Cluster Adaptive Training Weights as Features in SVM-Based Speaker Verification

Hao Yang[2], Yuan Dong[1,2], Xianyu Zhao[1], Jian Zhao[2], Liang Lu[2], Haila Wang[1]

[1]France Telecom Research & Development Center, Beijing, 100083, China
[2]Beijing University of Posts and Telecommunications, Beijing, 100876, China
{haoyang.hy, michaeljianzhao, luliang07}@gmail.com
{xianyu.zhao, yuan.dong, haila.wang}@orange-ft.com

Abstract

In this paper, we propose the use of cluster adaptive training (CAT) weights as features in support vector machine (SVM) based text-independent verification task. The speaker utterance is characterized by a vector of cluster weights, which are extracted during the cluster adaptive training process. The effects of the number of classes, which are obtained by partitioning the components of the model, and the number of clusters on the verification performance are investigated. To remove session variability due to influences of microphone, environment, etc, Nuisance Attribute Projection (NAP) is also evaluated. Experimental results in a NIST SRE 2006 task show that this CAT weights SVM system achieves comparable performance to a state-of-the-art cepstral GMM-UBM verification system, and their fusion can give further performance gains.

Index Terms: CAT, SVM, NAP, GMM-UBM, fusion

1. Introduction

For the task of text-independent speaker verification, the most prevalent framework is the Gaussian Mixture Model – Universal Background Model (GMM-UBM) framework [1], where the speaker model is constructed by Maximum a Posterior (MAP) adaptation of the means of the UBM. In recent years, lots of alternate speaker modeling methods have been proposed. Among these techniques, reference clusters or speakers based adaptation methods (e.g. Clustering Adaptive Training [2], Eigenvoice modeling [3], Reference Speaker Weighting [4], Anchor modeling [5], etc.) are studied extensively both in speech recognition [2, 3, 4] and speaker recognition [5, 6]. In the reference clusters based method, a model is built for each cluster, and then a new speaker model is constructed by a linear interpolation of all the cluster parameters. The aim of this method is to map the enrolled speaker to a new space expanded by reference clusters, in which there may be different discriminative capability.

Support vector machines (SVMs) have become one of the most popular classification techniques for speaker recognition, e.g. [8]–[11]. SVMs work on a high-dimensional feature space which is derived by a nonlinear mapping of the input space. To address the performance degradation introduced by session variability (e.g. microphone, environment, etc.), Nuisance Attribute Projection was developed in [10] to remove dimensions from the SVM expansion space that are irrelevant to the classification problem.

In this study, we investigated the use of CAT weights as features in SVM based speaker verification. In this method, the characteristic of a speaker is modeled using the CAT weights, which are stacked into a vector in the space spanned by a set of pre-selected reference clusters. NAP is then applied to reduce session variability in the cluster weight vectors. Fusion of this new system with conventional GMM-UBM system is also investigated in this study.

The remainder of this paper is organized as follows. In Section 2, we will review cluster adaptive training. In Section 3, we present the use of vectors of CAT weights vectors in SVM based speaker verification. In Section 4, we report experimental results in a NIST speaker recognition evaluation (SRE) 2006 task. Conclusion will be given at the end of this paper.

2. Cluster Adaptive Training

CAT is a popular rapid speaker adaptation algorithm which is conventionally used in speech recognition [2]. As for CAT, a set of clusters should be selected, and then a number of models, each corresponding to one particular cluster, are trained. CAT assumes that mean parameters of an enrolled speaker model are determined by a linear combination of these cluster means, while the Gaussian component variances and prior weights are left to be the same across all clusters.

To describe the speaker's characteristic more precisely, multiple groups of cluster weights could be used in CAT. In this fashion, the Gaussian components in a given model are classified into $G(G>1)$ groups according to certain predetermined criterion, and a separate set of cluster weights will be calculated for each group. For the *g*-th group of Gaussian components in a given cluster ***c***, $G_c^g = \left\{ w_{ci}^g, \mu_{ci}^g, \Sigma_{ci}^g; i = 1, \cdots, K \right\}$, all the Gaussian means are concatenated into a supervector

$$V_c^g = \left[\left(\mu_{c1}^g \right)^T, \left(\mu_{c2}^g \right)^T, \cdots, \left(\mu_{cK}^g \right)^T \right]^T \quad (1)$$

where μ_{ci}^g is the *i*-th Gaussian component in the *g*-th group of cluster ***c***.

Assume the supervector for the *g*-th group of components in a target speaker model ***s*** is Λ_s^g. As all the models, including models for clusters and models for speakers, share the same partition criterion, Λ_s^g can be denoted as a linear combination of the pre-selected clusters [2], which is

$$\Lambda_s^g = \sum_{c=1}^{C} \lambda_{sc}^g V_c^g \quad (2)$$

Therefore, mean vector of the *k*-th component in this group can be written as

$$\bar{\mu}_{sk}^{g} = \sum_{c=1}^{C} \lambda_{sc}^{g} \mu_{ck}^{g} \tag{3}$$

where λ_{sc}^{g} is the c-th element of the cluster weight vector for current group. The Maximum Likelihood (ML) training of cluster weight vector $\bar{\lambda}_{s}^{g} = \{\lambda_{s1}^{g}, \lambda_{s2}^{g}, ..., \lambda_{sC}^{g}\}$ is implemented through the Expectation-Maximum (EM) algorithm, the auxiliary function is defined as follows [2]-[4], ignoring the standard constants and terms independent of $\bar{\lambda}_{s}^{g}$:

$$\begin{aligned} &Q\left(\bar{\lambda}_{s}^{g}, M; \hat{\bar{\lambda}}_{s}^{g}, \hat{M}\right) \\ &= -\frac{1}{2}\sum_{t=1}^{T}\sum_{k=1}^{K}\gamma_{k}(t)(\bar{x}_{t} - \bar{\mu}_{sk}^{g})^{T}\Sigma_{sk}^{g^{-1}}(\bar{x}_{t} - \bar{\mu}_{sk}^{g}) \end{aligned} \tag{4}$$

where $\gamma_{k}(t)$ is the posterior probability of Gaussian component k at time t calculated using the old model parameters $\hat{\bar{\lambda}}_{s}^{g}$ and $\hat{M}$, $\bar{\mu}_{sk}^{g}$ and Σ_{sk}^{g} are the mean vector and covariance matrix of component k.

Let $\partial Q / \partial \lambda_{sc}^{g} = 0, c = 1, 2, \cdots, C$, we get the update equation for each $\lambda_{sc}^{g}, c = 1, 2, ..., C$.

$$\begin{aligned} &\sum_{t=1}^{T}\sum_{k=1}^{K}\gamma_{k}(t)\left(x_{t}\right)^{T}\Sigma_{sk}^{g^{-1}}\bar{\mu}_{sk}^{g} \\ &= \sum_{t=1}^{T}\sum_{k=1}^{K}\gamma_{k}(t)\sum_{c=1}^{C}\lambda_{sc}^{g}\left(\mu_{ck}^{g}\right)^{T}\Sigma_{sk}^{g^{-1}}\bar{\mu}_{sk}^{g} \ . \end{aligned} \tag{5}$$

3. SVM Based Speaker Verification Using Cluster Adaptive Training Weights

3.1. Support Vector Machines

SVM based speaker verification [7] has been increasingly popular these years. SVMs is a two-class classifier defined as

$$f(\mathbf{v}) = \sum_{i=1}^{M}\alpha_{i}k(\mathbf{v}, \overline{\mathbf{v}}_{i}) + d \tag{6}$$

where $k(\bullet, \bullet)$ is a kernel function. The d and $\{\alpha_{i}, \overline{\mathbf{v}}_{i}; i = 1, \cdots, M\}$ are obtained through a training process that maximizes the margin between two classes (positive vs. negative).

The kernel $k(\mathbf{v}_{1}, \mathbf{v}_{2})$ can be expressed as

$$k(\mathbf{v}_{1}, \mathbf{v}_{2}) = b(\mathbf{v}_{1})^{T} b(\mathbf{v}_{2}) \tag{7}$$

where $b(\bullet)$ is a mapping from the input feature space to the SVM expansion space.

3.2. SVM Kernel for SuperVector Constructed from CAT Weights

Suppose we have two sets of CAT weights, $\{\bar{\lambda}_{a}^{g}, g = 1, 2, ..., G\}$ and $\{\bar{\lambda}_{b}^{g}, g = 1, 2, ..., G\}$, which are calculated from two sentences ***a*** and ***b***. The approximate distance of these two sentences can be formulated as

$$d(a,b) = \sum_{g=1}^{G} w_{g}(\bar{\lambda}_{a}^{g} - \bar{\lambda}_{b}^{g})^{T} C_{g}^{-1}(\bar{\lambda}_{a}^{g} - \bar{\lambda}_{b}^{g}) \tag{8}$$

where w_{g} is the weight for the g-th group, and C_{g} is the covariance of weight vectors in g-th group. The use of weighted Mahalanobis distance could serve to place different emphasis on groups with different discriminative capability. (In this study, w_{g} is set to be equal for all groups.)

From the distance given above, we can get the kernel function for super vectors constructed from CAT weights:

$$\begin{aligned} k(\mathbf{v}_{a}, \mathbf{v}_{b}) &= \sum_{g=1}^{G} w_{g}(\bar{\lambda}_{a}^{g})^{T} C_{g}^{-1}(\bar{\lambda}_{b}^{g}) \\ &= \sum_{g=1}^{G}\left(\sqrt{w_{g}} C_{g}^{-\frac{1}{2}}\bar{\lambda}_{a}^{g}\right)^{T}\left(\sqrt{w_{g}} C_{g}^{-\frac{1}{2}}\bar{\lambda}_{b}^{g}\right) \\ &= b(\mathbf{v}_{a})^{T} b(\mathbf{v}_{b}) \end{aligned} \tag{9}$$

where

$$b(\mathbf{v}_{i})^{T} = \left\{\sqrt{w_{1}} C_{1}^{-\frac{1}{2}}\bar{\lambda}_{i}^{1}, ..., \sqrt{w_{G}} C_{G}^{-\frac{1}{2}}\bar{\lambda}_{i}^{G}\right\} \tag{10}$$

The kernel in eq (9) is linear, and the expansion from CAT weights supervector to SVM space can be calculated efficiently if we use only diagonal covariance matrix for each group.

In order to address performance degradation imposed by session variability, nuisance attribute projection method [9, 10] is proposed in SVM based speaker recognition. NAP aims to remove subspaces that cause variability in the kernel. NAP constructs a new kernel as

$$\begin{aligned} k(\mathbf{v}_{a}, \mathbf{v}_{b}) &= \left(Pb(\mathbf{v}_{a})\right)^{T}\left(Pb(\mathbf{v}_{b})\right) \\ &= \left((I - UU^{T})b(\mathbf{v}_{a})\right)^{T}\left((I - UU^{T})b(\mathbf{v}_{b})\right) \end{aligned} \tag{11}$$

where I is identity matrix and U is a matrix whose columns are composed of eigenvectors corresponding to the top M eigenvalues of the within-class covariance matrix [8], $\mathbf{W}$, which is calculated as follows:

$$\mathbf{W} = \sum_{j=1}^{J} p_{j}\Sigma_{j} \tag{12}$$

where p_{j} and Σ_{j} represent respectively the priori probability and covariance matrix of the j-th class, J is the total number of classes.

4. Experimental Results

In this section, we will report experimental results on SVM based speaker verification system using CAT weights as features. Section 4.1 presents the datasets used in our experiments. Section 4.2 reports the results of several CAT-SVM systems.

4.1. Datasets Description

Experiments were performed on the NIST2006 SRE corpus [13]. We focus on male part of the single-side 1 conversation train, single-side 1 conversation task, which contains 1570 true trials and 20561 false trials, while the female part is left to train fusion parameters.

The dataset used for cluster selection consisted of speech data of 268 male speakers and 349 female speakers chosen from both Switchboard-I and Switchboard-II.

The development set for background training consisted of 1790 conversation sides from NIST2004 SRE corpus (the single-side, 1 conversation train, single-side, 1 conversation

test part). This set was also used to estimate the within-class covariance matrix in NAP.

For CAT-SVM systems and GMM-UBM systems in this study, 13-dimensional PLP vectors were extracted from the silence removed speech signal every 10ms using 25ms window. Bandlimiting was performed by only retaining the filterbank outputs from the frequency range 300Hz-3400Hz. Cepstral features were processed with RASTA filtering to eliminate channel distortion. Delta, acceleration and triple-delta coefficients were then computed over ± 2 frames span and appended to the static coefficients, producing a 52 dimensional feature vector. Feature mapping and histogram equalization (HEQ) were applied to improve channel and noise robustness. Heteroscedastic linear discriminant analysis (HLDA) was then used to decorrelate the features and reduce the dimensionality from 52 to 51 (1 dimension is rejected as nuisance).

Gender-independent UBM with 2048 Gaussians was used in all the GMM-UBM baselines. Both speaker GMM models and cluster models were adapted from UBM by MAP-adaptation with the relevance fact set to be 16.

4.2 CAT-SVM System Results

The calculation of CAT weights is highly related to the cluster number, thus for a CAT-SVM system, it is important to determine an appropriate cluster number which can ensure a balance between system performance and computational cost. If the number of clusters is larger, it will make the computational cost incredible, while contributing marginal to further improvement of system performance. For this matter, the relationship between the number of clusters and system performance was firstly investigated. The cluster selection was based on principle component projection (PCA). Table 1 gives the results for different number of clusters in terms of both minimum detection cost function (DCF) and equal error rate (EER). DCF is the Bayesian risk function defined by NIST as DCF = 0.1 * Pr(miss) + 0.99 * Pr(false_alarm).

Table 1. *Results for systems with different number of clusters*

Cluster Number	EER	MinDCF
100	11.93%	0.519
200	10.44%	0.462
300	9.94%	0.434
400	9.66%	0.424
500	9.56%	0.420

From Table 1, we could find that system performance would become better if the number of clusters is increased. However, when the cluster number increases beyond 300, the improvement of system performance becomes marginal. In order to reach a compromise between performance and computational cost, 400 was used in the following experiments in this study.

As presented in Section 2, typically multiple groups of cluster weights are used in CAT [2] to model the speaker's character more precisely. However, to estimate the CAT weights accurately with limited training data in speaker verification, the group number can not be too large. Table 2 summarizes the results for different number of groups. The partition scheme used in this paper is similar to the clustering method in acoustic space described in [7]. As shown in Table 2, 8 is the optimum choice in the single-side 1 conversation train, single-side 1 conversation task. If the number of groups is too small, the speaker can not be modeled thoroughly; while too large the number will lead to imprecise CAT parameters estimated by insufficient training data.

In Figure 1, we compare CAT-SVM systems with state-of-the-art GMM-UBM systems. In Figure 1, the GMM-UBM baseline, labeled as "Cepstral GMM-UBM", is a standard MAP adaptation system without ATNorm [14]. "Cepstral GMM-UBM with ATNorm" stands for cepstral GMM-UBM system with ATNorm. The set of imposter speakers used for ATNorm is the same as the set used for cluster selection. The number of speaker adaptive cohort models for ATNorm is set to be 55. For CAT-SVM systems, 400 clusters and 8 groups were used, resulting in 400×8 CAT weights for each utterance.

Table 2. *Results for systems with different number of groups*

Group Number	EER	MinDCF
1	9.66%	0.424
4	8.58%	0.385
8	8.25%	0.374
12	8.36%	0.378

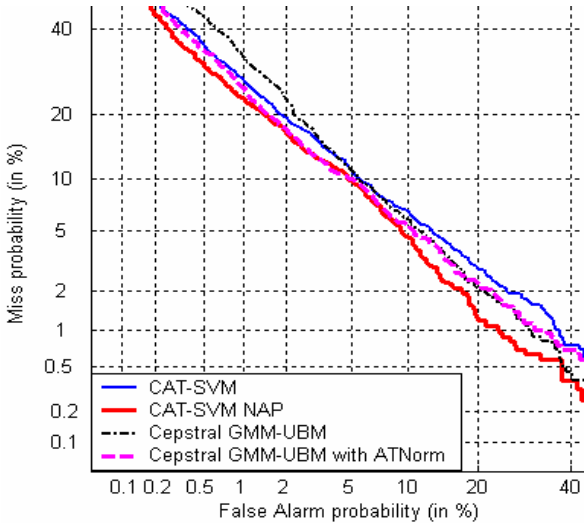

Figure 1: *Comparison of CAT-SVM systems with Cepstral GMM-UBM systems*

From Figure 1, it can be seen that performance of CAT-SVM baseline (named "CAT-SVM") is comparable to that of GMM-UBM baseline. NAP could provide significant performance improvement to CAT-SVM system. After reducing session variability in the weights supervector through NAP, the CAT weights becomes more stable. The result of CAT-SVM system with NAP is slightly better than performance of the GMM-UBM system with ATNorm.

Fusion of "CAT-SVM NAP" and "Cepstral GMM-UBM with ATNorm" is given in Figure 2, which was implemented using the logistic regression fusion developed by Brümmer [15], and the fusion parameters were trained on the female part of the single-side 1 conversation train, single-side 1 conversation task of the NIST2006 SRE. As show in Figure 2, the fusion system can provide further performance gains, the EER of the fusion system reduces from 7.13% to 5.67% when

compared with "Cepstral GMM-UBM with ATNorm", while the minimum DCF drops from 34.3×10^{-3} to 27.3×10^{-3}.

The complementary information may due to two factors. The first one is that mapping the speaker to a new space spanned by reference clusters may provide different discriminative capability compared with GMM-UBM system. The second factor may lie in that there is certain difference between the detection of SVM classifier and GMM-UBM classifier. Therefore, the combination of these two systems can give significant performance improvement.

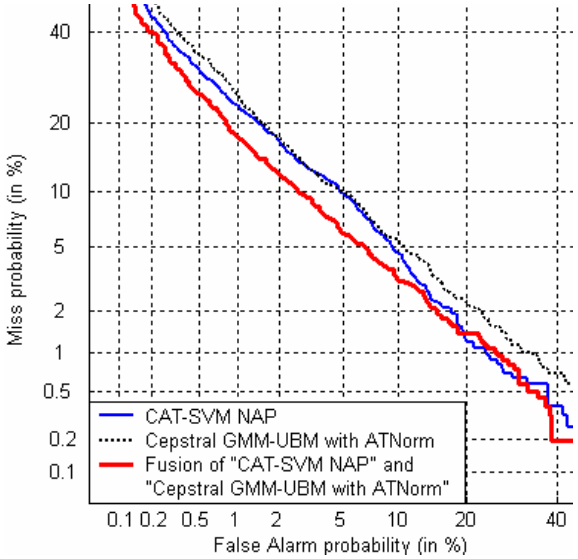

Figure 2: *Fusion of "CAT-SVM NAP" with "Cepstral GMM-UBM with ATNorm"*

5. Conclusion

Cluster adaptive training weights as features in SVM based speaker verification is investigated in this paper. By mapping an utterance used in training or testing phases onto a new space spanned by a set of properly selected reference clusters, the underlying speaker is modeled as a vector of cluster adaptive training weights. This enables the use of SVM as classifier for speaker verification purpose.

Experimental results on the NIST2006 SRE corpus show that with appropriate group number and cluster number, the CAT weights based verification system could give comparable performance with state-of-the-art Cepstral GMM-UBM systems. In addition, complementary information is provided by CAT weights based system, the fusion of these two systems can achieve additional performance gains. Further experiments with more extensive datasets and investigation on the distance weights mentioned in eq (8) are planned in the future.

6. 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] M. J. F. Gales, "Cluster Adaptive Training of Hidden Markov Models," *IEEE Trans. Speech and Audio Processing*. vol.8, no.6, pp. 417-428, July. 2000.

[3] 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.

[4] T. Hazen, "The use of speaker correlation information for automatic speech recognition," Ph.D. Thesis, Mass.Inst.Technol., Cambridge, Jan. 1998.

[5] D. Sturim, D. Reynolds, E. Singer, and J. P. Campbell, "Speaker indexing in large audio database using anchor models," in Proc. ICASSP'2001, pp. 429-432, 2001.

[6] Y. Mami, D. Charlet, "Speaker recognition by location in the space of reference speakers," *Speech Communication*, vol.48, pp. 127-141, 2006.

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

[8] A. O. Hatch, A. Stolcke, "Generalized linear kernels for one-versus-all classification: application to speaker recognition," in *Proc. ICASSP'2006*, 2006.

[9] A. O. Hatch, S. Kajarekar and A. Stolcke, "Within-Class Covariance Normalization for SVM-based Speaker Recognition," in *Proc. ICSLP'2006*, 2006.

[10] A. Solomonoff, W. M. Campbell and I. Boardman, "Advances in channel compensation for SVM speaker recognition," in *Proc. ICASSP'2005*, 2005.

[11] Campbell, W.M., Sturim, D.E., Reynolds, D.A., Solomonoff, A., "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation", in *Proc. of ICASSP' 2006*, 2006.

[12] R. Collobert, S. Bengio, "SVMTorch: Support vector machines for large–scale regression problems," *Journal of Machine Learning Research*, vol. 1, pp. 143-160, 2001.

[13] "The NIST 2006 speaker recognition evaluation plan," http://www.nist.gov/speech/tests/spk/spk/2006/ .

[14] D. E. Sturim, D. A. Reynolds, "Speaker adaptive cohort selection for TNorm in text-independent speaker verification," in *Proc. ICASSP*, pp. 741-744, 2005.

[15] N. Brümmer, J. Preez, "Application-independent evaluation of speaker detection," in *Computer, Speech and Language*, vol. 20, pp. 230-275, 2006.