EUSIPCO 2013 1569744297

COMBINATION OF SVM AND LARGE MARGIN GMM MODELING 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, Toulouse, France
[2]GeoStat Group, INRIA Bordeaux-Sud Ouest
200 avenue de la vieille tour, 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

Most state-of-the-art speaker recognition systems are partially or completely based on Gaussian mixture models (GMM). GMM have been widely and successfully used in speaker recognition during the last decades. They are traditionally estimated from a world model using the generative criterion of Maximum A Posteriori. In an earlier work, we proposed an efficient algorithm for discriminative learning of GMM with diagonal covariances under a large margin criterion. In this paper, we evaluate the combination of the large margin GMM modeling approach with SVM in the setting of speaker identification. We carry out a full NIST speaker identification task using NIST-SRE'2006 data, in a Symmetrical Factor Analysis compensation scheme. The results show that the two modeling approaches are complementary and that their combination outperforms their single use.

Index Terms— Large margin training, Gaussian mixture models, discriminative learning, Support vector machines, speaker recognition.

1. INTRODUCTION

Most state-of-the-art speaker recognition systems are based on Gaussian mixture models (GMM). These systems model target speakers by GMM or fuse them with other modeling approaches. GMM are traditionally estimated from a world model using the generative criterion of Maximum A Posteriori (MAP). A speaker-independent model or Universal Background Model (UBM) is first trained with the Expectation-Maximization (EM) algorithm using various speech recordings gathered from a large speaker population. When enrolling a new speaker to the system, the parameters of the UBM are MAP adapted to the feature distribution of the new speaker. Traditionally, in this GMM–UBM approach, the target speaker GMM is derived from the UBM model by updating only the mean parameters, while the (diagonal) covariances and the weights remain unchanged [1].

In speaker recognition applications, mismatch between the training and testing conditions can decrease considerably the performances. The session variability remains the most challenging problem to solve. The Factor Analysis techniques [2, 3], e.g., Symmetrical Factor Analysis (SFA) [4, 5], were proposed to address that problem in GMM based systems.

Generative training does not however directly addresses the classification problem because it uses the intermediate step of modeling system variables, and because classes are modeled separately. For this reason, discriminative training approaches have been an interesting and valuable alternative since they focus on adjusting boundaries between classes [6, 7], and lead generally to better performances than generative methods. In the literature, various works have dealt with different variants on GMM front-end, combined with Support Vector Machines (SVM) back-end; for example [8, 9, 10, 11]. SVM combined with GMM supervectors are among state-of-the-art discriminative approaches in speaker recognition [12, 13].

In earlier works [14, 15], we proposed an efficient algorithm for discriminative learning of GMM with diagonal covariances under a large margin criterion. Our modeling is based on a recent discriminative approach for multiway classification that has been used in speech recognition, the Large Margin Gaussian mixture models (LM-GMM) [16, 17]. In our LM-dGMM modeling, we separate the classes by defining a large margin criterion on the distances between the feature vectors and the models mean vectors. Our discriminative models outperform the traditional generative GMM.

In this paper, we study and evaluate the combination of the SVM and the LM-dGMM modeling approaches in the setting of speaker recognition.

We carry out a full NIST speaker recognition task using NIST-SRE'2006 (core condition) data [18], in a Symmetri-

cal Factor Analysis compensation scheme; SFA can be seen here as a preprocessing step in the LM-dGMM modeling. We compare the performances of GMM–SFA[1], LM-dGMM–SFA[2], (GMM–SFA) + SVM and (LM-dGMM–SFA) + SVM systems[3]. The results show that LM-dGMM and SVM are complementary and that their combination improves the classification performance.

The paper is organized as follows. After an overview on GMM supervector linear kernel SVM modeling in section 2, we describe our efficient training algorithm of LM-dGMM models in section 3. The experimental results and their discussions are then presented in section 4.

2. GMM SUPERVECTOR LINEAR KERNEL SVM SYSTEM

In this section we briefly describe the GMM supervector linear kernel SVM system (GSL)[12].

Given an M-components GMM trained by MAP adaptation from a world model, one forms a GMM supervector by stacking the D-dimensional mean vectors, leading to an MD supervector. This GMM supervector can be seen as a mapping of variable-length utterances into a fixed-length high-dimensional vector, through GMM modeling.

For two utterances x and y, a kernel distance based on the Kullback-Leibler divergence between the GMM models $\{\mu_{xm}, \Sigma_m, w_m\}$ and $\{\mu_{ym}, \Sigma_m, w_m\}$ trained on these utterances, is defined as:

$$\mathrm{K(x,y)} = \sum_{m=1}^{M} \left(\sqrt{w_m}\Sigma_m^{-(1/2)}\mu_{xm}\right)^T \left(\sqrt{w_m}\Sigma_m^{-(1/2)}\mu_{ym}\right). \tag{1}$$

The UBM weight and variance parameters, i.e., w_m and Σ_m, are used to normalize the Gaussian means μ_{cm} before feeding them into a linear kernel SVM training [12].

3. LM-dGMM MODELING

3.1. LM-dGMM training with k-best Gaussian

In Large Margin diagonal GMM (LM-dGMM) [14], each class (speaker) c is initially modeled by a GMM with M diagonal mixtures, trained by MAP adaptation of a world model. For each class c, the m^{th} Gaussian is parameterized by a mean vector μ_{cm}, a diagonal covariance matrix $\Sigma_m = diag(\sigma^2_{m1}, ..., \sigma^2_{mD})$ and a scalar factor $\theta_m = \frac{1}{2}\left(D\log(2\pi) + \log|\Sigma_m|\right) - \log(w_m)$, where w_m is the weight of the Gaussian and D is the dimension of the observations.

[1]GMM models trained in an SFA compensation scheme.

[2]SFA compensated LM-dGMM models.

[3]GMM–SFA and LM-dGMM–SFA supervectors linear kernel SVM systems.

For each training example o_n 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_n of the Gaussian component of the GMM modeling the class y_n which has the highest posterior probability. This index is called *proxy label*. We select too the set S_n of the k-best UBM Gaussian components, i.e., the indices of the k UBM Gaussian components with the highest posterior probabilities.

For each observation o_n, the goal of the training algorithm is to force the log-likelihood of its proxy label Gaussian m_n to be at least one unit greater than the log-likelihoods of the k-best Gaussian components of all competing classes. This one unit minimal margin to satisfy is an arbitrary choice. We note that by using some other experimental values, the experiments show that the performances increase. One of our actual works consists in improving margin selection.

Given the training examples $\{(o_n, y_n, m_n, S_n)\}_{n=1}^N$, we seek mean vectors μ_{cm} that satisfy the large margin constraints in Eq. (2) [15]:

$$\begin{aligned} &\forall c \neq y_n, \forall m \in S_n, \\ &\Big(d(o_n, \mu_{cm}) + \theta_m\Big) \;\geq\; 1 \;+\; \Big(d(o_n, \mu_{y_n m_n}) + \theta_{m_n}\Big), \end{aligned} \tag{2}$$

where $d(o_n, \mu_{cm}) = \sum_{i=1}^{D} \frac{(o_{ni} - \mu_{cmi})^2}{2\sigma^2_{mi}}$. Eq. (2) states that for each competing class $c \neq y_n$ the match (in term of normalized Euclidean distance) of the k nearest centroids in class c is worse than the target centroid by a margin of at least one unit.

Afterward, these k constraints are fold into a single one using the softmax inequality $\min_m a_m \geq -\log \sum_m \exp(-a_m)$. The large margin constraints become thus:

$$\begin{aligned} &\forall c \neq y_n, \quad -\log \sum_{m \in S_n} \exp(-d(o_n, \mu_{cm}) - \theta_m) \\ &\geq\; 1 + d(o_n, \mu_{y_n m_n}) + \theta_{m_n}. \end{aligned} \tag{3}$$

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

$$\begin{aligned} Ł \;=\; \sum_{n=1}^{N} \sum_{c \neq y_n} max\Bigg(0 \,,\; 1 + d(o_n, \mu_{y_n m_n}) + \theta_{m_n} \\ + \log \sum_{m \in S_n} \exp(-d(o_n, \mu_{cm}) - \theta_m)\Bigg). \end{aligned} \tag{4}$$

During test, we compute a match score depending on both the target model $\{\mu_{cm}, \Sigma_m, \theta_m\}$ and the UBM $\{\mu_{Um}, \Sigma_m, \theta_m\}$ for each test hypothesis. For each test frame o we use the UBM to select the set E of k-best scoring proxy labels and

compute the average log likelihood ratio using these k labels:

$$\begin{aligned} LLR_{avg} &= log \sum_{m \in E} \exp(-d(o, \mu_{cm}) - \theta_m) \\ &\quad - log \sum_{m \in E} \exp(-d(o, \mu_{Um}) - \theta_m). \end{aligned} \tag{5}$$

This quantity provides a score for the test segment to be uttered by the target model/speaker c. The higher the score is, the greater the probability that the test segment was uttered by the target speaker is.

3.2. Segmental training

In speaker recognition, the decision is known to be taken on a sequence of feature vectors belonging to a speech segment. The processing is done on a segmental manner. We rewrite thus the previous frame-based formulas in the segmental training scheme, to apply collectively to multiple consecutive analysis frames. Let t index the T_n frames belonging to the n^{th} segment (i.e. n^{th} speaker training data) $\{o_{n,t}\}_{t=1}^{T_n}$.

The segment-based large margin constraints, loss function and decision rule are thus:

$$\begin{aligned} &\forall c \neq y_n, \\ &\frac{1}{T_n} \sum_{t=1}^{T_n} \left(- \log \sum_{m \in S_{n,t}} \exp(-d(o_{n,t}, \mu_{cm}) - \theta_m) \right) \\ &\geq 1 + \frac{1}{T_n} \sum_{t=1}^{T_n} d(o_{n,t}, \mu_{y_n m_{n,t}}) + \theta_{m_{n,t}}, \end{aligned} \tag{6}$$

$$\begin{aligned} \text{Ł} &= \sum_{n=1}^{N} \sum_{c \neq y_n} max\left(0\ ,\ 1 + \frac{1}{T_n} \sum_{t=1}^{T_n} \left(d(o_{n,t}, \mu_{y_n m_{n,t}}) \right.\right. \\ &\quad \left.\left. + \theta_{m_{n,t}} + \log \sum_{m \in S_{n,t}} \exp(-d(o_{n,t}, \mu_{cm}) - \theta_m) \right)\right), \end{aligned} \tag{7}$$

$$\begin{aligned} LLR_{avg} &= \frac{1}{T} \sum_{t=1}^{T} \left(log \sum_{m \in E_t} \exp(-d(o_t, \mu_{cm}) - \theta_m) \right. \\ &\quad \left. - log \sum_{m \in E_t} \exp(-d(o_t, \mu_{Um}) - \theta_m) \right). \end{aligned} \tag{8}$$

3.3. Handling of outliers

Outliers are feature vectors that lie considerably on the wrong side of a decision boundary, i.e., feature vectors that are closer to the competing class centroids than the target centroid. They are common in speech corpora.

We adopt the strategy of [16] to detect the outliers that occur in the training data. Using the initial GMM models trained by MAP of the UBM, we compute the accumulated hinge loss incurred by violations of the large margin constraints in (6):

$$\begin{aligned} h_n &= \sum_{c \neq y_n} max\left(0\ ,\ 1 + \frac{1}{T_n} \sum_{t=1}^{T_n} \left(d(o_{n,t}, \mu_{y_n m_{n,t}}) + \right.\right. \\ &\quad \left.\left. \theta_{m_{n,t}} + log \sum_{m \in S_{n,t}} exp(-d(o_{n,t}, \mu_{cm}) - \theta_m) \right)\right). \end{aligned} \tag{9}$$

h_n measures the decrease in the loss function Ł when an initially misclassified segment is corrected during the course of learning. We associate outliers with values of $h_n > 1$, and in this case we multiply the hinge loss term by the weight $s_n = \frac{1}{h_n}$ (the correctly classified examples have a unit weight $s_n = 1$). We compute the weighting factors using the GMM–UBM models and then we hold them fixed during the discriminative training. The new loss function becomes thus:

$$\begin{aligned} \text{Ł} &= \sum_{n=1}^{N} s_n \sum_{c \neq y_n} max\left(0\ ,\ 1 + \frac{1}{T_n} \sum_{t=1}^{T_n} \left(d(o_{n,t}, \mu_{y_n m_{n,t}}) \right.\right. \\ &\quad \left.\left. + \theta_{m_{n,t}} + \log \sum_{m \in S_{n,t}} \exp(-d(o_{n,t}, \mu_{cm}) - \theta_m) \right)\right). \end{aligned} \tag{10}$$

We minimize this loss function using the second order optimizer LBFGS [19].

In summary, the 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 $\{m_{n,t}\}$ using these GMM,
- select the set $S_{n,t}$ of k-best UBM Gaussian components for each training frame,
- compute the weights s_n,
- minimize the objective function according to equation Eq. (10)

$$min \quad \text{Ł}. \tag{11}$$

4. EXPERIMENTAL RESULTS

We perform experiments using data of the NIST-SRE'2006 [18] speaker recognition task and compare the performances of GMM supervectors based SVM systems. We evaluate linear kernel SVM systems trained on session-variability compensated GMM supervectors:

- GMM–SFA supervectors, i.e., supervectors of generative GMM models trained in a Symmetrical Factor Analysis (SFA) compensation scheme [4, 5],

- LM-dGMM–SFA supervectors, i.e., supervectors of SFA compensated LM-dGMM models. The LM-dGMM–SFA models are initialized by model domain compensated GMM, which are then discriminated using feature domain compensated data.

The comparisons are made on the male part of the NIST-SRE'2006 core condition 1conv4w-1conv4w. 349 target speakers are included in this large-scale application, with 22123 trials to process involving 1601 test segments, which represents significant evaluation conditions. Performances are measured in terms of equal error rate (EER) and minimum of detection cost function (minDCF). The latter is calculated following NIST criteria [20].

For front-end processing, we follow the same procedure as in [5]. The feature extraction is carried out by the filter-bank based cepstral analysis tool Spro [21]. 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) [22]. 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 [23].

After that feature 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 [24, 5] for GMM–SFA modeling. To do so, A male-dependent UBM is trained using all the telephone data from the NIST-SRE'2004, and a session variability matrix $\mathbf{U}$ of rank $R = 40$ is estimated on NIST-SRE'2004 data using 2934 utterances of 124 different male speakers. The SVM training uses as a blacklist a list of 200 impostor speakers from the NIST-SRE'2004.

Table 1 provides the EERs and minDCFs of the GMM–SFA and LM-dGMM–SFA supervectors linear kernel SVM systems, for models with 512 Gaussian components ($M = 512$). All the large margin results are obtained with the 10 best proxy labels selected using the UBM, $k = 10$. We also report in the table the performances of the standalones systems GMM–SFA and LM-dGMM–SFA.

The results of Table 1 show that the SVM post-classification of the (generative and discriminative) compensated GMM supervectors leads to better performances. Indeed, the combination with SVM reduces the EER of the GMM–SFA and LM-dGMM–SFA systems by respectively $19,17\%$ and $12,55\%$, which is statistically significant. The systems combination substantially improves the recognition results.

As expected, the results of Table 1 confirm that our discriminative learning approach improves the performances of the GMM models in the two cases, with and without the combination with SVM. Moreover, they suggest that the two modeling approaches LM-dGMM and SVM are complementary. This can be explained first by the fact that the GMM obtained with our large margin approach can directly be used in a SVM classifier. Moreover, the re-estimation of the GMM mean vectors under the large margin criterions leads to a more distant (a more separated) supervectors in the (high dimensional) feature space which is beneficial to the SVM classifier. We also emphasize that the combination of LM-dGMM and SVM accelerates our speaker models evaluation during the scoring phase.

System	EER	minDCF(x100)
GMM–SFA	5.53%	2.18
LM-dGMM–SFA	5.02%	2.18
(GMM–SFA) + SVM	4,47%	2.17
(LM-dGMM–SFA) + SVM	**4.39%**	**2.16**

Table 1. EER(%) and minDCF(x100) performances for GMM–SFA, LM-dGMM–SFA, (GMM–SFA) + SVM and (LM-dGMM–SFA) + SVM systems, using models with 512 components.

5. CONCLUSION

We have proposed an efficient algorithm for discriminative learning of GMM under a large margin criterion. Our algorithm is suitable for the SFA channel compensation paradigm and achieves better performances than the standard generative GMM models. We carried out experiments on the male speaker recognition task under the NIST-SRE'2006 core condition. Combined with SVM classifiers, the resulting system outperforms the state-of-the-art speaker recognition discriminative approach of GMM supervector linear kernel SVM. our (LM-dGMM–SFA) + SVM system achieves 4.39% equal error rate and $2.16 * 10^{-2}$ minDCF value. Our future work will consist in applying the large margin concept in the Total Variability space instead of the actual feature space. In the Total Variability space [25, 26], the speakers are assumed to be represented by identity vectors (i-vectors) containing discriminative speaker specific informations. Thus, the definition of large margin constraints on the i-vectors is very promising, and we expect it to improve performances.

6. 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] P.J. Kenny, G. Boulianne, and P. Dumouchel, "Eigen-

voice modeling with sparse training data," *IEEE transactions on speech and audio processing*, vol. 13, no. 3, pp. 345–354, 2005.

[3] P.J. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker and session variability in GMM-based speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1448–1460, 2007.

[4] D. Matrouf, N. Scheffer, B.G.B. Fauve, and J.-F. Bonastre, "A straightforward and efficient implementation of the factor analysis model for speaker verification," in *Proc. of Interspeech*, 2007, pp. 1242–1245.

[5] B.G.B. Fauve, D. Matrouf, N. Scheffer, J.-F. Bonastre, and J.S.D. 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.

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

[7] J. Louradour, K. Daoudi, and F. Bach, "Feature space mahalanobis sequence kernels: Application to svm speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 8, pp. 2465–2475, 2007.

[8] M. Liu, B. Dai, Y. Xie, and Z. Yao, "Improved GMM-UBM/SVM for speaker verification," in *Proc. of ICASSP*, 2006, vol. 1, pp. I–925–I–928.

[9] N. Krause and R. Gazit, "SVM-based speaker classification in the GMM models space," in *Proc. of Odyssey*, 2006.

[10] N. Dehak and G. Chollet, "Support vector GMMs for speaker verification," in *Proc. of Odyssey*, 2006.

[11] C.H. You, K.A. Lee, and H. Li, "GMM-SVM Kernel With a Bhattacharyya-Based Distance for Speaker Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1300–1312, 2010.

[12] 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.

[13] W.M. Campbell, D.E. Sturim, D.A. Reynolds, and A. Solomonoff, "Svm based speaker verification using a gmm supervector kernel and nap variability compensation," in *Proc. of ICASSP*, 2006, vol. 1, pp. I–97–I–100.

[14] 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.

[15] R. Jourani, K. Daoudi, R. André-Obrecht, and D. Aboutajdine, "Discriminative speaker recognition using Large Margin GMM," *Neural Computing & Applications*, vol. 22, no. 7-8, pp. 1329–1336, 2013.

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

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

[18] *The NIST Year 2006 Speaker Recognition Evaluation Plan*, 2006, Online: *www*.itl.nist.gov/iad/mig/tests/spk/2006/sre-06_evalplan-v9.pdf.

[19] J. Nocedal and S.J. Wright, *Numerical optimization*, Springer verlag, New York, 1999.

[20] M. Przybocki and A. Martin, "NIST Speaker Recognition Evaluation Chronicles," in *Proc. of Odyssey*, 2004, pp. 15–22.

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

[22] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[23] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, no. 1-3, pp. 133–147, 1998.

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

[25] N. Dehak, R. Dehak, P.J. Kenny, N. Brümmer, P. Ouellet, and P. Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," in *Proc. of Interspeech*, 2009, pp. 1559–1562.

[26] N. Dehak, P.J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.