SOFT FRAME MARGIN ESTIMATION OF GAUSSIAN MIXTURE MODELS FOR SPEAKER RECOGNITION WITH SPARSE TRAINING DATA

Yan Yin and Qi Li

Li Creative Technologies, Inc.
Florham Park, New Jersey 07932, USA
{yyin,li}@licreativetech.com

ABSTRACT

Discriminative Training (DT) methods for acoustic modeling, such as MMI, MCE, and SVM, have been proved effective in speaker recognition. In this paper we propose a DT method for GMM using soft frame margin estimation. Unlike other DT methods such as MMI or MCE, the soft frame margin estimation attempts to enhance the generalization capability of GMM to unseen data in case the mismatch exists between training data and unseen data. We define an objective function which integrates multi-class separation frame margin and loss function, both as functions of GMM likelihoods. We propose to optimize the objective function based on a convex optimization technique, semidefinite programming. As shown in our experimental results, the proposed soft frame margin discriminative training with semidefinite programming optimization (SFME-SDP) is very effective for robust speaker model training when only limited amounts of training data are available.

Index Terms— robust speaker recognition, soft margin discriminative training, gaussian mixture model

1. INTRODUCTION

Gaussian Mixture Model (GMM) has been widely used as the probabilistic model in most automatic speaker recognition systems [1], and the parameters can be estimated by the EM algorithm under the ML objective. Discriminative training (DT) of GMM has been proved an effective way to improve the performance from ML training, e.g. maximum mutual information (MMI) [2], and minimum classification error (MCE) [3]. An issue with traditional DT approaches is the limited capability of performance gain carryover from training data to unseen test data. The power to deal with possible mismatches between the training and testing conditions can often be measured by the generalization ability of the machine learning algorithms [4]. To address this, the concept of a large margin classifier has been developed. The support vector machines (SVMs) [4] developed under the concept has demonstrated the generalization ability and has been applied successfully in speaker recognition [5, 6].

Inspired by SVM, recently many attempts have been made to incorporate the principle of large margin into hidden Markov model (HMM) training in automatic speech and language recognition. For speech recognition, a large margin estimation method is proposed to maximize the minimum margin between HMMs [7]. A soft margin estimation of HMMs is proposed to minimize the empirical loss and maximize the separation margin together [8]. Some other attempts, such as LMMCE and Boosted MMI [9, 10], embedded the discriminative margin concept into traditional DT methods. A soft margin estimation is proposed to maximize the minimum margin between GMMs for spoken language recognition [11]. However, to our knowledge, few such attempts have been made in speaker recognition. All of the above have motivated our idea of soft frame margin discriminative training of GMM for speaker recognition.

Convex optimization has been applied successfully to HMM parameter estimation for DT methods. In [12], semidefinite programming (SDP) is used to estimate HMM parameters of a formulated large margin estimation method. In [13], second order cone programming (SOCP) is applied successfully for parameter optimization of formulated large margin estimation method. A convex optimization method is used to jointly optimize the mean and variance of large margin HMMs [14]. All these works have proved that convex optimization is more effective than traditional Extended Baum Welch (EBW) and Generalized Probabilistic Descent (GPD) methods.

In this paper we propose a soft frame margin estimation of GMM with SDP optimization (SFME-SDP) for speaker recognition. The objective function of SFME-SDP integrates the maximization of the frame margins over correct data near decision boundary and minimization of a loss function over error data. We propose to optimize the objective function with the SDP convex optimization method. Also in this paper, we focus on evaluating the proposed SFME-SDP method under the data sparseness condition. We conducted experiments on NTIMIT to evaluate the proposed SFME-SDP approach. The paper is organized as follows. Section 2 describes the concept of SFME of GMM. Section 3 presents the GMM parameter estimation with SDP convex optimization. Section 4 presents experimental setup and results. Finally, conclusions are drawn in section 5.

2. SOFT FRAME MARGIN GMM

Suppose there are K target speakers to be recognized. The training data set consists of a collection of speech segments $\mathcal{D} = \{X_n^k; n = 1, 2, \cdots, N, k = 1, 2, \cdots, K\}$, where each speech segment is a sequence of feature vector $X_n^k = \{x_{nt}^k; t = 1, 2, \cdots, T_n\}$. The GMMs for all K speakers are denoted as $\Lambda = \{\lambda_k, k = 1, 2, \cdots, K\}$. The frame level multi-class separation margin of speech segment X_n^k from speaker k is defined as,

This work was supported by US AFRL under the contract number FA8750-08-C0028.

$$d(X_n^k) = \frac{1}{T_n}\left[\mathcal{P}(X_n^k|\lambda_k) - \max_{j\in\Omega\ j\neq k}\mathcal{P}(X_n^k|\lambda_j)\right] = \min_{j\in\Omega\ j\neq k}\frac{1}{T_n}\left[\mathcal{P}(X_n^k|\lambda_k) - \mathcal{P}(X_n^k|\lambda_j)\right] \quad (1)$$

where Ω denotes a set of all speakers, and $\mathcal{P}(X_n^k|\lambda_j)$ denotes the log-domain likelihood scores of speech segment X_n given speaker model λ_j. From the definition, if $d(X_n^k) \leq 0$, X_n^k is incorrectly recognized by the GMM set Λ; if $d(X_n^k) > 0$, X_n^k is correctly recognized by the GMM set Λ. A subset $\mathcal{S}$ of $\mathcal{D}$ is defined as,

$$\mathcal{S} = \{X_n^k \mid X_n^k \in \mathcal{D} \text{ and } 0 \leq d(X_n^k) \leq \epsilon\} \quad (2)$$

where $\epsilon > 0$ is a pre-set positive number. $\mathcal{S}$ is called *support token set* and each speech segment X_n^k in $\mathcal{S}$ is called a support token. Each support token has small positive margin, thus is correctly identified and near the classification boundary. Furthermore, another subset $\mathcal{E}$ of $\mathcal{D}$ is defined as:

$$\mathcal{E} = \{X_n^k \mid X_n^k \in \mathcal{D} \text{ and } d(X_n^k) < 0\} \quad (3)$$

where $\mathcal{E}$ is called *error token set* and each speech segment in $\mathcal{E}$ is called a error token. Each error token has negative margin, thus is misclassified. To achieve better generalization power, it is desirable to adjust decision boundaries to make all support tokens as far from the decision boundaries as possible. While maximizing the separation margin, it is desirable to minimize the total error caused by the error token set $\mathcal{E}$. Suppose we define the error function $\xi(X_n^k)$ for a speech segment X_n^k in $\mathcal{E}$ as follows:

$$\xi(X_n^k) = \frac{1}{|\mathcal{E}|}\sum_{j\in\Omega}\left[\mathcal{P}(X_n^k|\lambda_j) - \mathcal{P}(X_n^k|\lambda_k)\right] \quad (4)$$

This leads to estimate the GMM models based on the objective of integrating minimum separation frame margin maximization and the average error minimization, which is named as *Soft Frame Margin Estimation* (SFME):

$$\tilde{\Lambda} = \arg\min_{\Lambda}\left[-\min_{X_n^k\in\mathcal{S}} d(X_n^k) + \frac{\eta}{|\mathcal{E}|}\sum_{X_n^k\in\mathcal{E}}\xi(X_n^k)\right] \quad (5)$$

where $\eta > 0$ is a pre-set positive constant to balance contribution from the minimum margin and the average error.

A margin term ρ is introduced as a common lower bound to represent the *min* part of all margin terms in 5. Also it is beneficial to impose a locality constraint on model parameters Λ to ensure that parameters do not deviate too much from their initial or current values. The locality constraint can be quantitatively computed based on relaxed Kullback-Leibler divergence (KLD). As a result, The constrained SFME problem is formulated as a minimization problem,

$$\tilde{\Lambda} = \arg\min_{\Lambda,\rho}\left[-\rho + \frac{\eta}{|\mathcal{E}||\Omega|}\cdot \sum_{X_n^k\in\mathcal{E}, j\in\Omega}\Big(\mathcal{P}(X_n^k|\lambda_j) - \mathcal{P}(X_n^k|\lambda_k)\Big)\right] \quad (6)$$

subject to:

$$\mathcal{P}(X_n^k|\lambda_j) - \mathcal{P}(X_n^k|\lambda_k) \leq -\rho\cdot T_n \quad (7)$$
$$\forall X_n^k \in \mathcal{S} \text{ and } j \in \Omega \text{ and } j \neq k$$
$$\|\Lambda - \Lambda^{(0)}\|_2 \leq \theta^2 \quad (8)$$
$$\rho \geq 0 \quad (9)$$

3. PARAMETER ESTIMATION WITH CONVEX OPTIMIZATION

Among all the optimization methods for discriminative training, convex optimization has been proved to be effective. It has been shown that SDP, although has very high computational complexity, is the most successfully convex optimization method for tasks with small model size [12]. In this work we adopt SDP to solve the constrained SFME problem formulated in section 2. The standard SDP problem is illustrated as,

$$\textit{Minimize} \quad \sum_{j=1}^{p} C_j \cdot X_j \quad (10)$$

subject to

$$\sum_{j=1}^{p} B_{ij}\cdot X_j \leq b_i, \quad i = 1,\cdots,m \quad (11)$$
$$X_j \succeq 0 \quad (12)$$

where $X_j \succeq 0$means each variable X_j is a positive semidefinite matrix. A_{ij}, C_j are real symmetric matrices with the same dimension as X_j, b_i is a scalar constant, and $X \cdot Y$ denotes the inner product of two symmetric matrices.

In our work, we only consider optimizing the mean parameters of GMM and leave other parameters unchanged. The formulation can be extend to deal with other GMM parameters as well. Suppose there are totally $\mathcal{L}$ Gaussian in the model set Λ, the normalized mean vector $\tilde{\boldsymbol{\mu}}_l$ for all $l \in \{1,\cdots,\mathcal{L}\}$ is defined as:

$$\tilde{\boldsymbol{\mu}}_l = \Big(\frac{\mu_{l1}}{\sigma_{l1}};\ \frac{\mu_{l2}}{\sigma_{l2}};\ \ldots;\ \frac{\mu_{lD}}{\sigma_{lD}}\Big). \quad (13)$$

Then, we construct a matrix U by concatenating all normalized Gaussian mean vectors as columns:

$$U = (\tilde{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\mu}}_2, \ldots, \tilde{\boldsymbol{\mu}}_{\mathcal{L}}). \quad (14)$$

when using the top Gaussian path to approximate the sum of all paths, the approximated GMM likelihood is formulated as,

$$\begin{aligned}\mathcal{P}(X_n^k|\lambda_j) &= c_j - \frac{1}{2}\sum_{t=1}^{T_n}\sum_{d=1}^{D}\frac{(x_{ntd}^k - \mu_{j_t^* d})^2}{\sigma_{j_t^* d}^2}\\ &= c_j - \frac{1}{2}\sum_{t=1}^{T_n}(\tilde{\mathbf{x}}_{nt}^k - \tilde{\boldsymbol{\mu}}_{j_t^*})'(\tilde{\mathbf{x}}_{nt}^k - \tilde{\boldsymbol{\mu}}_{j_t^*})\\ &= c_j - \frac{1}{2}\sum_{t=1}^{T_n}(\tilde{\mathbf{x}}_{nt}^k; \mathbf{e}_{j_t^*})'(I_D, U)'(I_D, U)(\tilde{\mathbf{x}}_{nt}^k; \mathbf{e}_{j_t^*})\\ &= -A_j\cdot Z + c_j \end{aligned} \quad (15)$$

where $p = \{(j_t^*)_1^{T_n}, j_t^* \in \{1,\cdots,\mathcal{L}\}\}$ denotes the viterbi Gaussian path for feature X_n^k and GMM model λ_j. e_i is a vector

with -1 at the i-th position, and zero everywhere else. I_D is D-dimensional identity matrix. $\tilde{\mathbf{x}}_{nt}^k$ denotes normalized feature vector,

$$\tilde{\mathbf{x}}_{nt}^k := \Big(\frac{x_{nt1}^k}{\sigma_{j_t^* 1}};\ \frac{x_{nt2}^k}{\sigma_{j_t^* 2}};\ \ldots;\ \frac{x_{ntD}^k}{\sigma_{j_t^* D}}\Big) \tag{16}$$

$$A_j = \frac{1}{2}\sum_{t=1}^{T}(\tilde{\mathbf{x}}_{nt}^k; \mathbf{e}_{j_t^*})(\tilde{\mathbf{x}}_{nt}^k; \mathbf{e}_{j_t^*})' \tag{17}$$

$$Z = \begin{pmatrix} I_D & U \\ U' & Y \end{pmatrix} \qquad Y = U'U. \tag{18}$$

Similarly the average error in (6), the margin constraint in (7), and the locality constraint in (8) are formulated as,

$$\mathcal{P}(X_n^k|\lambda_j) - \mathcal{P}(X_n^k|\lambda_k) = A_{kj}\cdot Z - c_{kj} \le -\rho\cdot T_n \tag{19}$$

$$\frac{\eta}{|\mathcal{E}||\Omega|}\sum_{X_n^k\in\mathcal{E}}\sum_{j\in\Omega}\Big(\mathcal{P}(X_n^k|\lambda_j) - \mathcal{P}(X_n^k|\lambda_k)\Big) = E\cdot Z \tag{20}$$

$$\|\Lambda - \Lambda^0\|_2 = \sum_{l=1}^{\mathcal{L}}\sum_{d=1}^{D}\frac{(\mu_{ld} - \mu_{ld}^{(0)})^2}{\sigma_{ld}^2} = Q\cdot Z \tag{21}$$

where $A_{kj} = A_k - A_j$, $E = \frac{\eta}{|\mathcal{E}||\Omega|}\sum_{X_n^k\in\mathcal{E}}\sum_{j\in\mathcal{E}} A_{kj}$, and $Q = \sum_{l=1}^{\mathcal{L}}(\tilde{\boldsymbol{\mu}}_l^{(0)}; \mathbf{e}_l)(\tilde{\boldsymbol{\mu}}_l^{(0)}; \mathbf{e}_l)'$. To formulate the constrained SFME problem into an SDP problem, all the constraints have to be convex. Then the relaxation is made for the constraint in (18),

$$Y = U'U \overset{relaxation}{\rightarrow} Y - U'U \succeq 0 \tag{22}$$

with which the non-convex constraint in (18) is relaxed to the convex constraint $Z = \begin{pmatrix} I_D & U \\ U' & Y \end{pmatrix} \succeq 0$. Finally the constraint SFME problem is formulated as an SDP problem and named as SFME-SDP,

$$\tilde{\Lambda} = \arg\min_{\Lambda,\ \rho} \quad -\rho \ +\ E\cdot Z \tag{23}$$

subject to

$$A_{kj}\cdot Z + T_n\cdot\rho \le c_{kj} \tag{24}$$
$$\forall X_n^k \in \mathcal{S} \text{ and } j\in\Omega \text{ and } j\neq k$$

$$Q\cdot Z \le \mathcal{L}Dr^2 \tag{25}$$

$$Z\succeq 0,\quad Z_{1:D,1:D} = I_D,\quad \rho\ge 0 \tag{26}$$

where locality constraint upper bound θ^2 in (8) is replaced by $\mathcal{L}Dr^2$ since the scaled r is much easier to tune in experiments.

4. EXPERIMENTS

The NTIMIT corpus is used to evaluate the effectiveness of the proposed SFME-SDP approach for robust speaker model training with sparse training data. A 168 speaker (112 males, 56 females) identification task from NTIMIT, referred to as NTIMIT168, is configured as the test set. In order to conduct parameter tuning and system optimization, a separate development set consisting of 38 speakers is used as our development set, which is referred to as NTIMIT38. For each speaker in both the NTIMIT168 and the NTIMIT38 tasks, 8 utterances are used for training, and 2 utterances are used for evaluation. The average duration of each test segment is about 3 seconds.

An SDP optimization problem is formulated based on the proposed SFME-SDP approach, open source convex optimization software DSDP [15] is used for the optimization of the formulated problem. All the system trainings are performed on the NTIMIT38 development set. The system parameters are tuned towards achieving the best performance on NTIMIT38 evaluation data. Then the tuned parameter settings are directly carried over to the NTIMIT168 test set to set up the system using NTIMIT168 training data, the performance of which on NTIMIT168 evaluation data is referred to as test performance. First, MFCC features are generated as the front-end for all systems. Then the GMM-UBM baseline system described in [1] is trained. The GMM-UBM baseline ID accuracy on NTIMIT38 development set is 80.26%. With the GMM-UBM baseline being used as the seed model, GMM speaker models based on the proposed SFME-SDP are trained and named as GMM-SFME-SDP. To compare the proposed SFME-SDP with other similar approaches, MMI and SVM are also evaluated. Similar to the GMM-SFME-SDP system, the GMM-MMI system is trained with the use of the GMM-UBM baseline as the seed model. For SVM training as described in [16], SVMTorch is used to train the SVM classifier. Gaussian kernel is used for the SVM classifier. We fine-tuned the GMM-MMI and SVM systems to achieve the best performance.

In this experiment, we slightly modify the selection of support token set defined in (2). The support token set is selected by including the top N correctly identified data closest to the decision boundary instead of the use of ϵ. Also, we realize that instead of imposing constraints in (24) for all other competing speakers to the SFME-SDP problem, it is sufficient to include those for the top M most confusable speaker candidates. Then we fine-tuned the three critical parameters: N, M, and r. Table 1 illustrates the effect of M on the GMM-SFME-SDP system ID accuracy, based on which M over 10 does not give any further gain. Table 2 shows the system performance for various N settings. Finally locality constraint threshold r is tuned based on the optimal top N and top M values, which is shown in Table 3. System comparison is illustrated in Figure 1. Given limited data, all systems improve the development data performance over the GMM-UBM baseline, while GMM-MMI performance drops quickly. Among all the systems GMM-SFME-SDP significantly outperforms the other two. The optimal parameter settings tuned with NTIMIT38 are carried

Table 1. GMM-SFME-SDP system performance with various top M settings for NTIMIT38 development set

	M = 5	M = 10	M = 15
Accuracy	82.90%	84.21%	84.21%

Table 2. GMM-SFME-SDP system performance with various top N settings for NTIMIT38 development set

	N = 20	N = 30	N = 40
Accuracy	81.58%	84.21%	82.90%

Table 3. The effect of locality constraint threshold r on GMM-SFME-SDP system ID accuracy for NTIMIT38 development set

	r = 0.02	r = 0.04	r = 0.06
Accuracy	84.21%	86.84%	84.21%

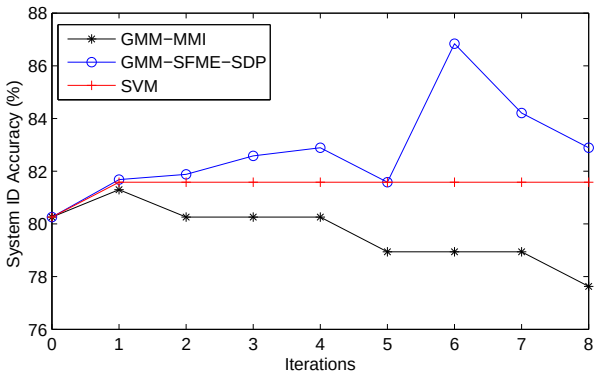

Fig. 1. Comparison of GMM-MMI, SVM, and GMM-SFME-SDP: SVM is not iterative approach, only one iteration SVM training is performed

over to the NTIMIT168 test set for the setup of all systems. The GMM-UBM baseline ID accuracy for the NTIMIT168 test set is 66.7%. The NTIMIT168 test data identification accuracy for various systems are listed in Table 4. When generilized to test set, MMI only maintains marginal improvement over baseline, mainly due to the sparseness of available training data. GMM-SFME-SDP, on the other hand, still generalize pretty well to test set, a 10% relative improvement over GMM-MMI. SVM performance on test set is not as good as expected, very likely due to the use of the one-versus-other approach instead of the pairwise one-versus-one approach.

Table 4. System comparison of GMM baseline (GMM-EM), GMM-MMI, SVM, and GMM-SFME-SDP on NTIMIT168 test set

Accuracy			
GMM-EM	GMM-MMI	SVM	GMM-SFME-SDP
66.7%	66.9%	64.8%	70.2%

5. CONCLUSIONS

In this paper we proposed the SFME method for speaker recognition. We introduced the SDP convex optimization for the formulated SFME problem. The proposed SFME-SDP methods greatly outperform other discriminative training methods such as MMI under data sparseness condition. We realized more experiments with SVM system is needed before we compare the SFME-SDP and SVM. Following this work we will set up SVM identification system based on the pairwise one-versus-one method, and even the hybrid GMM/SVM system. Also we will evaluate the proposed SFME-SDP method under sufficient training data condition. Other future works include refining the proposed SFME method with frame and utterance selection, and investigating other convex optimization techniques.

6. REFERENCES

[1] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," in *Digital Signal Processing*, 2000, vol. 10, pp. 19–41.

[2] L.R. Bahl, P.F. Brown, P.V. De Souza, and R.L. Mercer, "Maximum mutual information estimation of hidden markov model parameters for speech recognition," in *Proc. of IEEE (ICASSP86)*, Tokyo, Japan, April 1986, pp. 290–294.

[3] O. Siohan, A.E. Rosenberg, and S. Parthasarathy, "Speaker identification using minimum classification error training," in *Proc. of IEEE*, Seattle, WA, May 1998, pp. 109–112.

[4] V.N. Vapnik, *The nature of statistical learning theory*, Springer, New York, NY, 1995.

[5] M. Schmidt and H. Gish, "Speaker identification via support vector classifiers," in *Proc. of IEEE*, Atlanta, GA, May 1996, pp. 105–108.

[6] S. Fine, J. Navratil, and R.A. Gopinath, "A hybrid GMM/SVM approach to speaker identification," in *Proc. of IEEE*, Salt Lake City, UT, May 2001, pp. 417–420.

[7] H. Jiang, X. Li, and C. Liu, "Large margin hidden markov models for speech recognition," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1584–1595, September 2006.

[8] J. Li, M. Yuan, and C.H. Lee, "Soft margin estimation of hidden markov model parameters," in *Proc. of International Conference Spoken Language Processing*, Pittsburgh, USA, 2006, pp. 2422–2425.

[9] D. Yu, L. Deng, X. He, and A. Acero, "Use of incrementally regulated discriminative margins in MCE training for speech recognition," in *Proc. of International Conference Spoken Language Processing*, Pittsburgh, USA, 2006, pp. 2418–2421.

[10] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, "Boosted MMI for model and feature space discriminative training," in *Proc. of IEEE*, Las Vegas, NV, 2008, pp. 4057–4060.

[11] B. Ma, H. Li, and D. Zhu, "Soft margin estimation of gaussian mixture model parameters for spoken language recognition," in *Proc. of IEEE*, Dallas, TX, March 2010, pp. 4990–4993.

[12] H. Jiang and Y. Yin, "A fast optimization method for large margin estimation of HMMs based on second order cone programming," in *Proc. of Interspeech 2007*, Antwerp, Belgium, 2007, pp. 34–37.

[13] Y. Yin and H. Jiang, "A compact semidefinite programming (SDP) formulation for large margin estimation of HMMs in speech recognition," in *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Kyoto, Japan, December 2007, pp. 312 – 317.

[14] T.H. Chang, Z.Q. Luo, L. Deng, and C.Y. Chi, "A convex optimization method for joint mean and variance parameter estimation of large margin CDHMM," in *Proc. of IEEE*, Las Vegas, NE, April 2008, pp. 4053–4056.

[15] S. J. Benson and Y. Ye, "Dsdp5: Software for semidefinite programming," Tech. Rep. ANL/MCS-P1289-0905, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, September 2005.

[16] V. Wan and W. M. Campbell, "Support vector machines for speaker verification and identification," in *Proc. of IEEE Workshop Neural Networks for Signal Processing*, Sydney, Australia, December 2000, pp. 775–784.