Hidden Markov Model (HMM) in a large vocabulary continuous speech recognition (LVCSR) system [2, 3]. In this paper, we are interested in the transformation parameters derived from UBM.

MLLR was originally used for speaker adaptation in automatic speech recognition (ASR) [5]. In order to obtain more robust speaker adaptation, Maximum *a posteriori* Linear Regression (MAPLR) [6, 7, 8], which incorporates prior knowledge into the parameter estimation, has been proposed. Given its superior performance over MLLR transform in ASR tasks, the use of MAPLR transforms as features in speaker recognition has been proposed recently [9].

In MLLR and MAPLR, the Gaussian components are grouped into regression classes according to acoustic or phonetic similarity. For mean adaptation, the mean vectors in regression class *c* are linearly transformed using a square transformation matrix A_c and a bias vector b_c as follows

$$\hat{\mu} = A_c \mu + b_c = W_c \xi \quad (1)$$

where μ and $\hat{\mu}$ are respectively the original and transformed Gaussian mean vectors, $\xi = [1 \ \mu^T]^T$ is the extended mean vector, and $W_c = [b_c^T \ A_c^T]^T$ is the extended transform. For covariance adaptation, Gaussian covariance matrices are linearly transformed as follows

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

where Σ and $\hat{\Sigma}$ are respectively the original and transformed Gaussian covariance matrices, and H_c are the transformation matrix of class *c*. In constrained MLLR (CMLLR) and constrained MAPLR (CMAPLR) adaptation [10], mean and covariance parameters share the same transformation matrix (i.e. $A_c = H_c$). Single class/global transform and full transformation matrices are considered in this paper.

MLLR and MAPLR differ in the optimization criterion used in their parameter estimation as follows

$$MLLR: \quad \hat{W}_{ML} = \arg \max_w p(X | \Lambda, W) \quad (3)$$

$$MAPLR: \quad \hat{W}_{MAP} = \arg \max_w p(X | \Lambda, W) p(W) \quad (4)$$

where Λ is the original Gaussian mixture components. In MLLR, maximum likelihood (ML) estimation depends only on the adaptation data X and does not impose any constraint on the possible parameters in the transform W . But in MAPLR, Maximum *a Posteriori* (MAP) estimation depends not only on the adaptation data but also the prior density $p(W)$, which constrains the possible values of W .

This paper presents two different approaches in the estimation of transformation parameters. In MAPLR supervector estimation, inspired by the works of Ferras *et al* [4] and Zhu *et al* [11], we firstly propose to use the UBM which are re-estimated by CMAPLR transformed cepstral features. The purpose of the speaker adaptive training (SAT)-like scheme is to make UBM more speaker-independent. Secondly, we investigate whether incorporating covariance transformation parameters in mean-transform-related supervectors provides a significant improvement. Following the general practice in ASR tasks, mean or constrained adaptation is used in speaker recognition. To the best of our knowledge, using linear covariance transformation parameters to construct the supervector has not yet been studied in speaker recognition. However, since the covariance transform, which is estimated independently from the mean transform, can also increase the likelihood of the adaptation data given

the updated model, we believe it may contain useful speaker information and may be complementary to the mean transform from the point of view of a mixture of experts.

The remaining of this paper is organized as follows: Section 2 and 3 present MAPLR mean and CMAPLR transformations respectively, together with the relation with their corresponding MLLR solutions. Section 4 describes the integration of covariance transforms into the linear transformation supervectors. Section 5 briefly presents the supervector and SVM modeling. Section 6 describes the experimental setup and results respectively. Finally we conclude in Section 7.

2. Maximum a Posteriori Linear Regression Mean Transformation

In MAPLR mean adaptation [6], transform parameters are estimated using expectation maximization (EM) algorithm. Ignoring all the terms independent of W , the auxiliary function $Q_{MAP}(W|\bar{W})$ (applicable to constrained and covariance adaptation as well) can be defined as

$$Q_{MAP}(W|\bar{W}) = Q_{ML}(W|\bar{W}) + \log p(W) \quad (5)$$

$$Q_{ML}(W|\bar{W}) = -\frac{1}{2} \sum_{t,m} \gamma_m(t) \left[\log \left(\hat{\Sigma}^{(m)} \right) + (x(t) - \hat{\mu}^{(m)})^T \hat{\Sigma}^{(m)-1} (x(t) - \hat{\mu}^{(m)}) \right] \quad (6)$$

where $Q_{ML}(W|\bar{W})$ is the auxiliary function used in MLLR adaptation [5], $\gamma_m(t)$ is the posterior probability of being in Gaussian m at time t , $x(t)$ is the adaptation data of the hypothesized speaker at time t , $\hat{\mu}^{(m)}$ and $\hat{\Sigma}^{(m)}$ are the transformed mean vector and covariance matrix for Gaussian component m .

2.1. Selection and Estimation of Prior Distribution

A matrix version of multivariate normal distribution is usually chosen for the prior distribution of $p(W)$ as

$$p(W) \propto |E|^{-(D+1)/2} |\Phi|^{-(D)/2} \cdot \exp \left[-\frac{1}{2} \text{tr} (W-U)^T E^{-1} (W-U) \Phi^{-1} \right] \quad (7)$$

where $\{U, E, \Phi\}$ is the set of hyperparameters, U is a $D \times (D+1)$ matrix, $E = \text{diag}\{\varepsilon_1 \dots \varepsilon_D\}$ is a diagonal $D \times D$ matrix, Φ is a $(D+1) \times (D+1)$ matrix and $E, \Phi \geq 0$. Assuming Φ is an identity matrix, other hyperparameters can be estimated as

$$\hat{U} = \frac{1}{N} \sum_{i=1}^N W_i \quad (8)$$

$$\hat{E} = \frac{1}{N} \sum_{i=1}^N (W_i - \hat{U}) \Phi^{-1} (W_i - \hat{U})^T \quad (9)$$

where $\{W_1, \dots, W_N\}$ are the MLLR transforms obtained from a set of training data from different speakers. These training data are chosen from the same data used for UBM training in our experiments. In maximizing the auxiliary function, since E is assumed to be diagonal, differentiating $\log p(W)$ with

respect to each row of W , notated as w_i , leads to an extra term (ignoring all the terms independent of W , U and E) as

$$\frac{\partial \log p(w_i)}{\partial w_i} = -\frac{w_i}{\varepsilon_i} + \frac{u_i}{\varepsilon_i} \quad (10)$$

2.2. Maximization of the Auxiliary Function

Differentiating (5) with respect to each row of W and equating to zero, and assuming diagonal covariance matrices in the original Gaussian components, the extended transform W can be obtained by solving a system of $D \times (D+1)$ linear equations. Moreover, the solution differs from that in the corresponding MLLR case only in the additional terms related to the prior distribution.

3. Constrained Maximum a Posteriori Linear Regression Transformation

In the MAPLR based speaker recognition system in [9], speaker-independent (SI) UBM is used. Similar to using CMLLR adaptation for UBM re-estimation in [4], we propose to re-estimate the UBM by CMAPLR adaptation. This makes the cepstral features in each training speech utterance and the resultant UBM more speaker-independent. One iteration of UBM re-estimation is used in our experiment.

CMLLR and CMAPLR are the techniques originally designed for speaker adaptation in ASR. In CMLLR adaptation, when a transform matrix A is applied to both Gaussian mean and covariance, the auxiliary function (6) becomes

$$Q_{ML}(W|\bar{W}) = -\frac{1}{2} \sum_{t,m} \gamma_m(t) \left[\log \left(\Sigma^{(m)} \right) - \log \left(|A|^2 \right) + (\hat{o}(t) - \mu^{(m)})^T \Sigma^{(m)-1} (\hat{o}(t) - \mu^{(m)}) \right] \quad (11)$$

where

$$\hat{o}(t) = A o(t) + b = W \xi(t) \quad (12)$$

and $W = \begin{bmatrix} b^T & A^T \end{bmatrix}^T$ is the extended transformation matrix, $\xi(t) = \begin{bmatrix} 1 & o(t)^T \end{bmatrix}^T$ is the extended observation vector at time t , $\gamma_m(t)$ is again the posterior probability of being in Gaussian m at time t , $\mu^{(m)}$ and $\Sigma^{(m)}$ are the mean vector and the covariance matrix for Gaussian component m . The iterative solution for the transform is given by

$$w_i = (\alpha p_i + k^{(i)}) G^{(i-1)} \quad (13)$$

where $p_i = [0 \text{ cof}(A_{i1}) \dots \text{cof}(A_{iD})]$ is the extended cofactor vector,

$$G^{(i)} = \sum_m \frac{1}{\sigma_i^{(m)2}} \sum_t \gamma_m(t) \xi(t) \xi(t)^T, \quad (14)$$

$$k^{(i)} = \sum_m \frac{1}{\sigma_i^{(m)2}} \mu_i^{(m)} \sum_t \gamma_m(t) \xi(t)^T, \quad (15)$$

and α is solved from the following quadratic equation

$$\alpha^2 p_i^T G^{(i-1)} p_i + \alpha p_i^T G^{(i-1)} k^{(i)} - \sum_m \sum_t \gamma_m(t) = 0 \quad (16)$$

In CMAPLR adaptation [10] with the same choice of prior distribution as in Section 2.1, substituting (11) in (5), and using the same row-by-row differentiation technique and substituting (10) in it, the iterative solution of the transform is given by

$$\begin{aligned} w_i &= (\alpha p_i + k^{(i)}) G^{(i-1)} \\ &= \left(\alpha p_i + k^{(i)} + \frac{u_i}{\varepsilon_i} \right) \left(G^{(i)} + \frac{I}{\varepsilon_i} \right)^{-1} \end{aligned} \quad (17)$$

Again, the MAP solution differs from the corresponding ML solution only in the additional terms related to the prior density.

4. Supervector Incorporating Covariance Adaptation

Chou *et al* [12] derived the solution for MAPLR covariance adaptation with the original covariance in the form of Choleski factorization. However, using a diagonal transform is assumed in this approach. This assumption greatly decreases the number of parameters in the covariance transform (from D^2 to D) for characterizing a speaker and so this approach is not considered in this paper. We propose to use similar full transformation matrices, which are obtained from MLLR covariance adaptation.

MLLR covariance adaptation [5] is originally used to adapt covariance parameters based on the mean-adapted model for ASR. The corresponding auxiliary function can be defined as

$$\begin{aligned} Q_{ML}(W | \bar{W}) &= \\ &= -\frac{1}{2} \sum_{t,m} \gamma_m(t) \left[\log \left(\left| \Sigma^{(m)} \right| \right) - \log \left(\left| A \right|^2 \right) + (A o(t))^T \Sigma^{(m-1)} (A \hat{o}(t)) \right] \end{aligned} \quad (18)$$

where A is the inverse of the transform H_c in (2). Note that in our work for speaker recognition, covariance adaptation is performed independently from mean adaptation and so no mean transformation matrix can be seen in (18). Applying the same row-by-row differentiation techniques for maximizing (18), the following equation is obtained

$$\sum_{t,m} \gamma_m(t) \frac{p_i}{p_i a_i^T} - a_i G^{(i)} = 0 \quad (19)$$

where a_i is the i^{th} row vector of A , p_i is again an extended cofactor, and $G^{(i)}$ is defined as in (14). Rearranging (19) and using the fact that a_i is a scalar multiplication of $p_i G^{(i-1)}$, the solution is given by

$$a_i = p_i G^{(i-1)} \sqrt{\frac{\sum_{t,m} \gamma_m(t)}{p_i G^{(i-1)} p_i^T}} \quad (20)$$

If MAPLR covariance adaptation is considered, the maximization leads to

$$\sum_{t,m} \gamma_m(t) \frac{p_i}{p_i a_i^T} - a_i G^{(i)} - \frac{a_i}{\varepsilon_i'} + \frac{u_i'}{\varepsilon_i'} = 0 \quad (21)$$

where u_i' and ε_i' , which are respectively the row vectors of the hyperparameters estimated in (8) and (9) where $W_i = H_i^{-1}$. Because of the extra term related to the prior distribution, a_i being a scalar multiplication of $p_i G^{(i-1)}$ does

not always hold in (21). So it is not straightforward to derive the MAPLR solution (and the initialization step in the iterative process) as for the corresponding MLLR solution.

5. Linear Transformation Supervector and SVM Modeling

For each speech utterance, after its transformation parameters are estimated, they are concatenated to form a supervector. Following the approach in [3], transforms from matched-gender and opposite-gender UBMs are used in our experiments. If only MAPLR mean transforms are used, the supervector's dimension is $2D \times (D + 1)$. If MLLR covariance transforms are appended, the dimension becomes $2D \times (D + 1) + 2D^2$. An SVM is trained for each target speaker by taking the target speaker's training supervectors as positive examples, and the supervectors from the background set as negative examples. Our experiments are implemented using the LIBSVM package, in which a linear inner-product kernel function is adopted [13].

Before SVM training or testing, the supervectors are normalized to equal dynamic range. In our experiments, we use rank normalization, which replaces each value in the supervectors with its rank among the background data samples on a given dimension, and then scales the rank to a value between 0 and 1.

Recently, nuisance attribute projection (NAP) has been successfully used in the SVM framework to factor out the nuisance subspace from the original supervector space [14]. The subspace, aiming to model nuisance effects in speech, is trained on a set of training data recorded from many different speakers each speaking multiple speech segments in different channels. The NAP also exhibits remarkable effectiveness in our experiment. Finally, the output SVM scores are normalized with Tnorm which further compensates for the nuisance effects [15].

6. Experimental Setup and Results

We conducted our experiments on the core task of the 2006 NIST Speaker Recognition Evaluation (SRE). Given a speech segment extracted from a 5-minute conversation, the goal is to decide whether this segment is spoken by the target speaker or not. Equal error rate (EER) is used to evaluate system performance. Detection Error Tradeoff (DET) curve showing system behavior in the full range of operating points is also used.

In cepstral feature extraction, each speech utterance is converted to a sequence of 36-dimensional feature vectors including 12 MFCC coefficients and their first and second order derivatives, which are then filtered by a RASTA filter. An energy-based voice activity detection (VAD) process is then used to remove non-speech frames. Finally, the feature vectors are processed by mean and variance normalization.

The data for training the UBMs (512 mixtures and gender dependent), and the background data for training the SVM and NAP projection matrix comprises 383 2.5-minute speech utterances, recorded by 310 speakers, from the 2004 NIST SRE corpus. The NAP matrix has co-rank of 40. The cohort model for Tnorm score normalization is chosen from the 1-conversation training data in the 2005 NIST SRE corpus.

Table 1 compares three systems with different settings in UBM estimation and transformation supervectors. It is shown

that NAP and Tnorm consistently provide performance improvement in the three systems. Using CMAPLR adaptation in UBM re-estimation (system B) and adding MLLR covariance transform parameters (system C) provide consistent performance improvement in different normalization steps. Applying the two proposed techniques provides 25% relative EER reduction in total. It is observed that adding MLLR covariance transform parameters provide more performance gain than applying UBM re-estimation with CMAPLR adaptation. Figure 1 shows the DET curves of the 3 systems listed in Table 1 after NAP and Tnorm. Figure 1 shows that the performance improvement made by the two proposed techniques is consistent along most operating points. Moreover we observed that inclusion of the transformation from the opposite-gender UBM provides consistent improvement in our systems (approximately 1% absolute reduction in EER before NAP and Tnorm).

Table 1. Results (in EER) on the core task of 2006 NIST SRE

System ID	Supervector	UBM updated with CMAPLR	EER (%)		
			SVM	+NAP	+Tnorm
A	MAPLR-mean	No	8.80	8.10	8.00
B	MAPLR-mean	Yes	8.63	7.85	7.73
C	MAPLR-mean + MLLR-covariance	Yes	7.38	6.43	6.03

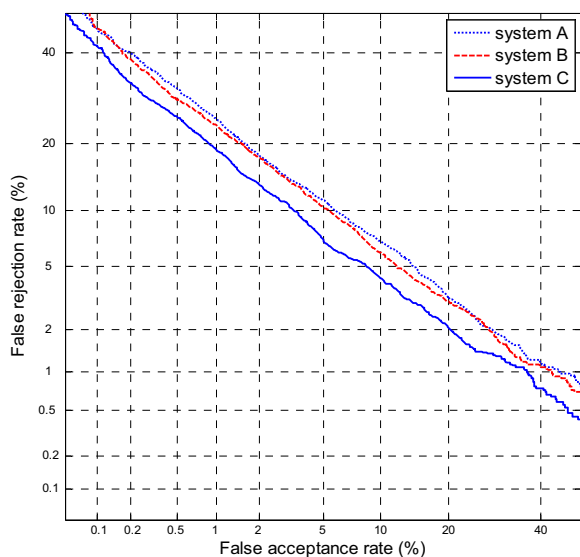


Figure 1. DET curves showing the performance of three systems on the core task of 2006 NIST SRE

7. Conclusions

In this paper we propose to improve an SVM based speaker recognition system using MAPLR mean transformation

supervectors by: 1) applying CMAPLR adaptation in UBM re-estimation and 2) incorporating covariance transformation parameters in the supervectors.

These two techniques consistently improve the system performance in consecutive normalization steps in our experiments. Auxiliary function maximization in MLLR and MAPLR covariance adaptation is studied. When utilizing information in covariance transformation, MLLR adaptation is employed, but an obvious performance improvement is still obtained. In our future work, multi-class adaptation will be investigated.

8. References

- [1] Campbell W. M., Sturim D. E., Reynolds D. A. and Solomonoff A., "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in Proc. ICASSP, 2006, pp. 97-100.
- [2] Stolcke A., Ferrer L., Kajarekar S., Shriberg E. and Vekataraman A., "MLLR transforms as features in speaker recognition," in Proc. Eurospeech, 2005, pp. 2425-2428.
- [3] Stolcke A., Kajarekar S. S., Ferrer L., and Shriberg E., "Speaker recognition with session variability normalization based on MLLR adaptation transforms", IEEE Trans. Audio, Speech, and Lang. Process., vol.15, pp. 1987-1998, Sep. 2007
- [4] Ferras M., Leung C. C., Barras C. and Gauvain J.-L., "Constrained MLLR for speaker recognition," in Proc. ICASSP, 2007, pp. 53-56.
- [5] Gales M. J. F., "Maximum likelihood linear transformations for HMM-based speech recognition," Computer Speech and Language, vol. 12, pp.75-98, 1998.
- [6] Chesta C., Siohan O., and Lee C.-H., "Maximum a posteriori linear regression for hidden markov model adaptation," in EUROSPEECH, 1999.
- [7] Chou W., "Maximum a posteriori linear regression with elliptically symmetric matrix variate priors," in EUROSPEECH, 1999.
- [8] Siohan O., Myrvoll T. A. and Lee C.-H., "Structural maximum a posteriori linear regression for fast HMM adaptation," Computer, Speech and Language, vol. 16, pp.5-24, 2002.
- [9] Zhang X., Wang H., Xiao X., Zhang J., Yan Y., "Maximum a posteriori linear regression for speaker recognition", in Proc. ICASSP, 2010, pp. 4542-4545.
- [10] Lei X., Hamaker J., and He X., "Robust feature space adaptation for telephony speech recognition," in Proceedings of the 9th International Conference on Spoken Language Processing (ICSLP '06), vol. 2, pp. 773-776, Pittsburgh, Pa, USA, September 2006.
- [11] Zhu D., Ma B. and Li H., "Using MAP estimation of feature transformation for speaker recognition", in Proc. Interspeech, 2008.
- [12] Chou W., and He X., "Maximum a posteriori linear regression based variance adaptation of continuous density HMMs," in EUROSPEECH, 2003.
- [13] Chang C.-C., and Lin C.-J., LIBSVM: a library for support vector machines, 2001. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [14] Solomonoff A., Campbell W. M., and Boardman I., "Advances in channel compensation for SVM speaker recognition," in Proc. ICASSP, 2005, pp. 629-632.
- [15] Auckenthaler R., Carey M., and Lloyd-Thomas H., "Score normalization for text-independent speaker verification systems," Digital Signal Process., vol. 10, pp. 42-54, Jan. 2000.