MULTIGRAINED MODEL ADAPTATION WITH MAP AND REFERENCE SPEAKER WEIGHTING FOR TEXT INDEPENDENT SPEAKER VERIFICATION

Xianyu Zhao[1], Yuan Dong[1,3], Jun Luo[2],Hao Yang[3], Haila Wang[1]

[1]France Telecom R&D Center (Beijing), Beijing, 100080, P.R.China
{xianyu.zhao, yuan.dong, haila.wang}@rd.francetelecom.com
[2]Department of Electronics Engineering, Tsinghua University, Beijing, 100084, P.R.China
luoj97@mails.tsinghua.edu.cn
[3]Beijing University of Posts and Telecommunications, Beijing, 100876, P.R.China
yuandong@bupt.edu.cn

ABSTRACT

When traditional Maximum a Posteriori (MAP) adaptation is used to adapt a universal background model (UBM), some model components with little or no enrollment data would remain unchanged in the derived speaker model. These model components would have weak discriminative capability over the background model, and would impair subsequent verification performance. In this paper, we present a new speaker adaptation method which combines MAP and Reference Speaker Weighting (RSW) adaptation in a hierarchical, multigrained mode. It enables all model components to be updated in a way that strikes a good balance between model complexity and available data. The experimental results of NIST speaker recognition evaluation confirmed the effective performance increase with this new method compared with using MAP or RSW adaptation techniques alone.

1. INTRODUCTION

The Gaussian Mixture Model – Universal Background Model (GMM-UBM) system [1] is widely used for text independent speaker verification tasks. In this approach, the UBM is a single, speaker-independent GMM with a large number of mixture components (1024-2048) trained from a vast amount of speech data of many speakers. And a target speaker model is derived from adapting the parameters of the UBM using the speaker's enrollment data. A form of Bayesian adaptation or Maximum a Posteriori (MAP) estimation is generally used for speaker adaptation [1]. In classical MAP adaptation, each Gaussian component in UBM is updated individually. With limited adaptation data and imbalanced coverage of model components, some components with little or no adaptation data would remain unchanged in the derived speaker model and lose discriminative capability over UBM. This would impair subsequent verification performance.

To address these problems, one prominent class of methods (e.g. Eigenvoice modeling [2], Reference Speaker Weighting [3], Maximum Likelihood Model Interpolation [4] etc.) is to represent the target speaker's model as a linear combination of a set of reference acoustic models. In this case, all model components could be adapted through a set of combination or interpolation coefficients. However, since all speaker models are constrained to lie in the linear space spanned by the reference models, model complexity and discriminative capabilities are restricted. This would also deteriorate verification performance.

In this paper, a new hybrid method combining MAP and RSW adaptation in a hierarchical and multigrained mode is proposed. Depending on the amount of adaptation data assigned to each model component, different adaptation strategies are applied to guarantee the discriminative capability of each component and the whole model. Through properly organizing the model components with insufficient adaptation data into a regression class tree and applying hierarchical RSW to them, a balance between model complexity and available adaptation data is well achieved. The new adaptation method was tested with the NIST speaker recognition evaluation [5]. Experimental results showed that speaker models adapted with this new method had improved verification performance compared with conventional MAP and RSW adaptation techniques.

2. MULTIGRAINED ADAPTATION WITH MAP AND REFERENCE SPEAKER WEIGHTING

2.1. MAP adaptation

In the context of a GMM-UBM system for text-independent speaker recognition, consider a UBM with K components in which the model parameters λ are defined as $\lambda = \{w_i, \mu_i, \Sigma_i; i = 1, \cdots, K\}$, where w_i, μ_i and Σ_i are the

component weight, mean vector and covariance matrix of the *i*-th component respectively. The MAP adapted speaker model is obtained through using new sufficient statistics collected from the adaptation data, $X=\{x_1,\cdots,x_T\}$, to update the old background model sufficient statistics for mixture components [1], i.e., (in our following discussion and experiments, we focus on and do model mean vector adaptation only)

$$\hat{\mu}_i=\alpha_i E_i(X)+(1-\alpha_i)\mu_i, \tag{1}$$

where the new sufficient statistics are calculated as

$$n_i=\sum_{t=1}^{T}P(i|x_t)$$

$$E_i(X)=\sum_{t=1}^{T}P(i|x_t)x_t\Big/n_i.$$

In the above equations, $P(i|x_t)$ is the a posteriori probability of the *i*-th component given observation data x_t, which is

$$P(i|x_t)=\frac{w_i N(x_t;\mu_i,\Sigma_i)}{\sum_j w_j N(x_t;\mu_j,\Sigma_j)}.$$

The adaptation coefficient that controls the weight of a priori information is α_i and is defined to be

$$\alpha_i=n_i/(n_i+f), \tag{2}$$

where f is a fixed "relevance" factor (chosen to be 16 in our experiments that follow).

From the above derivation, it can be seen clearly that mixture components with large amounts of adaptation data would rely more on the new statistics and be well adapted to the target speaker; while for components with little data, they would be dominated by old statistics for the UBM and be poorly adapted.

2.2. RSW adaptation

For RSW adaptation, we begin with S reference speakers and train a model $m^s=\{w_i^s,\mu_i^s,\Sigma_i^s;i=1,\cdots,K\}$ for each of them. All of the component mean vectors in model m^s are concatenated into a supervector:

$$M^s=\left[(\mu_1^s)^T,(\mu_2^s)^T,\cdots,(\mu_K^s)^T\right]^T,$$

where μ_k^s is the mean vector of *k*-th component in the *s*-th reference model, $(\cdot)^T$ stands for vector transpose.

Let the supervector of component mean vectors of the target speaker be Λ, i.e.

$$\Lambda=\left[\hat{\mu}_1{}^T,\hat{\mu}_2{}^T,\cdots,\hat{\mu}_K{}^T\right]^T.$$

In RSW, it is assumed that Λ is a linear combination of the S supervectors of the reference speaker models, i.e.,

$$\Lambda=\beta_1 M^1+\beta_2 M^2+\cdots+\beta_S M^S. \tag{3}$$

The Maximum Likelihood (ML) estimation of these combination coefficients $\vec{\beta}=\{\beta_1,\cdots,\beta_S\}$ aims to maximize the likelihood of $p(X|\vec{\beta})$ with respect to $\vec{\beta}$. This is done through an EM algorithm, i.e., iteratively optimizing an auxiliary function $Q(\vec{\beta},\vec{\beta}')$ with respect to $\vec{\beta}$ [2, 3]

$$Q(\vec{\beta},\vec{\beta}')=\sum_{t=1}^{T}\sum_{i=1}^{K}P(i|x_t,\vec{\beta}')\log p(x_t|i,\vec{\beta}), \tag{4}$$

where $\vec{\beta}'$ is the current estimate of combination coefficients and $P(i|x_t,\vec{\beta}')$ is the a posteriori probability of the *i*-th model component give the observation data x_t and the current estimate $\vec{\beta}'$,

$$P(i|x_t,\vec{\beta}')=\frac{w_i N\left(x_t;\sum_{s=1}^{S}\beta_s'\mu_i^s,\Sigma_i\right)}{\sum_{j=1}^{K}w_j N\left(x_t;\sum_{s=1}^{S}\beta_s'\mu_j^s,\Sigma_j\right)}. \tag{5}$$

Let $\partial Q/\partial\beta_s=0, s=1,2,\cdots,S$. We obtain the update equation for each $\beta_s, s=1,2,\cdots,S$:

$$\begin{aligned}&\sum_{t=1}^{T}\sum_{i=1}^{K}P(i|x_t,\vec{\beta}')(x_t)^T\Sigma_i^{-1}\mu_i^s\\&=\sum_{t=1}^{T}\sum_{i=1}^{K}P(i|x_t,\vec{\beta}')\sum_{l=1}^{S}\beta_l(\mu_i^l)^T\Sigma_i^{-1}\mu_i^s.\end{aligned} \tag{6}$$

In RSW adaptation the statistics for all model components are used to estimate a set of global combination parameters, so that even if some model components have little (or no) adaptation data they still can be updated. But the constraint of all speaker models to lie in the linear space spanned by a limited number of reference models restricts model complexity and impairs discriminative capabilities.

2.3. Multigrained model adaptation with MAP and RSW

As mentioned above, MAP adaptation has good asymptotic property, but the model components could not be updated in a balanced way with limited adaptation data; on the other hand, RSW adaptation can be used to adapt a model globally with a small amount of data, but has constrained model complexity which also impairs discriminative capabilities among speaker models. In order to further improve the discriminative capability of adapted target speaker models in the GMM-UBM system, we combined them in a multigrained mode.

For those components that acquire a large amount of adaptation data (above a threshold), we let them adapt individually in a MAP mode, i.e.,

$$\hat{\mu}_i=\alpha_i E_i(X)+(1-\alpha_i)\mu_i,\quad if\quad n_i>C_t, \tag{7}$$

where C_t is the threshold.

In order to well update the model as a whole, RSW is applied to other model components with little adaptation data. The simple and direct way is to treat these components as a whole, and obtain one set of RSW combination coefficients to update them all.

However, in order to increase the model complexity of the RSW adapted model, RSW adaptation is applied to these model components in a hierarchical, multigrained mode. Like those used for tree-based MLLR [6] or hierarchical Eigenvoice [7], these model components are clustered into a binary regression class tree. Instead of using a global set of combination parameters for them, components associated with a particular regression class are assumed to share a particular set of combination parameters.

A simple binary regression class tree with four base classes is shown in Fig.1. A regression class tree, T, consists of a hierarchy of regression classes, $\{r_1, r_2, r_3\}$ and a set of base classes, $\{r_4, r_5, r_6, r_7\}$. In each base class, there is a model component, v, and its adaptation data acquired, D_v. For a regression class, the model components and adaptation data in it are defined to be those contained in its children.

For a group of components in node r, $\{v_1, v_2, \cdots, v_L\} \in r$, the combination parameters for these components, $\vec{\beta}(r) = \{\beta_1(r), \cdots, \beta_S(r)\}$, can be obtained through the following set of equations, which is a slight modification of equation (6),

$$\sum_{t=1}^{T} \sum_{i \in \{v_1, \cdots, v_L\}} P\left(i \middle| x_t, \vec{\beta}'(r)\right) (x_t)^T \Sigma_i^{-1} \mu_i^s = \sum_{t=1}^{T} \sum_{i \in \{v_1, \cdots, v_L\}} P\left(i \middle| x_t, \vec{\beta}'(r)\right) \sum_{l=1}^{S} \beta_l(r) \left(\mu_i^l\right)^T \Sigma_i^{-1} \mu_i^s , \quad (8)$$

where $\vec{\beta}'(r)$ is the current estimate of combination parameters for the regression class node r, and $\vec{\beta}(r)$ is to be estimated in this turn of iteration. The adapted mean vector for component v_g is calculated as

$$\hat{\mu}_{v_g} = \beta_1(r) \mu_{v_g}^1 + \beta_2(r) \mu_{v_g}^2 + \cdots + \beta_S(r) \mu_{v_g}^S . \quad (9)$$

To achieve reliable model adaptation, the amount of adaptation data and model complexity (the number of adaptation parameters) have to be balanced. So, a procedure called hierarchical RSW adaptation is undertaken to select nodes in the regression class tree to do RSW adaptation based on the amount of available adaptation data. Hierarchical RSW adaptation transverses the regression tree in a post-order manner. For a regression class node, RSW adaptations are firstly tried for its left and right child nodes. If in these child nodes there are sufficient adaptation data for reliable estimation of combination parameters, model components in these nodes are updated using equations (8) and (9); otherwise, the parent regression class in the higher layer is selected for RSW adaptation. For Gaussian components that have already been adapted in regression classes in lower layers, they are kept unchanged during adaptations in higher level classes. For example, in the regression class tree shown in Fig.1, if regression class r_2 has sufficient data to estimate RSW combination parameters $\vec{\beta}(r_2)$, it is selected to adapt Gaussian components $\{v_1, v_2\}$. If regression class r_3 does not have sufficient data, then it resorts to its parent class. In this case, all adaptation data of Gaussian components $\{v_1, v_2, v_3, v_4\}$ are used for the estimation of RSW combination parameters of class r_1, $\vec{\beta}(r_1)$. If this is successful, the adaptation of Gaussian components $\{v_3, v_4\}$ is carried out through $\vec{\beta}(r_1)$; while for components $\{v_1, v_2\}$ already adapted in child class r_2, $\vec{\beta}(r_1)$ is ignored.

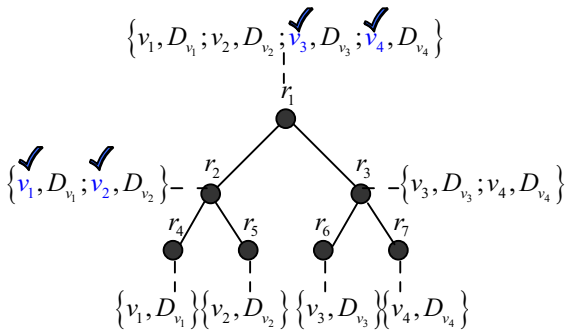

Fig. 1. An exemplar regression class tree and hierarchical RSW adaptation procedure

3. EXPERIMENTAL RESULTS

In this section, we report on speaker verification experiments conducted on the male part of NIST 1998 speaker recognition evaluation dataset. The evaluation includes 250 speakers. For each speaker, approximately 2 minutes of speech from a single telephone call is used for enrollment, i.e. one-session training condition. Verification utterances are normally 30 seconds in duration. There are 2500 verification utterances. Each verification utterance is scored against 10 putative speaker models.

Speech is sampled at 8 KHz and the 28-dimensional feature vector is formed by 14 MFCC's plus their first order differentials. A 20-ms window length and a 10-ms frame shift are used. RASTA and Feature mapping are applied as in [8,9]. We use a GMM consisting of 1024 Gaussians as the UBM.

Results are presented using Detection Error Tradeoff (DET) plots. Performance is computed after collecting all verification scores. Along with Equal Error Rate (EER), the minimum decision cost function (DCF), defined by NIST as

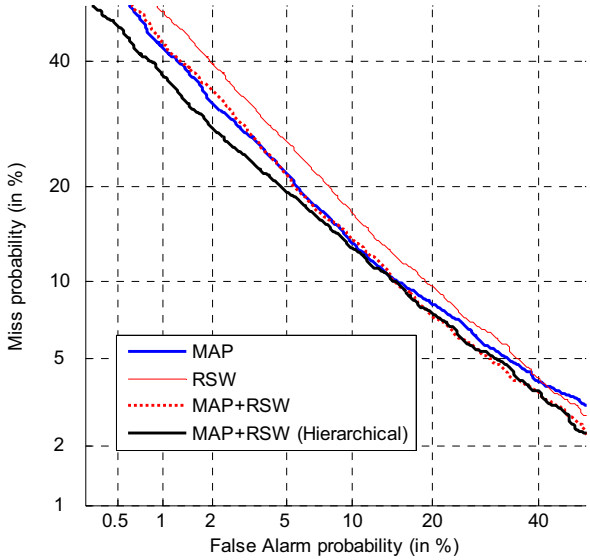

Fig. 2. DET curves for 4 different adaptation schemes

DCF = 0.1 * Pr(miss) + 0.99 * Pr(false_alarm), is also used as an overall performance measure.

In Fig.2, we show DET curves for four speaker model adaptation schemes. In this figure, "MAP" and "RSW" correspond to MAP and RSW adaptation respectively. For RSW adaptation, 230 male speakers in NIST SRE'99 were used as reference speakers. "MAP+RSW" stands for the combination scheme which is not hierarchical; i.e., we treat those components not MAP adapted as whole and get a single set of RSW combination coefficients to update them. "MAP+RSW (Hierarchical)" is the case of applying RSW in a hierarchical and multigrained mode. In our experiments that combining MAP and RSW, the component's MAP adaptation threshold (see equation (7)), C_t, is chosen to be 10. With hierarchical RSW, the candidate node in the regression tree for adaptation was required to have at least 50 mixture components and 50 frames of adaptation data in it. This kind of requirement improves the robustness of and speeds up the hierarchical RSW adaptation.

From this figure, it can be seen that the RSW adapted models provide worse verification performance than using MAP due to the model complexity constraint. After combining MAP and RSW, "RSW+MAP" achieves comparable performance with MAP and shows some improvements in the area of high false alarm and low missing. Furthermore, through the combination of MAP and RSW in a hierarchical and multigrained mode, this combination strategy achieves better performance than using MAP adaptation alone, especially in the area of low false alarm and high missing. Compared with MAP, the EER of "MAP+RSW (Hierarchical)" drops slightly from 11.8% to 11.6%, while the minimum DCF value drops from 51×10^{-3} to 46×10^{-3}. This result confirms the advantage of combining MAP and RSW properly.

4. CONCLUSION

A speech adaptation scheme that combines MAP and RSW in a hierarchical and multigrained mode was developed in this paper. Model components which acquire sufficient adaptation data are chosen to perform MAP adaptation to guarantee good asymptotic behavior of Bayesian adaptation; other model components with insufficient data are grouped together through a regression class tree, and RSW adaptations are applied to them in a hierarchical mode. Through these means, all model components are updated well with a balance between model complexity and available data. The experimental results using the NIST speaker recognition evaluation dataset show that better verification performance is obtained with models adapted through this new method. Further experiments with more extensive datasets and different configurations of adaptation controlling parameters are planned in future work.

5. REFERENCES

[1] D. Reynolds, T. Quatieri and R. Dunn, "Speaker Verification using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol.10, pp. 19-41, Jan. 2000.

[2] R. Kuhn, J.-C. Junqua, P. Nguyen and N. Niedzielski, "Rapid Speaker Adaptation in Eigenvoice Space," *IEEE Trans. Speech and Audio Processing*, vol.8, no.6, pp. 695-707, Nov. 2000.

[3] T. Hazen, "The Use of Speaker Correlation Information for Automatic Speech Recognition," Ph.D. Thesis, Mass.Inst.Technol., Cambridge, Jan. 1998.

[4] Zuoying Wang, Feng Liu, "Speaker Adaptation using Maximum Likelihood Model Interpolation," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing '99*, pp.753-756.1999.

[5] The NIST 1998 Speaker Recognition Evaluation Plan. [Online]. Available: http://www.nist.gov/speech/tests/spk/

[6] M.J.F. Gales, "The Generation and Use of Regression Class Trees for MLLR Adaptation," Tech-Report-263, Cambridge University, Aug. 1996.

[7] Yoshifumi Onishi, Ken-ichi Iso, "Speaker Adaptation by Hierarchical Eigenvoice," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing '2003*, pp. 576-579, 2003.

[8] H. Hermansky, N. Morgan, "RASTA Processing of Speech," *IEEE Trans. on Speech and Audio Processing*, vol.2, no.4, pp. 578~589, Oct. 1994.

[9] D. Reynolds, "Channel Robust Speaker Verification via Feature Mapping," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing '2003*, pp. 53-56, 2003.