JOINT MAP ADAPTATION OF FEATURE TRANSFORMATION AND GAUSSIAN MIXTURE MODEL FOR SPEAKER RECOGNITION

Donglai Zhu, Bin Ma, Haizhou Li

Institute for Infocomm Research, A*Star, 1 Fusionopolis Way, Singapore 138632
{dzhu,mabin,hli}@i2r.a-star.edu.sg

ABSTRACT

This paper extends our previous work on feature transformation-based support vector machines for speaker recognition by proposing a joint MAP adaptation of feature transformation (FT) and Gaussian Mixture Models (GMM) parameters. In the new approach, the prior probability density functions (PDFs) of FT and GMM parameters are jointly estimated using the background data under the maximum likelihood criteria. In this way, we derive a generic prior GMM that is more compact than the Universal Background Model due to the reduction of speaker variations. With the prior PDFs, we construct a supervector to characterize a speaker using FT and GMM parameters. We conducted experiments on NIST 2006 Speaker Recognition Evaluation (SRE06) data set. The results validated the effectiveness of the joint MAP adaptation approach.

Index Terms— speaker recognition, feature transformation, maximum a posteriori, support vector machine

1. INTRODUCTION

In state-of-the-art approaches to the text-independent speaker recognition, the Gaussian mixture model (GMM) and support vector machine (SVM) have been successfully used as classifiers. An advantage of the SVM-based approach is the flexibility of designing the supervectors, e.g. the generalized linear discriminant sequence (GLDS) [1], the concatenation of GMM mean vectors [2], and parameters of maximum likelihood linear regression (MLLR) [3].

In our previous works, we proposed using parameters of a *feature transformation* (FT) function to construct the SVM supervectors [4]. The FT function was designed to convert speaker-dependent features to speaker-independent features, and therefore, its parameters can be used to characterize the speakers. with reference to a speaker-independent universal background model (UBM), the FT parameters are estimated with the maximum-a-posteriori (MAP) criteria. The approach provides a flexibility of clustering transformation matrices and bias vectors with two separate regression classes.

In this paper, we present a joint MAP adaptation method on both FT and GMM parameters for speaker recognition. It has two advantages over the previous FT method. Firstly, speakers are modeled by both FT and GMM parameters that are estimated by a *joint MAP adaptation* process. The speaker characteristics in both FT transformations and the speaker-dependent GMM will be modeled. Secondly, the prior probability density functions (PDFs) of FT and GMM parameters are jointly estimated. A *generic prior* GMM, which is more compact than the UBM due to the normalization of speaker variations in the model, is produced for the MAP adaptation. The generic prior model has demonstrated the capability of improving speaker adaptation performance in speech recognition by modeling separately the phonetic variation and speaker variation, and reducing the variation and overlap of the acoustic models [5].

This paper is organized as follows. Section 2 presents the speaker modeling with FT and GMM. Section 3 presents the joint MAP adaptation of both FT and GMM parameters. Section 4 presents the experimental results. Finally, we conclude in Section 5.

2. GMM WITH FEATURE TRANSFORMATION

Let's assume that a speech utterance spoken by a speaker s has been represented by a sequence of feature vectors $Y^{(s)} = \{y_t^{(s)}\}_{t=1}^{T}$, where $y_t^{(s)}$ is a D-dimensional vector. We define a feature transformation (FT) function that maps the speaker's feature vector $y^{(s)}$ to a pseudo feature vector $x^{(s)}$ as follows:

$$x^{(s)} \triangleq \mathcal{F}(y^{(s)}; \Theta^{(s)}) = A_k^{(s)} y^{(s)} + b_l^{(s)} , \qquad (1)$$

where $A_k^{(s)}$ is a nonsingular $D \times D$ matrix, $b_l^{(s)}$ is a D-dimensional vector, and $\Theta^{(s)} = \{A_k^{(s)}, b_l^{(s)}; k = 1, \cdots, K; l = 1, \cdots, L\}$.

Let's model $x^{(s)}$ with a GMM of which parameters are denoted as $\Lambda^{(s)} = \{c_m^{(s)}, \mu_m^{(s)}, \Sigma_m^{(s)}; m = 1, \cdots, M\}$, where M is the number of Gaussian components, $c_m^{(s)}$'s are Gaussian mixture weights, $\mu_m^{(s)} = [\mu_{m1}^{(s)}, \cdots, \mu_{mD}^{(s)}]^T$ is a D-dimensional mean vector, and $\Sigma_m^{(s)} = diag\{\sigma_{m1}^{(s)2}, \cdots, \sigma_{mD}^{(s)2}\}$ is a diagonal covariance matrix.

The transformation classes k and l are associated with Gaussian components in GMM by sharing the transformation

across mixture components. A centroid splitting algorithm with Euclidean distance measure is used to map each component m to two separate classes: $\mathcal{C}_k = \{m|k_m = k\}; k = 1, \cdots, K$ and $\mathcal{C}_l = \{m|l_m = l\}; l = 1, \cdots, L$.

3. JOINT MAP ADAPTATION OF PARAMETERS

Given a speaker's data $Y^{(s)}$, we estimate the FT parameters $\Theta^{(s)}$ and GMM parameters $\Lambda^{(s)}$ with the MAP criteria. Suppose the prior PDFs of $A_k^{(s)}$, $b_l^{(s)}$ and $\Lambda^{(s)}$ are respectively known as $p(A_k^{(s)}|\Gamma_A), p(b_l^{(s)}|\Gamma_b)$ and $p(\Lambda^{(s)}|\Gamma_\Lambda)$, where $\Gamma_A, \Gamma_b, \Gamma_\Lambda$ are hyperparameters in the prior PDFs. The MAP adaptation of $\Theta^{(s)}$ and $\Lambda^{(s)}$ is to maximize the following posteriori PDF:

$$\mathcal{P}(\Theta^{(s)}, \Lambda^{(s)}) = p(\mathcal{F}(Y^{(s)}; \Theta^{(s)})|\Lambda^{(s)}) \times p(\Lambda^{(s)}|\Gamma_\Lambda) \prod_{k=1}^{K} p(A_k^{(s)}|\Gamma_A) \prod_{l=1}^{L} p(b_l^{(s)}|\Gamma_b) \ . \quad (2)$$

By strict Bayesian learning definition, the hyperparameter set $\Gamma = \{\Gamma_A, \Gamma_b, \Gamma_\Lambda\}$ is assumed known based on some subjective knowledge of $\Theta^{(s)}$ and $\Lambda^{(s)}$. In reality, it is difficult to possess a complete knowledge of the prior distributions. We adopt the *empirical Bayes* approach [6] to derive the hyperparameters from the background data recorded by a number of speakers. The estimated prior PDFs can model the information of the variability of $\Theta^{(s)}$ and $\Lambda^{(s)}$ among different speakers. As we can pretrain $\Theta^{(s)}$ and $\Lambda^{(s)}$ on the background data in the absence of the data of target speakers, we use a parameter estimation method called τ-*initialization* [7], in which the hyperparameter set Γ is specified directly from the pretrained models with the assistance of a user-defined control parameter τ.

3.1. Choice of prior PDFs

The prior PDF of $A_k^{(s)}$ is defined as a matrix variate normal PDF [8]:

$$p(A_k^{(s)}|\Gamma_A) \propto |\Xi|^{-D/2}|\Phi|^{-D/2} \exp[-\frac{1}{2}\mathrm{tr}(A_k^{(s)} - U_k)^T \Xi^{-1}(A_k^{(s)} - U_k)\Phi^{-1}] \ , \quad (3)$$

where $\Gamma_A = \{U_k, \Xi, \Phi\}$ is the set of hyperparameters, with $U_k \in \mathbb{R}^{D\times D}$, $\Xi \in \mathbb{R}^{D\times D}$, $\Xi \geq 0$, and $\Phi \in \mathbb{R}^{D\times D}$, $\Phi \geq 0$. We fix $\Xi = cI$ and $\Phi = I$, where c is a scalar control parameter and I is an identity matrix. When the value of c gets smaller, the MAP estimation of $A_k^{(s)}$ becomes closer to the prior parameter U_k; on the contrary the MAP estimation gets closer to the ML estimation.

The prior PDF of $b_l^{(s)}$ is defined as a normal PDF:

$$p(b_l^{(s)}|\Gamma_b) \propto \exp\left[-\frac{\tau_b}{2}(b_l^{(s)} - \rho_l)^T \Psi_l^{-1}(b_l^{(s)} - \rho_l)\right] \ , \quad (4)$$

where $\Gamma_b = \{\Psi_l, \rho_l; l = 1, \cdots, L\}$ is the set of hyperparameters, with $\tau_b > 0$, ρ_l being a D-dimensional vector, and $\Psi_l = diag\{\psi_{l1}^2, \cdots, \psi_{lD}^2\}$ being a diagonal covariance matrix. We fix τ_b to be a constant.

The prior PDF of $\Lambda^{(s)}$ is composed of a Dirichlet PDF for mixture weights and a normal-Wishart PDF for means and variances [7]:

$$p(\Lambda^{(s)}|\Gamma_\Lambda) \propto \prod_m c_m^{(s)\nu-1}|\Sigma_m^{(s)}|^{\frac{D-\alpha}{2}} \exp[-\frac{\tau}{2}(\mu_m^{(s)} - \mu_m)^T \Sigma_m^{(s)-1}(\mu_m^{(s)} - \mu_m) - \frac{1}{2}tr(\Sigma_m\Sigma_m^{(s)-1})] \ , \quad (5)$$

where $\Gamma_\Lambda = \{\nu, \tau, \alpha, \mu_m, \Sigma_m; m = 1, \cdots, M\}$ is the set of hyperparameters, with $\nu > 0$, $\tau > 0$, $D - \alpha < -1$; μ_m and Σ_m are respectively the mth mean and variance in a generic GMM that models variability among speakers.

3.2. Prior PDF estimation

We estimate the hyperparameter set Γ with the maximum likelihood (ML) estimation on the background data $\mathcal{Y}$. As the mean matrix in the prior PDF of $A_k^{(s)}$ is U_k (Eq. 3) and the prior PDF of $b_l^{(s)}$ follows Eq. (4), the likelihood of $\mathcal{Y}$ given Γ can be written as follows:

$$p(\mathcal{Y}|\Gamma) = \prod_t \sum_m c_m |U_k| \mathcal{N}(U_k y_t; \mu_m - \rho_l, \Sigma_m + \tau_b^{-1}\Psi_l) \ . \quad (6)$$

This likelihood is maximized by using an *alternative estimation* procedure of the hyperparameter set Γ as below:

Step 1: *Initialization*

The initial values of Gaussian mixture parameters (c_m, μ_m and Σ_m) are set to be the parameters of the UBM trained on the background data. U_k's are initialized to be identity matrices. ρ_l's are set to be zero vectors and Ψ_l's are set to be diagonal matrices with small values (e.g. 0.01).

Step 2: *Estimating* Γ_A *by fixing* Γ_b *and* Γ_Λ

In the hyperparameter set Γ_A, U_k can be estimated by fixing the other hyperparameters Γ_b and Γ_Λ. The inference of update formula of U_k is similar to CMLLR [9]. Several EM iterations can be performed.

Step 3: *Estimating* Γ_b *and* Γ_Λ *by fixing* Γ_A

Given parameters of Γ_A, we estimate parameters of Γ_b and Γ_Λ with several EM iterations. For simplicity, let's define $z_t = U_k y_t$. The pseudo feature vector after transformation (Eq. 1) is then rewritten as $x_t = z_t + b_t$. Given z_t as observation vectors, the integrated model of x_t and b_t is estimated with the EM algorithm [10][11][12][13]. In the E-step of the

EM algorithm, the sufficient statistics of Γ_b and Γ_Λ are estimated as follows:

$$\begin{aligned}
\tilde{x}_m(t) &= E(x_t|z_t,m) = \mu_m + \Delta_m^{(1)}\epsilon_m(t) \ , \\
\tilde{b}_m(t) &= E(b_t|z_t,m) = \rho_{l_m} - \Delta_m^{(2)}\epsilon_m(t) \ , \\
\tilde{L}_m(t) &= E(x_t x_t^T|z_t,m) = \tilde{x}_m(t)\tilde{x}_m^T(t) + \Delta_m^{(3)} \ , \\
\tilde{V}_m(t) &= E(b_t b_t^T|z_t,m) = \tilde{b}_m(t)\tilde{b}_m^T(t) + \Delta_m^{(3)} \ ,
\end{aligned}$$

where

$$\begin{aligned}
\Delta_m^{(1)} &= \Sigma_m(\Sigma_m + \tau_b^{-1}\Psi_{l_m})^{-1} \ , \\
\Delta_m^{(2)} &= I - \Delta_m^{(1)} \ , \\
\Delta_m^{(3)} &= \tau_b^{-1}\Delta_m^{(1)}\Psi_{l_m} \ , \\
\epsilon_m(t) &= z_t + \rho_{l_m} - \mu_m \ ,
\end{aligned}$$

and I is a $D \times D$ identity matrix. The re-estimation formulae of Γ_b and Γ_Λ are as follows:

$$\bar{c}_m = \frac{\sum_t \gamma_m(t)}{\sum_t \sum_m \gamma_m(t)} \ , \quad (7)$$

$$\bar{\mu}_m = \frac{\sum_t \gamma_m(t)\Delta_m^{(1)}\epsilon_m(t)}{\sum_t \gamma_m(t)} + \mu_m \ , \quad (8)$$

$$\bar{\Sigma}_m = diag\{\frac{\sum_t \gamma_m(t)\tilde{L}_m(t)}{\sum_t \gamma_m(t)} - \bar{\mu}_m\bar{\mu}_m^T\} \ , \quad (9)$$

$$\bar{\rho}_l = \frac{-\sum_t \sum_{m\in\mathcal{C}_l} \gamma_m(t)\Delta_m^{(2)}\epsilon_m(t)}{\sum_t \sum_{m\in\mathcal{C}_l} \gamma_m(t)} + \rho_l \ , \quad (10)$$

$$\bar{\Psi}_l = \tau_b diag\{\frac{\sum_t \sum_{m\in\mathcal{C}_l} \gamma_m(t)\tilde{V}_m(t)}{\sum_t \sum_{m\in\mathcal{C}_l} \gamma_m(t)} - \bar{\rho}_l\bar{\rho}_l^T\} \quad (11)$$

Step 4: *Repeating* **Step 2** *and* **Step 3** *several times until a pre-set criterion is satisfied*.

3.3. MAP estimation of FT and GMM parameters

Given the prior PDFs (Eq. 3,4,5), we estimate the FT parameters $\Theta^{(s)}$ and the GMM parameters $\Lambda^{(s)}$ to maximize the posterior PDF (Eq. 2). The estimation is still an *alternative estimation* procedure as follows.

Step 1: *Initialization*

The transformation matrices $A_k^{(s)}$'s are initialized to be identity matrices. The bias vectors $b_l^{(s)}$'s are initialized to be zero vectors. The GMM parameters $\Lambda^{(s)}$ are initialized to be parameters of the generic prior GMM Λ.

Step 2: *Estimating* $\Theta^{(s)}$ *by fixing* $\Lambda^{(s)}$

In this step, we estimate the FT parameters $\Theta^{(s)}$ by fixing the GMM parameters $\Lambda^{(s)}$. The estimation includes two sub-steps. In the first sub-step, the transformation matrices $A_k^{(s)}$ are estimated by fixing the bias vectors $b_l^{(s)}$. In the second sub-step, $b_l^{(s)}$ are estimated by fixing $A_k^{(s)}$. In both sub-steps, several EM iterations can be performed. The update formulae of $A_k^{(s)}$ and $b_l^{(s)}$ are similar to those in our previous paper [4], while the only difference is that the GMM parameters are the re-estimated $\Lambda^{(s)}$ instead of the UBM parameters.

Step 3: *Estimating* $\Lambda^{(s)}$ *by fixing* $\Theta^{(s)}$

Given the updated $\Theta^{(s)}$, we can estimate $\Lambda^{(s)}$ in a similar way to the MAP estimation of Gaussian densities [7]. The update formula of mean vectors is as follows:

$$\bar{\mu}_m^{(s)} = \frac{\tau\mu_m + \sum_t \gamma_m(t)(A_k^{(s)}y^{(s)} + b_l^{(s)})}{\tau + \sum_t \gamma_m(t)} \ . \quad (12)$$

Step 4: *Repeating* **Step 2** *and* **Step 3** *several times until a pre-set criterion is satisfied*.

4. EXPERIMENTS

Our evaluation data is the core test condition (1-conversation training, 1-conversation test, all trials) of the 2006 NIST SRE (SRE06). The background data consists of 8000 speech utterances of 2.5 minutes duration from the Switchboard corpora, which cover a number of speakers (female and male) and channels. The Nuisance Attribute Projection (NAP) [14] training data includes 3383 speech utterances of 2.5 minutes duration, recorded by 310 speakers, from the 2004 NIST SRE corpus. The 1-conversation training data in the 2005 NIST SRE corpus are used for training cohort models in Tnorm score normalization [15]. The input 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. With the background data, we train two gender-dependent GMMs each including 512 Gaussian components.

We set $K = 1$ and $L = 512$ in the joint MAP adaptation of FT and GMM (JMAP). In prior PDF of $A_k^{(s)}$ (Eq. 3), we set $c = 100$. In Eq. (4) and (5), we set $\tau_b = \tau = 20$. In both the estimation of prior PDFs (Section 3.2) and the estimation of speaker-dependent parameters (Section 3.3), we perform one EM iteration in Step 2 and Step 3, and skip Step 4. The JMAP method is compared with other two methods: 1) the MAP-MEAN SVM method in which only the MAP adaptation of GMM means is performed [2], and 2) the MAP-FT SVM method in which only the MAP adaptation of FT parameters is performed [4].

In JMAP and MAP-FT methods, supervectors are normalized with a rank normalization to equate their dynamic ranges [3]. NAP (with a matrix rank of 40) is performed to reduce the nuisance effects in speech [14]. An SVM is trained for

each target speaker by regarding the target speaker's training supervectors as positive examples, and the supervectors from the background data as negative examples. Finally, the SVM scores are normalized with Tnorm.

We consider two kinds of JMAP-adapted parameters as the SVM supervectors for speaker recognition, one denoted by JMAP(I) consisting of $\Theta^{(s)}$ parameters only, and another denoted by JMAP(II) consisting of both $\Theta^{(s)}$ and $\Lambda^{(s)}$ parameters. JMAP(I) differs from MAP-FT in adopting the joint estimation of prior PDFs. Table 1 shows the EER and minimum DCF values of four types of supervectors at the stages of SVM, NAP and Tnorm, respectively, on the SRE06 data of male speakers. In comparison with MAP-FT and MAP-MEAN, JMAP(I) and JMAP(II) effectively improve the performance. There is slight difference between the performance of JMAP(I) and JMAP(II), which indicates that the JMAP-adapted GMM means are insignificant in the supervector. Considering its comparable results and low computational cost, we suggest using JMAP(I) as the SVM supervectors. Table 2 summarizes the results on the data of both female and male speakers. Comparing with MAP-FT SVM, JMAP method reduces the EER and minimum-DCF by 5.3% and 3.6% relatively. Comparing with MAP-MEAN SVM, the relative reductions are 19.5% and 17.3%, respectively.

5. CONCLUSIONS

We presented a joint MAP adaptation method of feature transformation (FT) and GMM. Speakers are modeled by GMMs with FT, which combines feature-space and model-space modeling. The prior PDFs of FT and GMM parameters are jointly estimated on the background data, which yields a generic prior GMM that is more compact than the UBM. Results on the 2006 NIST SRE corpus show that the method effectively improves the performance over the methods using only FT or GMM means. In future, we will study the contributions of transformation matrices and bias vectors by setting different numbers of classes for them, and applying FT to other speaker recognition methods such as the GMM-UBM.

Table 1. *Results on the core test condition of SRE06 (male speakers, all trials). The upper row (in italics) in each table cell is the EER (%). The lower row is the minimum DCF value (X100).*

Supervector	SVM	+NAP	+Tnorm
MAP-MEAN	*7.26*	*4.68*	*4.14*
	3.60	2.39	2.15
MAP-FT	*6.67*	*4.05*	*3.57*
	3.44	2.09	1.91
JMAP (I)	*6.12*	*3.73*	*3.31*
	3.18	1.95	1.81
JMAP (II)	*6.10*	*3.73*	*3.32*
	3.19	1.96	1.79

Table 2. *Results on the core test condition of SRE06 (all speakers, all trials). The upper row (in italics) in each table cell is the EER (%). The lower row is the minimum DCF value (X100).*

Method	Female	Male	All
MAP-MEAN	*5.68*	*4.14*	*5.12*
SVM	2.90	2.15	2.60
MAP-FT	*4.85*	*3.57*	*4.35*
SVM	2.47	1.91	2.23
JMAP (I)	*4.61*	*3.31*	*4.12*
SVM	2.36	1.81	2.15

6. REFERENCES

[1] Campbell W. M., "Generalized linear discriminant sequence kernels for speaker recognition," in *Proc. ICASSP*, 2002, pp. 161-164.

[2] 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.

[3] 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.

[4] Zhu D., Ma B. and Li H., "Using MAP estimation of feature transformation for speaker recognition," in *Proc. Interspeech*, 2008.

[5] Anastasakos T., McDonough J., Schwartz R. and Makhoul J., "A compact model for speaker adaptive training," in *Proc. ICSLP*, 1996, pp. 1137-1140.

[6] Carlin B. P. and Louis T. A., "Bayes and Empirical Bayes Methods for Data Analysis," Chapman & Hall, London, 1996.

[7] Gauvain J.-L. and Lee C.-H., "Bayesian learning for hidden Markov models with Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 291-298, Apr. 1994.

[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, no. 1, pp. 5-24, 2002.

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

[10] Rose R. C., Hofstetter E. M. and Reynolds D. A., "Integrated models of signal and background with application to speaker identification in noise," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 245-257, Apr. 1994.

[11] Sankar A. and Lee C.-H., "A maximum-likelihood approach to stochastic matching for robust speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 190-202, May 1996.

[12] Afify M., Gong Y. and Haton J.-P., "A general joint additive and convolutive bias compensation approach applied to noisy lombard speech recognition, " *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 524-538, Nov. 1998.

[13] Wu J. and Huo Q., "A switching linear Gaussian hidden Markov model and its application to nonstationary noise compensation for robust speech recogntion," in *Proc. ICASSP*, 2003. pp. 977-980.

[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.