Chinese Journal of Electronics
Vol.19, No.4, Oct. 2010

Multi-Class Maximum A Posteriori Linear Regression for Speaker Verification*

ZHANG Xiang, XIAO Xiang, WANG Haipeng, ZHANG Jianping and YAN Yonghong

(*ThinkIT Speech Laboratory, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China*)

Abstract — Maximum likelihood linear regression (MLLR) transforms have proven useful for text-independent speaker recognition systems. These systems use the parameters of MLLR transforms as features for SVM modeling and classification. In this paper, we focus on calculating affine transforms based on a GMM Universal background model (UBM). Rather than estimating transforms using maximum likelihood criterion, we propose to use Maximum a posteriori linear regression (MAPLR) for feature extraction. This work is enriched by a multi-class technique, which clusters the Gaussian mixtures into regression classes and estimates a different transform for each class. The transforms of all classes are concatenated into a supervector for SVM classification. Besides, a further accuracy boost is obtained by combining supervectors derived from both female and male UBMs into a larger supervector. Experiments on a NIST 2008 SRE corpus show that the MAPLR system outperforms MLLR and the multi-class approaches can also bring significant gains.

Key words — Speaker recognition, Maximum likelihood linear regression (MLLR), Maximum a posteriori linear regression (MAPLR), Support vector machine, Supervector.

I. Introduction

Support vector machine (SVM) has become a popular and useful tool for speaker recognition. In any SVM based speaker recognition system, it is important to choose an appropriate SVM feature expansion, which maps a given utterance to a feature vector in a high-dimensional space for SVM classification.

Maximum likelihood linear regression (MLLR) is a commonly used speaker adaptation approach. The primary goal here is to capture the speaker-independent to speaker-dependent difference (in the form of an affine transform). The concatenation of the transformation parameters can be seen as a kind of mapping from the given utterance to a high-dimensional space. A system proposed in Ref.[1] first uses the MLLR transforms employed in speech recognition systems as the features for SVM based speaker recognition. However, this approach typically requires the use of large-vocabulary word recognition systems, and the computational Complexity is extraordinarily high. Another system uses Constrained MLLR (CMLLR) to adapt the means of a GMM UBM to a given utterance, and uses the entries of the transform as features for SVM classification[2]. MLLR in a GMM framework is also introduced for the task of speaker recognition in Ref.[3], which outperforms CMLLR and is useful for the system fusion. Besides, the approach in Ref.[4] groups the adaptation data based on broad phonetic classes into multiple classes in order to get multi-class transforms. This method can improve the performance of the MLLR system significantly.

In MLLR, parameters are estimated with the Maximum likelihood (ML) criterion, which is well known for its poor asymptotic properties. It may encounter numerical problems when the adaptation data is insufficient[5]. A possible solution to this problem is to introduce some constraints on the possible values of the transformation parameters. Maximum a posteriori linear regression (MAPLR)[6] is such an adaptation approach, which inserts the priori information of the transforms in the estimation process using Maximum a posteriori (MAP) as the estimation criterion to derive the transformation parameters η:

$$\hat{\eta} = \arg\max_{\eta} p(\eta|X,\lambda) = \arg\max_{\eta} p(X|\lambda,\eta)p(\eta) \qquad (1)$$

where $p(\eta)$ is the priori distribution of the parameters η, X is the adaptation features and λ represents the speaker-independent model.

MAPLR can generate transforms showing better adaptation performance than MLLR. In this study, we use MAPLR for SVM based speaker recognition, and introduce it into a GMM framework which avoids the need for transcripts. The transformation parameters are estimated based on a GMM UBM and concatenated into supervectors for SVM classification. Then, a multi-class MAPLR is proposed to improve the performance. Rather than using phonetic recognition, the proposed multi-class approach clusters the mixtures of UBM into classes based on a likelihood measure. For each class, a single transformation matrix is estimated using the mixtures within it. All the adaptation transforms of all mixture classes are concatenated into a single feature vector and modeled using SVMs.

A further improvement for the proposed multi-class

*Manuscript Received Nov. 2009; Accepted Apr. 2010.

MAPLR technique is also presented, which calculates transforms relative to multiple UBMs. In our study, we process all speakers with both male and female UBMs, and get two kinds of supervectors for each utterance, one of which is based on the male UBM, and another based on the female UBM. These two kinds of supervectors are also concatenated into a larger supervector for SVM modeling and scoring.

II. MAPLR for Speaker Recognition

1. Model transformation function

In MAPLR adaptation, the mean vectors of the Gaussian mixtures are also adapted using an affine transform as MLLR:

$$\hat{\mu}_m = A\mu_m + b = W\xi_m \tag{2}$$

where μ_m is the mean vector of the UBM, $\hat{\mu}_m$ is the adapted mean vector and ξ_m is the extended mean vector, defined as $\xi_m = (\mu_m, 1)$. Thus, the transformation parameters can be denoted by $\eta = \{A, b\} = \{W\}$.

We first assume that all the mean vectors share the same transform. Given some adaptation data of a hypothesized speaker, $X = \{x_1, x_2, \cdots, x_T\}$, the objective of the MAPLR for speaker recognition is to derive an affine transform η using a MAP estimation criterion as described by Eq.(1).

2. Definition of the auxiliary function

Commonly, the maximization of Eq.(1) cannot be carried out directly. The maximization problem is traditionally addressed by solving an auxiliary and simpler problem having the same solution, using the EM algorithm. Let $\lambda = \{\omega_m, \mu_m, \Sigma_m\}$, $m = 1, 2, \cdots, M$ denote the UBM, according to Ref.[6], the auxiliary function in a GMM framework can be defined as following:

$$\begin{aligned} Q(\eta|\bar{\eta}) &= E\{\log p(X, L|\lambda, \eta) + \log p(\eta)|X, \lambda, \bar{\eta}\} \\ &= \sum_L p(L|X, \lambda, \bar{\eta}) \log p(X|L, \lambda, \eta) p(L|\lambda, \eta) \\ &\quad + \log p(\eta) \end{aligned} \tag{3}$$

where $L = \{l_t\}$ represents the mixture sequence, and $\bar{\eta}$ is the current value of the transformation parameters. By iteratively maximizing Eq.(3) over η until a fixed point is reached, it can be shown that the obtained η also maximizes Eq.(1) locally.

The auxiliary function can be rewritten as:

$$\begin{aligned} Q(\eta|\bar{\eta}) = &\sum_L p(L|X, \lambda, \bar{\eta}) \sum_{t=1}^{T} [\log p(l_t|\lambda, \eta) \\ &+ \log p(x_t|l_t, \lambda, \eta)] + \log p(\eta) \end{aligned} \tag{4}$$

Let $l_t = m$, we have $\log p(l_t|\lambda, \eta) = \log \omega_m$, which is independent of η, and in Eq.(4), $p(L|X, \lambda, \bar{\eta})$ is also independent of η. Thus, Eq.(4) can be rewritten as:

$$Q(\eta|\bar{\eta}) = \sum_{t=1}^{T} \sum_{m=1}^{M} \gamma_t(m) \log p(x_t|\eta, \mu_m, \Sigma_m) + \log p(\eta) + \Psi \tag{5}$$

where $\gamma_t(m) = p(l_t = m|X, \lambda, \bar{\eta})$ is the posteriori probability of being on mixture m of the UBM, given the feature vector x_t. Ψ represents all the terms independent of η.

3. Estimation of the prior density $p(W)$

We choose a special case of elliptical distribution for the prior density $p(W)$, namely a matrix variate normal prior density, which can be seen as a matrix version of a multivariate normal distribution:

$$\begin{aligned} P(W) = &|R|^{-(p+1)/2} |\Phi|^{-p/2} \\ &\cdot \exp\left\{ -\frac{1}{2} tr(W - M)' R^{-1} (W - M) \Phi^{-1} \right\} \end{aligned} \tag{6}$$

where p is the feature dimension, W and M are two $p \times (p+1)$ matrices, R is a $p \times p$ matrix, Φ is $(p+1) \times (p+1)$ and $\Phi \geq 0$.

After the choice of the form of the distribution of W, a crucial issue is to get an estimation of the hyperparameters of the prior distribution. The basic principle is to first generate a set of N transformation matrices, each of which can be seen as a sample drawn from the prior distribution $p(W)$. Given this set of matrices, it is possible to derive an estimate of the hyperparameters using an ML estimation scenario.

We first select a set of training data containing N speech utterances with many different speakers and calculate the transform for each utterance using MLLR directly from the UBM. Thus, we can get a set of N transformation matrices, $\{W_1, W_2, \cdots, W_N\}$. Then we use an ML approach to derive the hyperparameters M, R, Φ. Under the assumption that Φ is the identity matrix, the hyperparameter estimates are:

$$\hat{M} = \frac{1}{N} \sum_{i=1}^{N} W_i \tag{7}$$

$$\hat{R} = \frac{1}{N} \sum_{i=1}^{N} (W_i - \hat{M}) \Phi^{-1} (W_i - \hat{M})' \tag{8}$$

4. Maximization of the auxiliary function

Given the hyperparameters of $p(W)$, we differentiate the Eq.(5) with regard to each element of W and equate the results to zero. The following systems of $p \times (p+1)$ linear equations are obtained:

$$\begin{aligned} &\sum_{k=1}^{p} \sum_{l=1}^{p+1} w_{kl} \left[\sum_{m=1}^{M} \left(\sum_{t=1}^{T} \gamma_t(m) \right) \sigma_{ik} \tilde{\mu}_l \tilde{\mu}_j \right. \\ &\left. + \frac{1}{2} r_{ki} \phi_{jl} + \frac{1}{2} r_{ik} \phi_{lj} \right] \\ =&z_{ij} \end{aligned} \tag{9}$$

where w_{ij}, σ_{ij}, r_{ij} and ϕ_{ij} are the (i, j)th components of matrices W, Σ_m, R and Φ, and $\tilde{\mu}_i$ is the ith component of μ_m. Let $x_i(t)$ denote the ith component of the feature vector $x(t)$, and m_{ij} denote the (i, j)th component of the hyperparameter matrix M, z_{ij} is defined as:

$$\begin{aligned} z_{ij} = &\sum_{k=1}^{p} \sum_{l=1}^{p+1} \left[\sum_{m=1}^{M} \left(\sum_{t=1}^{T} \gamma_i(m) x_k(t) \right) \sigma_{ik} \tilde{\mu}_j \right. \\ &\left. + \frac{1}{2} r_{ki} m_{kl} \phi_{jl} + \frac{1}{2} r_{ik} m_{kl} \phi_{lj} \right] \end{aligned} \tag{10}$$

It is worth noting that equations of MAPLR are very similar to the MLLR solution except for the additional terms related to the prior density. The matrix W can be obtained by

solving p systems of $p+1$ linear equations described by Eqs.(9) and (10).

5. Multi-class MAPLR adaptation

The Gaussian mixtures can be considered to be modeling some underlying broad phonetic sounds that characterize a person's voice. We assume that the mixtures representing acoustic classes located in similar acoustic space could be grouped into the same class. Each class has a single transformation matrix associated with it, and all the mixtures within that class share the same matrix. Thus, using multi-class technique for MAPLR adaptation allows for more freedom in adapting the GMM, since all the means are not constrained to move the same way.

In this work, the novel affinity propagation algorithm[7] is firstly used to cluster the mixtures of UBM into regression classes based on the Bhattacharyya distance. After the clustering procedure, a single transformation matrix is estimated for each class using the given utterance. We give the two-class MAPLR adaptation as an example. Let Θ_1 and Θ_2 denote the mixture index sets for class one and class two, the UBM can be redefined as:

$$g(x) = \sum_{m\in\Theta_1} \omega_m N(x;\mu_m,\Sigma_m) + \sum_{m\in\Theta_2} \omega_m N(x;\mu_m,\Sigma_m) \quad (11)$$

where $N(x;\mu_m,\Sigma_m)$ is a Gaussian mixture of UBM with mean μ_m and covariance Σ_m. Adapting the means of the UBM via two-class MAPLR to a given utterance produces a transformation matrix $W_1 = \{A_1, b_1\}$ using Eqs.(9) and (10) for mixtures assigned to class one, and $W_2 = \{A_2, b_2\}$ for mixtures assigned to class two. They can be used to adapt the means of the mixtures in the corresponding classes:

$$\hat{\mu}_m = A_1\mu_m + b_1, \quad \forall m \in \Theta_1 \quad (12)$$

$$\hat{\mu}_m = A_2\mu_m + b_2, \quad \forall m \in \Theta_2 \quad (13)$$

The transforms W_1 and W_2 characterize two kinds of speaker-independent to speaker-dependent differences, and they can complement with each other. We concatenate the parameters of the two transforms into a supervector to pursue higher accuracy performance for speaker recognition.

6. Further improvement

In multi-class MAPLR, multiple transforms are used to project the reference speaker onto the new speaker. The transforms are dependent on the UBM relative to which they are computed. In general, different UBMs are not just linear transforms of each other. The availability of different gender dependent UBMs raises the possibility of expanding the feature space by computing the transforms relative to an array of UBMs and concatenating the resulting feature vectors into a larger supervector. Thus, we can expect the corresponding sets of MAPLR transformation features to afford different, not entirely redundant "views" of observation space, and the resulting combined feature vector to yield higher accuracy.

In our study, MAPLR transforms are calculated using both male and female UBMs. We can get two kinds of transforms for each utterance, one of which is based on the male UBM, and another based on the female UBM. These two kinds of transforms for each utterance are concatenated into a supervector for SVM classification. Experiments show that a further accuracy boost is obtained when we combine the supervectors derived from the two UBMs into a larger supervector.

7. Feature extraction and SVM modeling

The MAPLR parameters from one or more transforms are concatenated into a single supervector consisting of $K \times N \times p \times (p+1)$ elements and modeled using SVMs, where K is the number of mixture classes of UBM, N is the number of UBMs, and p is the cepstral feature dimension.

Rank normalization[8] is used to normalize the supervectors to equate their dynamic ranges. Rank normalization warps the distribution to be approximately uniform, which may result in better robustness for SVM classifier. Besides, Nuisance attribute projection (NAP)[9] is applied on supervectors to project out the subspace of maximum intra-speaker variability, thus compensating inter-session variability. Our experiments are implemented using the SVMTorch with a linear inner-product kernel function. The output SVM scores are normalized with Znorm which further compensates for nuisance effects.

III. Experiments

We performed experiments on the NIST 2008 SRE corpus. For this corpus, we focused on the single-side 1 conversation train, single-side 1 conversation test, and multi-language telephone task, which is one part of the core test condition. This setup resulted in 2678 true trials and 33218 false trials. We used Equal error rate (EER) and the minimum decision cost value (minDCF) as metrics for evaluation[10].

For cepstral feature extraction, a 20-ms Hamming window with 10 ms shifts is used. Each utterance is converted into a sequence of 36-dimensional feature vectors, each consisting of 12 MFCC coefficients and their first and second derivatives. An energy-based speech detector is applied to discard vectors from low-energy frames. To mitigate channel effects, RASTA and feature warping are applied to the features.

The GMM UBM consists of 1024 mixtures, which is trained using EM with the data from the corpora: NIST 01, 02, 04, and 05. The background data is the same with UBM. A training set is selected for estimating the prior density of the transforms as well as NAP training, the data of which are from the corpora: NIST 04, 05, 06, and the Switchboard Cellar Part I. Another training set is used for Znorm score normalization, recorded by 628 female speakers and 481 male speakers from NIST 05 and 06 corpora.

Table 1. Comparison of results for MLLR and MAPLR. The upper row in each table cell is the EER (%). The lower row is the minDCF value. Rank normalization, NAP and Znorm are not used

System	Female	Male
MLLR	14.80 0.0625	12.50 0.0490
MAPLR	13.69 0.0602	9.85 0.0440

In Table 1, we compare the results of MAPLR system to MLLR system. We can see that MAPLR system produces bet-

ter performance than MLLR on both male and female data. It leads to gains of 21% on EER and 10% on minDCF on male speakers, and 7% on EER and 4% on minDCF on female speakers.

Table 2 presents the comparison of EER and minDCF values between the global single class MAPLR and the multi-class MAPLR systems on the male speakers. We can see that the multi-class technique can improve the performance of the MAPLR system significantly. The performance of the system increases with the number of Gaussian mixture classes. When the number of the Gaussian mixture classes is 8, and the multi-class MAPLR system achieves a 17% improvement on EER and an 8% improvement on minDCF over the single class system. It should be noted that there is no further improvement for the sixteen-class MAPLR system. The lack of improvement for the sixteen-class is most likely due to the fact that as the number of classes increases the amount of adaptation data assigned to each class decreases. This leads to instance where there is not enough adaptation data to obtain a good transform for a given class.

Table 2. Comparison of the global single class MAPLR system with the multi-class MAPLR systems (male speakers, all trials). Rank normalization, NAP and Znorm are not used

System	EER (%)	minDCF
1C_MAPLR	9.85	0.0440
2C_MAPLR	8.58	0.0413
4C_MAPLR	8.48	0.0416
8C_MAPLR	8.14	0.0406
16C_MAPLR	8.23	0.0410

Table 3. Results of the eight-class MAPLR system (male speakers, all trials)

Matric	SVM	+Ranknorm	+NAP	+Znorm
EER (%)	8.14	8.01	7.78	7.36
minDCF	0.0406	0.0396	0.0381	0.0368

Table 3 lists the results of the eight-class MAPLR system on male speakers' data in the test corpus. Besides the results of SVM classifier, we also present the results after rank normalization process, after the rank normalization plus NAP process as well as after rank normalization and NAP plus Znorm process. It can be seen that all the three technique are effective in the eight-class MAPLR system.

Table 4. Results of the eight-class MAPLR system using different UBMs

System	Female	Male
Using female UBM	11.08 0.0508	9.37 0.0440
Using male UBM	11.15 0.0489	7.36 0.0413
Using both UBMs	9.67 0.0461	7.10 0.0365

Table 4 lists the performance using different UBMs. EER and minDCF values are observed on both male and female speakers' data. Compared to using single UBM, a better performance can be achieved when using multiple UBMs. This can leads to 14.5% improvement on EER and 9.2% improvement on minDCF for female data, and 3.5% improvement on EER and a little improvement on minDCF for male data.

Table 5 shows the results of the eight-class MAPLR system, the GMM-SVM system and the combination of the two systems. GMM-SVM is one of the best systems submitted for NIST 2008 SRE. We can see that the performance of MAPLR system is comparable to that obtained by GMM-SVM system, especially on the female data. Besides, it can bring significant gains on EER for both male and female speakers when combining the two systems, and it can also achieve performance improvement on minDCF. Compared to GMM-SVM, the combining system can lead to 6.2% EER and 3.0% minDCF reduction for female speakers, and 16.2% EER and 6.8% minDCF reduction for male speakers.

Table 5. Results of the eight-class MAPLR system, GMM-SVM system and the combination of the two systems on the whole test corpus (all speakers, all trials)

System	Female	Male
8C_MAPLR	9.47 0.0461	7.10 0.0365
GMM-SVM	9.85 0.0487	6.40 0.0346
Fusion	8.88 0.0447	5.95 0.0340

IV. Conclusion

In this paper, MAPLR is introduced for speaker recognition. Two kinds of multi-class techniques are proposed for further improving the MAPLR approach. Finally, the parameters of all the transforms for each utterance are concatenated into a supervector, which is used as the feature for SVM classification. Experiments show that MAPLR based speaker recognition shows better performance than MLLR, and the proposed multi-class MAPLR approach is quite effective for system fusion.

References

[1] A. Stolcke, L. Ferrer, S. Kajarekar, E. Shriberg and A. Venkataraman, "MLLR transforms as features in speaker recognition", *Proc. of Ninth Europe Conference on Speech Communication and Technology*, pp.2425–2428, 2005.

[2] M. Ferras, C. Leung, C. Barras and J. Gauvain, "Constrained MLLR for speaker recognition", *Proc. of ICASSP'07*, pp.53–56, 2007.

[3] M. Ferras, C. Leung, C. Barras and J. Gauvain, "MLLR techniques for speaker recognition", *Proc. of IEEE Odyssey Speaker and Language Recognition Workshop*, 2008.

[4] Z. Karam and W. Campbell, "A multi-class MLLR kernel for SVM speaker recognition", *Proc. of ICASSP'08*, pp.4117–4120, 2008.

[5] M.J.F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition", *Computer Speech and Language*, Vol.12, pp.75–98, 1998.

[6] C. Chesta, O. Siohan and C. Lee, "Maximum a posteriori linear regression for hidden Markov model adaptation", *Proc. of Sixth European Conference on Speech Communication and Technology*, Budapest, pp.211–214, 1999.
[7] X. Zhang, J. Gao, P. Lu and Y. Yan, "A novel speaker clustering algorithm via supervised affinity propagation", *Proc. of ICASSP'08*, pp.4369–4372, 2008.
[8] A. Stolcke, S. Kajarekar and L. Ferrer, "Nonparametric feature normalization for SVM-based speaker verification", *Proc. of ICASSP'08*, pp.1577–1580, 2008.
[9] A. Solomonoff, Q. Campbell and I. Boardman, "Advances in channel compensation for SVM speaker recognition", *Proc. of ICASSP'05*, 2005.
[10] "The NIST 2008 Speaker Recognition Evaluation Plan", *http://www.nist.gov/speech/tests/spk/2008/index.html*, 2008.

ZHANG Xiang received B.E. degree in Electronic Information Engineering from Shangdong University in 2006. Now he is a doctor candidate of ThinkIT Speech Laboratory, Institute of Acoustics, Chinese Academy of Sciences. His research interests include speaker recognition, language identification, speaker diarization, and audio watermarking. (Email: xzhang@hccl.ioa.ac.cn)

XIAO Xiang received B.E. degree in Electrical Engineering from University of Science and Technology of China in 2004. Now he is a doctor student of Institute of Acoustics, Chinese Academy of Sciences. His research interests include speaker recognition, language identification.

WANG Haipeng received B.S. degree from the Department of Electronic Science and Engineering in Nanjing University. Now he is a M.S. student in Institute of Acoustics, Chinese Academy of Sciences. His research interests include speech recognition, speaker recognition and language identification.

ZHANG Jianping received Ph.D. degree in Electronic Engineering from Tsinghua University in 1999. He is currently an Associated Professor in Institute of Acoustics, Chinese Academy of Sciences. His research interests include language/speaker recognition, and spoken language understanding. He had published nearly 10 papers in periodicals and conferences.

YAN Yonghong received B.E. degree from Tsinghua University in 1990, and Ph.D. degree from Oregon Graduate Institute (OGI). He worked in OGI as an Assistant Professor (1995), Associate Professor (1998) and Associate Director (1997) of Center for Spoken Language Understanding. He worked in Intel from 1998–2001, chaired Human Computer Interface Research Council, worked as Principal Engineer of Microprocessor Research Laboratory and Director of Intel China Research Center. Currently he is a professor and director of Think IT Laboratory. His research interests include speech processing and recognition, language/speaker recognition, and human computer interface. He has published more than 100 papers and holds 40 patents.