Fast training of Large Margin diagonal Gaussian mixture models for speaker identification

Reda Jourani, Khalid Daoudi, Régine André-Obrecht, Driss Aboutajdine

Fast training of Large Margin diagonal Gaussian mixture models for speaker identification

Reda Jourani [1,3], Khalid Daoudi [2], Régine André-Obrecht [1] and Driss Aboutajdine [3]
[1] SAMoVA Group, IRIT - UMR 5505 du CNRS
University Paul Sabatier, 118 Route de Narbonne, F-31062 Toulouse Cedex 9, France
[2] INRIA Bordeaux-Sud Ouest
351, cours de la libération. 33405 Talence. France
[3] Laboratoire LRIT. Faculty of Sciences, Mohammed 5 Agdal University
4 Av. Ibn Battouta B.P. 1014 RP, Rabat, Morocco
{jourani, obrecht}@irit.fr, khalid.daoudi@inria.fr, aboutaj@fsr.ac.ma

Abstract—**Gaussian mixture models (GMM) have been widely and successfully used in speaker recognition during the last decades. They are generally trained using the generative criterion of maximum likelihood estimation. In an earlier work, we proposed an algorithm for discriminative training of GMM with diagonal covariances under a large margin criterion. In this paper, we present a new version of this algorithm which has the major advantage of being computationally highly efficient. The resulting algorithm is thus well suited to handle large scale databases. We carry out experiments on a speaker identification task using NIST-SRE'2006 data and compare our new algorithm to the baseline generative GMM using different GMM sizes. The results show that our system significantly outperforms the baseline GMM in all configurations, and with high computational efficiency.**

Keywords-component; Large margin training; Gaussian mixture models; discriminative learning; speaker recognition; speaker identification

I. Introduction

Most of state-of-the-art speaker recognition systems rely on the generative training of Gaussian Mixture Models using maximum likelihood estimation and maximum a posteriori estimation [1]. This generative training does not however directly optimize the classification performance. For this reason, discriminative training approaches have been an interesting and valuable alternative to address directly the classification problem [2], and lead generally to better performances than generative methods. For instance, Support Vector Machines (SVM) combined with GMM supervectors are among state-of-the-art approaches in speaker recognition [3].

Recently a new discriminative approach for multiway classification has been proposed, the Large Margin Gaussian mixture models (LM-GMM) [4]. The latter have the same advantage as SVM in terms of the convexity of the optimization problem to solve. However they differ from SVM because they draw nonlinear class boundaries directly in the input space, and thus no kernel trick/matrix is required. While LM-GMM (and their LM-HMM extension [5]) have been used in speech recognition, they have not been used in speaker recognition (to the best of our knowledge). In an earlier work [6], we proposed a simplified version of LM-GMM which exploit the fact that traditional GMM systems use diagonal covariances and only the mean vectors are *maximum a posteriori* (MAP) adapted. The resulting training algorithm is more efficient than the original one, however we found that it is still not efficient enough to process large scale databases such as in NIST'SRE campaigns [7].

In this paper, we propose a new and fast training algorithm which is suited for such large scale applications. To do so, we exploit the fact that in general not all the components of the GMM are involved in the decision process, but only the *k*-best scoring components. We also exploit the property of correspondence between the MAP adapted GMM mixtures and the *Universal Background Model* (UBM) mixtures. We carry out experiments on a speaker identification task using NIST-SRE data. The results show that our new algorithm is not only highly efficient but also significantly outperforms the baseline GMM.

The paper is organized as follows. After a brief overview on Large-Margin GMM training with diagonal covariances in section 2, we describe our new fast training algorithm in section 3. Experimental results are then reported in section 4.

II. Overview on Large Margin GMM with diagonal covariances (LM-dGMM)

Most of state-of-the art speaker recognition systems use diagonal-covariances GMM. In these GMM based speaker recognition systems, a speaker-independent *world model* or UBM is first trained with the EM algorithm [8] from tens or hundreds of hours of speech data gathered from a large number of speakers. The background model represents speaker-independent distribution of the feature vectors. When enrolling a new speaker to the system, the parameters of the UBM are adapted to the feature distribution of the new speaker. The adapted model is then used as the model of that speaker. It is possible to adapt all the parameters, or only some of them from the background model. Traditionally, in the GMM-UBM approach, the target speaker GMM is derived from the UBM model by updating only the mean parameters using a MAP

algorithm [1], while the (diagonal) covariances and the weights remain unchanged.

Making use of this assumption of diagonal covariances, we proposed in [6] a simplified algorithm to learn GMM with a large margin criterion. This algorithm has the advantage of being more efficient than the original LM-GMM one [4], [9] while it still yielded similar or better performances on a speaker identification task. In our Large Margin diagonal GMM (LM-dGMM), the first (initialization) step is to model each class (speaker) c by a GMM with M mixtures using MAP adaptation of the UBM. The m^{th} Gaussian is parametrized by a mean vector μ_{cm}, a diagonal covariance matrix $\Sigma_m = diag(\sigma_{m1}^2,...,\sigma_{mD}^2)$ where D is the input space dimension, and a non negative scalar factor $\theta_m = \frac{1}{2}\left(D\log(2\pi)+\log\left|\Sigma_m\right|\right) - \log(w_m)$ which corresponds to the weight of the Gaussian.

Then, let $\left\{x_{nt}\right\}_{t=1}^{T_n}$ $\left(x_{nt}\in\mathbb{R}^D\right)$ be the T_n feature vectors of the n^{th} segment (i.e. n^{th} speaker training data). For each x_{nt} belonging to the class y_n, $y_n\in\{1,2,...,C\}$ where C is the total number of classes, we determine the index m_{nt} of the Gaussian component of the GMM modeling the class y_n which has the highest posterior probability. This index is called *proxy label*. For each example x_{nt}, the goal of the training algorithm is to force the log-likelihood of its proxy label Gaussian m_{nt} to be at least one unit greater than the log-likelihood of each Gaussian component of all competing classes. That is, given the training examples $\left\{\left(x_{nt},y_n,m_{nt}\right)\right\}_{n=1}^{N}$, the LM-dGMM criterion is given as:

$$\begin{aligned}&\forall\, c\neq y_n,\ \forall m,\\ &d(x_{nt},\mu_{cm})+\theta_m \ \geq\ 1+d(x_{nt},\mu_{y_n m_{nt}})+\theta_{m_{nt}},\end{aligned} \tag{1}$$

where

$$d(x_{nt},\mu_{cm}) = \sum_{i=1}^{D}\frac{(x_{nti}-\mu_{cmi})^2}{2\sigma_{mi}^2}\ .$$

Then, these multiple constraints are fold into a single one using the softmax inequality

$$\min_m a_m \ \geq\ -\log\sum_m e^{-a_m}\ .$$

In a segmental training scheme, the LM-dGMM criterion is written as:

$$\begin{aligned}&\forall\, c\neq y_n,\\ &\frac{1}{T_n}\sum_{t=1}^{T_n}\left(-\log\sum_{m=1}^{M}\exp(-d(x_{nt},\mu_{cm})-\theta_m)\right)\\ &\geq\ 1\ +\ \frac{1}{T_n}\sum_{t=1}^{T_n}d(x_{nt},\mu_{y_n m_{nt}})+\theta_{m_{nt}}.\end{aligned} \tag{2}$$

The loss function to minimize for LM-dGMM is then given by:

$$\begin{aligned}\mathbf{L} \ =\ \sum_{n=1}^{N}\sum_{c\neq y_n} max(\,0,1+\frac{1}{T_n}\sum_{t=1}^{T_n}(\,d(x_{nt},\mu_{y_n m_{nt}})\\ +\ \theta_{m_{nt}}+\log\sum_{m=1}^{M}\exp(-d(x_{nt},\mu_{cm})-\theta_m))).\end{aligned} \tag{3}$$

III. LM-DGMM TRAINING WITH *K*-BEST GAUSSIANS

A. Description of the new LM-dGMM training algorithm

Despite the fact that our LM-dGMM is computationally much faster than the original LM-GMM of [4], [9], we still encountered efficiency problems when dealing with high number of Gaussian mixtures. Indeed, even for an easy 50 speakers identification task as the one presented in [6], we could not run the training in a relatively short time with our current implementation. This would imply that large scale applications such as NIST-SRE would be infeasible in reasonable time. In order to develop a fast training algorithm which could be used in large scale applications, we propose to drastically reduce the number of constraints to satisfy in (2). By doing so, we would drastically reduce the computational complexity of the loss function and its gradient, which are the quantities responsible for most of the computational time. To achieve this goal we propose to use another property of state-of-the-art GMM systems, that is, decision is not made upon all mixture components but only using the *k*-best scoring Gaussians.

In other words, for each x_n and each c, instead of summing over the M mixtures in the left side of (2), we would sum only over the k Gaussians with the highest posterior probabilities selected using the GMM of class c. In order to further improve efficiency and reduce memory requirement, we exploit the property reported in [1] about correspondence between MAP adapted GMM mixtures and UBM mixtures. We use the UBM to select one unique set S_{nt} of k-best Gaussian components per frame x_{nt}, instead of $(C\text{-}1)$ sets. This leads to a $(C\text{-}1)$ times faster and less memory consuming selection. Thus, the higher the number of target speakers is, the greater computation and memory saving is.

More precisely, we now seek mean vectors μ_{cm} that satisfy the large margin constraints in (4) :

$$\begin{aligned}&\forall\, c\neq y_n,\\ &\frac{1}{T_n}\sum_{t=1}^{T_n}\left(-\log\sum_{m\in S_{nt}}\exp(-d(x_{nt},\mu_{cm})-\theta_m)\right)\\ &\geq\ 1\ +\ \frac{1}{T_n}\sum_{t=1}^{T_n}d(x_{nt},\mu_{y_n m_{nt}})+\theta_{m_{nt}}.\end{aligned} \tag{4}$$

The loss function becomes:

$$\mathbf{L} = \sum_{n=1}^{N} \sum_{c \neq y_n} max(0, 1 + \frac{1}{T_n} \sum_{t=1}^{T_n} (d(x_{nt}, \mu_{y_n m_{nt}}) + \theta_{m_{nt}} + \log \sum_{m \in S_{nt}} \exp(-d(x_{nt}, \mu_{cm}) - \theta_m))). \quad (5)$$

This loss function remains convex and can still be solved using dynamic programming. During test, we use again the same principle to achieve fast scoring. Given a test segment of T frames, for each test frame x_t we use the UBM to select the set E_t of k-best scoring proxy labels and compute the LM-dGMM likelihoods using only these k labels. The decision rule is thus given as:

$$y = \underset{c}{argmin} \left\{ \sum_{t=1}^{T} -\log \sum_{m \in E_t} \exp(-d(x_t, \mu_{cm}) - \theta_m) \right\}. \quad (6)$$

B. Handling of outliers

We adopt the strategy of [4] to detect outliers and reduce their negative effect on learning. Outliers are detected using the initial GMM models. We compute the accumulated hinge loss incurred by violations of the large margin constraints in (4):

$$h_n = \sum_{c \neq y_n} max(0, 1 + \frac{1}{T_n} \sum_{t=1}^{T_n} (d(x_{nt}, \mu_{y_n m_{nt}}) + \theta_{m_{nt}} + \log \sum_{m \in S_{nt}} \exp(-d(x_{nt}, \mu_{cm}) - \theta_m))). \quad (7)$$

h_n measures the decrease in the loss function when an initially misclassified segment is corrected during the course of learning. We associate outliers with large values of h_n. We then re-weight the hinge loss terms in (5) by using segment weights $sw_n = min(1, 1/h_n)$:

$$\mathbf{L} = \sum_{n=1}^{N} sw_n h_n. \quad (8)$$

We solve this unconstrained non-linear optimization problem using the second order optimizer LBFGS [10].

In summary, our new and fast training algorithm of LM-dGMM is the following:

- For each class (speaker), initialize with the GMM trained by MAP of the UBM,
- select Proxy labels using these GMM,
- select the set of k-best UBM Gaussian components for each training frame,
- compute the segment weights,
- using the LBFGS algorithm, solve the unconstrained non-linear optimization problem according to (8)

$$min \ \mathbf{L}. \quad (9)$$

IV. Experimental results

We carry out experiments using NIST-SRE'2004 and 2006 data and compare the performances of the baseline GMM and our new LM-dGMM system on a speaker identification task. The feature extraction is carried out by the filter-bank based cepstral analysis tool Spro [11]. Bandwidth is limited to the 300-3400Hz range. 24 filter bank coefficients are first computed over 20ms Hamming windowed frames at a 10ms frame rate and transformed into Linear Frequency Cepstral Coefficients (LFCC). Consequently, the feature vector is composed of 50 coefficients including 19 LFCC, their first derivatives, their 11 first second derivatives and the delta-energy. The LFCCs are preprocessed by Cepstral Mean Subtraction and variance normalization. We applied an energy-based voice activity detection to remove silence frames, hence keeping only the most informative frames. Finally, the remaining parameter vectors are normalized to fit a zero mean and unit variance distribution.

We use the state-of-the-art open source software ALIZE/Spkdet [12], [13] for GMM modeling. A male-dependent UBM is trained using all the telephone data from the NIST-SRE'2004. Then we train a MAP adapted GMM for each speaker of 50 male target speakers belonging to the NIST-SRE'2006 primary task (1conv4w-1conv4w). The corresponding list of 11600 trials are used for test. Session variability modeling and score normalization techniques are not used in our experiments. The so MAP adapted GMM define the baseline GMM system, and are used as initialization for the LM-dGMM one.

TABLE I. Speaker identification rates with GMM and Large Margin diagonal GMM models

System	GMM	LM-dGMM
16 Gaussians	61.6%	67.7%
32 Gaussians	68.1%	69.8%
64 Gaussians	70.7%	72.8%
128 Gaussians	74.6%	77.2%
256 Gaussians	73.3%	**77.6%**

Table I shows the speaker identification accuracy scores of the two systems using different number of GMM mixtures (*M*

= 16, 32, 64, 128, 256). All these scores are obtained with the 10 best proxy labels selected using the UBM, *k=10*.

The results of Table I show that the LM-dGMM algorithm yields significantly better accuracy than the GMM system in all configurations. In particular, our best system achieves *77.6%* speaker identification rate, while the best GMM achieves *74.6%*. This is a *3%* improvement which shows that our discriminative training significantly outperforms the classical generative one.

Moreover, the LM-dGMM system with 256 Gaussians required less than 2 hours for train and test (estimated on an Intel XEON 64bits 3.16GHz Processor, with 6MB of L2 cache and 24GB of RAM). This shows that by relaxing the margin constraints, we could not only develop a relatively accurate algorithm for speaker identification, but also a highly efficient one.

V. Conclusion

We presented a new simplified algorithm to train Large-Margin GMM by using the *k*-best scoring Gaussians selected form the UBM. This algorithm is highly efficient which makes it well suited to process large scale databases such as in NIST'SRE. We carried out experiments on a speaker identification task using NIST'SRE data. The results show that we achieve significantly better accuracy than the baseline GMM system with high computational efficiency. These encouraging results suggest that this framework should be further investigated and applied to large scale applications. Our future work will be to apply this new algorithm to speaker verification under NIST'SRE conditions.

References

[1] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[2] J. Keshet and S. Bengio, *Automatic speech and speaker recognition: Large margin and kernel methods*, Wiley, 2009.

[3] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, 2006.

[4] F. Sha and L. K. Saul, "Large margin Gaussian mixture modeling for phonetic classification and recognition," in *Proc. of ICASSP*. IEEE, 2006, vol. 1, pp. 265–268.

[5] F. Sha and L. K. Saul, "Large margin hidden Markov models for automatic speech recognition," *Advances in Neural Information Processing Systems*, vol. 19, pp. 1249–1256, 2007.

[6] R. Jourani, K. Daoudi, R. André-Obrecht, and D. Aboutajdine, "Large Margin Gaussian mixture models for speaker identification," in *Proc. of Interspeech*, 2010, pp. 1441–1444.

[7] http://www.itl.nist.gov/iad/mig//tests/sre/ .

[8] C. M. Bishop, *Pattern recognition and machine learning*, Springer Science+Business Media, LLC, New York, 2006.

[9] F. Sha, *Large margin training of acoustic models for speech recognition*, Ph.D. thesis, University of Pennsylvania, 2007.

[10] J. Nocedal and S. J. Wright, *Numerical optimization*, Springer verlag, 1999.

[11] G. Gravier, *SPro: "Speech Signal Processing Toolkit"*, 2003, Online: https://gforge.inria.fr/projects/spro.

[12] J.-F. Bonastre, N. Scheffer, D. Matrouf, C. Fredouille, A. Larcher, A. Preti, G. Pouchoulin, N. Evans, B. Fauve, and J. Mason, "ALIZE/SpkDet: a state-of-the-art open source software for speaker recognition," in *Proc. of Odyssey-The Speaker and Language Recognition Workshop*, 2008.

[13] B. Fauve, D. Matrouf, N. Scheffer, J.-F. Bonastre, and J. Mason, "State-of-the-Art Performance in Text-Independent Speaker Verification through Open-Source Software," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, Issue 7, pp. 1960–1968, 2007.