A NEW HYBRID INFINITE GMM-SVM SYSTEM FOR SPEAKER VERIFICATION

Souad Friha[a], *Nora Mansouri*[b] *and Abdlmalik Taleb Ahmed*[c]

[a] Faculté des Sciences et Technologie, Université Abbas Laghrour, Khenchela, Algeria.

[b] Laboratoire d'Automatique et de Robotique, Université de Mentouri, Constantine, Algeria.

[c] LAMIH, Le Mont Houy, Université de Valenciennes et du Hainaut Cambrésis, Valenciennes, France.

ABSTRACT

A new method for speaker text-independent verification that combines the Infinite Gaussian Mixture Models (IGMM) with Support Vector Machine (SVM) is described. Infinite GMM supervectors are constructed by stacking the means of the adapted mixture then they are trained via an SVM classifier. This allows overcoming the problem of fixing a priori the number of the involved Gaussians. Experiments showed a relative gain of about 12% in terms of the Equal Error Rate (EER) and about 59% in terms of the minimum detection cost Function (min DCF). Moreover, more improvement in terms of both EER and min DCF can be noticed when time increases with a lower number of components for comparable performance with GMM models.

Key words— *Speaker Verification, Infinite GMM, Dirichlet Process, Gibbs Sampling, Supervector.*

1. INTRODUCTION

We consider the problem of text independent speaker verification. The standard approach to this problem is to model the speaker using a Gaussian Mixture Model (GMM). [1-3]. Another state of the art technique, widely adopted within the speaker recognition domain, is the Support Vector Machine (SVM) discrimination [4-6]

Dealing with GMM models in combination with SVM is an attractive way to model systems, but one has to fix a priori the number of Gaussians involved. This has been an open problem for many years and some research works were carried out in order to estimate the optimal number [7].This has been resolved elegantly by Rasmussen in his original paper [8] within a general framework of the so-called Dirichlet Process Mixture (DPM) model which extends the finite mixture model to an infinite one. Inference in this model is done using Gibbs sampling. Literature on Dirichlet Process and Gibbs sampling is abundant (see [8,14] for example).

In this paper we propose an IGMM-SVM approach rather than the traditional GMM-SVM approach. Combining SVM this way with the IGMM models instead of the finite traditional GMM will take benefit from both the robustness of SVM and the appropriate IGMM estimation of the model order. This is supposed to speed up the convergence of the algorithm.

The outline of the paper is as follows. Section 2 describes the basic theoretical framework for IGMM introduced via GMM principle. In section 3 supervectors for respectively GMM-SVM and IGMM-SVM systems are presented. Section 4 contains a description of the conducted experiments with results and comments. Conclusion follows in section 5.

2. FINITE VERSUS INFINITE GAUSSIAN MIXTURE MODELING

The infinite Gaussian Mixture Model (GMM) is an example of Dirichlet Process mixture Models (DPM). These are mixture based models built using the Dirichlet Process. The Dirichlet Process (DP) mixture model itself is a nonparametric Bayesian model for clustering problems involving multiple groups of data. Each group of data is modeled with a mixture, with the number of components being open-ended and inferred automatically by the model. Further, components can be shared across groups, allowing dependencies across groups to be modeled effectively as well as conferring generalization to new groups [14]. In the case of speaker verification groups stand for speakers. We chose DPM because within this model the number of clusters is open-ended which is particularly interesting to our problem where there is no prior knowledge about the system complexity. Furthermore, the DPMs offer an alternative to the drawbacks of the Gaussian Mixture Models (GMM) which tend to smear multimodal behavior through averaging [15]. Within a DPM model framework, individual characteristics can be balanced with global behavior without weakening the quality of the individual models. What the DPM model attempts to do is to preserve unique behaviors through use of an infinite mixture model. Another interesting point is that under the Dirichlet process, the number of clusters grows logarithmically with the number of data points [16].

A comprehensive discussion of alternative perspectives on the Dirichlet process mixtures can be found in [9,14]. Within our paper, the concept is introduced through the finite Gaussian mixture model, whose mixing weight is given by a Dirichlet Process prior. The infinite Gaussian mixture model is then derived by considering the situation where the number of mixtures tends to infinity [8,10,14].

2.1. Finite Gaussian Mixture Model

The probability density function of data, $x=\{x_1, \dots, x_n\}$ can be modeled by finite mixtures of Gaussian distributions with k components:

$$p(x|\mu,s,\pi)=\sum_{j=1}^{k}\pi_j\, G\left(\mu_j,s_j^{-1}\right) \qquad (1).$$

where $\mu=\{\mu_1, \dots, \mu_k\}$ are the means, $\mathbf{s}=\{s_1, \dots, s_k\}$ are the precisions (inverse variances), $\pi=\{\pi_1, \dots, \pi_k\}$ are the mixing weights (which must be positive and sum to one) and G is a Gaussian distribution.

Given a set of training data with N observations, $x=\{x_1,\dots x_N\}$ the the goal is to estimate the GMM parameters $(\boldsymbol{\mu}, \boldsymbol{s}, \boldsymbol{\pi})$. Within a Bayesian framework the inference is performed with respect to the posterior probability of the parameters. As a reminder for a model M Bayes' rule is:

$$P(M|Y)\propto P(Y|M)P(M) \qquad (2).$$

Where P(M|Y) is the posterior probability of the model M given a set of observations Y, P(Y|M) is the likelihood of the observations under the model and P(M) is the prior probability of the model M.

In general, priors for the model parameters are specified via *hyper-parameters*, which themselves are given higher level priors. The inference of samples from the posterior distribution is implemented using Gibbs sampling [10].

Component means are given Gaussian priors:

$$p\left(\mu_j\middle|\lambda,r\right) \sim G(\lambda,r^{-1}) \qquad (3).$$

where prior mean λ and prior precision r, are hyper-parameters that are common to all components. The hyper-parameters themselves are given vague Gaussian and Gamma hyper-priors:

$$p(\lambda)=G\left(\mu_x,\sigma_x^2\right) \qquad (4).$$

$$p(r)=G_a\left(1, \sigma_x^{-2}\right)\propto r^{-\frac{1}{2}}\exp\left(-\frac{r\sigma_x^2}{2}\right) \qquad (5).$$

Where μ_x and σ_x^2 are the mean and the variance of the training data points.

To make inferences with respect to component means, the conditional posterior distributions from μ_j are obtained by multiplying the likelihood in (9) by the prior in (11), resulting in a Gaussian distribution:

$$p\left(\mu_j|c,x,s_j,\lambda,r\right) \sim G\left(\frac{\bar{x}_j N_j s_j+\lambda r}{N_j s_j+r}, \frac{1}{N_j s_j+r}\right) \qquad (6).$$

where $\bar{x}_j$ and N_j are the mean and the number of data points belonging to component j, respectively. The latent indicator variable $c=\{c_1\dots c_N\}$ is introduced to indicate that the data point x_n belongs to mixture component c_n.

Component precisions are given Gamma priors (15) and as for the general case of mixture models the mixing weights are given Dirichlet priors with concentration parameter $\frac{\alpha}{k}$ in (16) (See [10] for more details).

$$p(s_j|c,x,\mu_j,\beta,\omega)\sim Ga\left(\beta+N_j, \left[\frac{\omega\beta+\sum_{i=c_i:j}\left(x_i-\mu_j\right)^2}{\beta+N_j}\right]^{-1}\right) \qquad (7).$$

$$p(\pi_1, \dots,\pi_k| c)\sim \text{Dirichlet}\left({}^{\alpha}/_{k}, \dots, {}^{\alpha}/_{k}\right) \qquad (8).$$

To use Gibbs sampling, a very wide used MCMC method, for the discrete indicators c_i, the conditional prior for a single indicator, given all the other indicators, is required and can be obtained as follows:

$$p\left(c_n=j\middle|c_{-n},\alpha\right)=\frac{N_{-i,j}+\alpha/k}{N-1+\alpha} \qquad (9).$$

where the subscript –i indicates all indices except i and $N_{-n,\, j}$ is the number of data points, excluding x_n which belongs to mixture j. The conditional posterior of each c_n are given by the multiplication of the likelihood and the prior:

$$p\left(c_n=j\middle|c_{-n},\mu_j,s_j,\alpha\right)\propto\frac{N_{-n,j}+\frac{\alpha}{k}}{N-1+\alpha}s_j^{\frac{1}{2}}\exp\left(-\frac{s_j\left(x_i-\mu_j\right)^2}{2}\right) \qquad (10).$$

2.2. Extension to the Infinite Gaussian Mixtures

The previous subsections have been restricted to a finite number of mixtures. In Bayesian methodology, inference is performed on an infinite number of mixtures. The computation with infinite mixtures is finite through the use of "represented" and "unrepresented" mixtures. Represented mixtures are those that have training data associated with them whilst unrepresented mixtures, which are of infinite number, have no training data associated with them.

Let $k\rightarrow\infty$ in (17):

$$p\left(c_n=j\middle|c_{-n},\alpha\right)=\begin{cases}\frac{N_{-n,j}}{N-1+\alpha} & \text{j is represented}\\ \frac{\alpha}{N-1+\alpha} & \text{j is unrepresented}\end{cases} \qquad (11).$$

The conditional posteriors of the indicator variables are as follows:

$$p\left(c_n=j\middle|c_{-n},\alpha\right)\propto\begin{cases}\frac{N_{-n,j}}{N-1+\alpha}s_j^{\frac{1}{2}}\exp\left(-\frac{s_j\left(x_n-\mu_j\right)^2}{2}\right) & \text{for represented j}\\ \frac{\alpha}{N-1+\alpha}\int p(x_n|\mu_j,s_j)p\left(\mu_j\middle|\lambda,r\right)p(s_j|\beta,\omega)d\mu_j ds_j & \text{for unrepresented j}\end{cases} \quad (12).$$

The time complexity for each iteration of Gibbs sampling is $O(N\, k_{rep})$.

3. SUPERVECTOR CONSTRUCTION

3.1. GMM-SVM Supervector

A speaker-independent GMM UBM is trained with acoustic data from a set of different speakers to represent general speech characteristics. During the enrollment phase, the GMM-SVM supervectors are constructed as follows:

-Acoustic feature vectors X are extracted from all available training utterances of the enrolling speaker.

-MAP adaptation is used to obtain a GMM model M with K=512 components from UBM model, only means are adapted.

-The N-dimensional means obtained are normalized by the corresponding deviation of each of the Gaussian mixtures in the adapted GMM model.

-A supervector V_x is constructed for speaker S by concatenating the N-dimensional means resulting in a KxN dimensional vector. An SVM classifier is trained using the target GMM supervector.

-An SVM classifier is trained using the target GMM supervectors V_x as positive examples (labeled as +1) and the SVM background, a set of imposter speaker vectors common to all enrollment speakers, as negative examples (labeled as -1).

-The UBM model and the SVM parameters are finally stored.

While verification for a given input speech utterance the first four steps are the same as before. Then, an SVM classifier allows to decide whether the enrolled speaker and the input speech came from the same speaker or not.

3.2. Infinite GMM-SVM Supervector

When dealing with the IGMM models the number of hidden classes is treated as a parameter to be learned. The construction strategy is nearly the same as before:

-Acoustic feature vectors X are extracted from all available training utterances of the enrolling speaker.

-Gibbs sampler is run upon every input feature utterance till convergence. This leads to a mixture model M with K_{rep} Gaussian mixtures.

-The N-dimensional means obtained are normalized by the corresponding deviation of each of the Gaussian mixtures.

-A supervector V_x is constructed for speaker S by concatenating the N-dimensional means resulting in a K_{rep}xN dimensional vector.

-An SVM classifier is trained using the target GMM supervectors V_x, as positive examples (labeled as +1) and the SVM background, a set of imposter speaker vectors common to all enrollment speakers, as negative examples (labeled as -1).

-The IGMM model and the SVM parameters are stored.

Given an input speech utterance during the verification stage, the first four steps are the same as before. The final step is the SVM classification.

Fig. 1 illustrates the global scheme of speaker verification within an infinite GMM scheme.

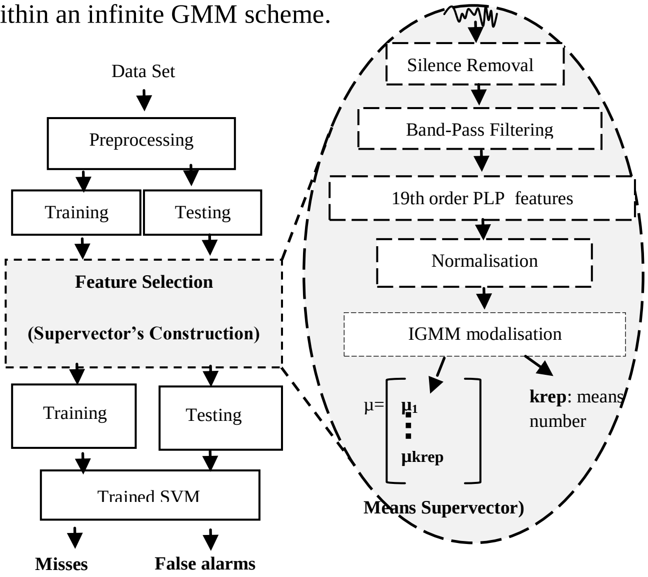

Fig. 1. A Modular representation of the infinite GMM system

4. EXPERIMENTS

Experiments are performed on the 2006 NIST (SRE) corpus. We use a speech feature representation based on a RASTA-PLP approach. This is an acronym for *Relative Spectral Transform-Perceptual Linear Prediction*. Both of these two techniques aim at computing the speech parameters similar to the way how a human perceives sounds.

A 19-dimensional PLP vector is computed from pre-emphasized speech every 20 ms using Hamming window. The first coefficient is then discarded and only 18 coefficients are kept. Delta cepstral coefficients are obtained and appended to the cepstra resulting in a 36 dimensional feature vector. An energy-based speech detector is applied to discard silence and noise frame. To mitigate the channel effects, RASTA and mean-variance normalization are applied to the features.

For the implementation of the SVM stage, we used an RBF kernel because it can handle the nonlinearities between class labels and attributes. For the two parameters (C,γ.), namely the cost (or the penality) parameter and the width of the Gaussian function, to be tuned we used the two following values: C=10 and γ=0.5 [5] as yielding to good results for speech applications. We have used the OSU-

SVM toolbox which is derived from the LIBSVM package which can be used within Matlab. All the experiments carried out are gender-independent.

In order to investigate the usefulness of the IGMM over the GMM system, comparison is made between the two techniques by using respectively the same training and test data.

The evaluation is carried out with a total of 38416 trials out of which 20105 trials belong to female speakers and 18311 trials belong to male speakers. The female subset is consisting of 1407 genuine trials and 18698 imposter trials. And the male subset is consisting of 1281 genuine trials and 17030 imposter trials. The two experiments hereafter are repeated many times on different subsets of nearly the same size. All curves look nearly like those of Fig. 2 and Fig. 3.
The GMM supervector is computed using MAP adaptation with 512 components as we have found that almost the best results are obtained for this value, only means are adapted [1,11]. The SVM training is performed by implementing kernel in (7) and using SVMTorch [12].

The SVM background is obtained by extracting 2039 GMM supervectors, respectively 1025 GMM-supervectors for male and 1014 for female from NIST SRE-2006 training database. Results can be evaluated through comparison with some recent studies like in [11]. Similar results obtained for GMM systems participating in the NIST SRE-2006 evaluation are also available online.

Coming to infinite GMM, we implement the Gibbs sampler as described above upon the same training data used for GMM. Unlike the previous GMM approach the number of mixtures is not fixed to 512 a priori and is allowed to be optimized progressively through the successive iterations.
The evaluation for both training and test is performed over a “core set” as defined by the NIST Evaluation Plan [13]. This is a two-channels (4-wire) excerpt from a conversation of approximately five minutes total duration from which 10 seconds of speech are extracted to be used in experiment 2.

4.1. Performance measure

Comparisons of performance are achieved through the calculation of the Equal Error Rate (EER) and minimum Detection Cost Function (DCF). These measures are derived from Detection Error Trade-off (DET) curves.

The (DCF) Detection Cost Function denoted C_{Det} is:

$$C_{Det}=C_{Miss}*P_{Miss|Target}*P_{Target} + C_{False\ Alarm}*P_{False\ Alarm|NonTarget}*(1-P_{Target}) \qquad (13).$$

With the following values recommended by NIST:
$C_{Miss}=10$, $C_{False\ Alarm}=1$ $P_{Target}=0.01$ respectively for the cost of a miss:, the cost of a false alarm and the a priori probability of the specified target speaker probability. Then, $P_{NonTarget}=(1-P_{Target})=0.99$.
For each test, a detection cost function is computed over the sequence of trials provided. Each trial is independently judged as “true” or “false” (according to the fact that the model speaker speaks in the test segment or not) [13].

4.2. Experiment 1: Comparing GMM/IGMM

For both GMM-SVM and IGMM-SVM systems, this experiment is done on the same NIST-SRE 06, core condition, male set (including 680 target tests and 821 nontarget tests). Table 1 shows the performance of the Infinite GMM model versus the finite GMM model. The relative gain is about 12% in terms of the Equal Error Rate (EER) and about 59% in terms of the minimum detection cost Function (min DCF). The corresponding results are depicted in Fig. 2.

Table 1.Performance of GMM-SVM system compared with the IGMM-SVM system

System	EER (%)°	Min DCF (X100)
GMM-SVM	2.35	2.30
IGMM-SVM	2.05	0.94

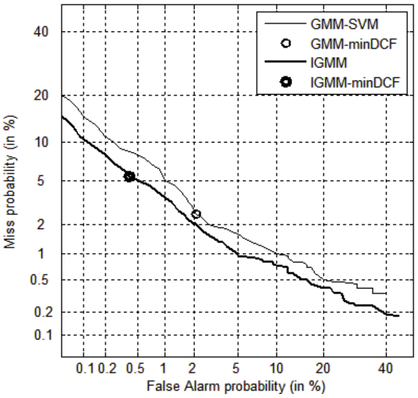

Fig. 2. DET curves for GMM-SVM system and IGMM-SVM system (experiment 1 conditions)

4.3. Experiment 2: Impact of the training duration

We consider the same data sets as in experiment 1 before. Only the training duration is different (respectively 3s, 6s and 10s). Table 2 shows the results in terms of EER and minDCF and Fig 3 illustrates the corresponding DET curves. An improvement in terms of both EER and min DCF can be noticed when time increases. This result goes with similar studies on the impact of the training data which is again expected.

Table 2.Performance of IGMM-SVM according to the training duration

System	EER (%)°	Min DCF (X100)
IGMM-SVM-3s	4.82	4.80
IGMM-SVM-6s	3.55	3.40
IGMM-SVM-10s	2.05	0.94

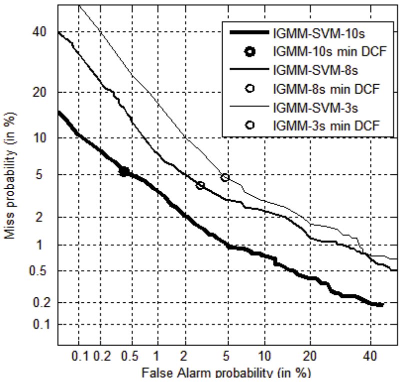

Fig. 3. DET curves for IGMM-SVM system versus training duration (experiment 2 conditions)

4.4. Number of components

Fig. 4 shows the normalized histogram of the number of components during IGMM modeling. This number is, in almost cases, well below the advocated value 512.

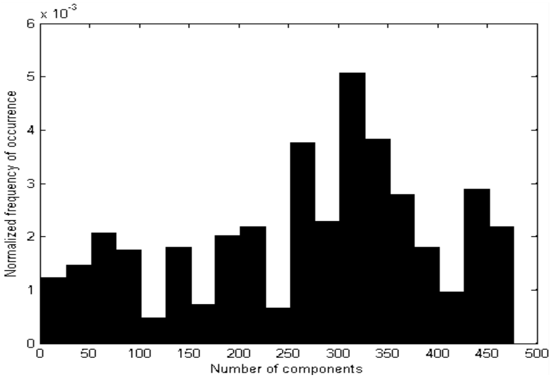

Fig. 4. Histogram of the components number during IGMM modeling

5. CONCLUSION

The effectiveness of replacing GMM by IGMM with a combination to an SVM classification for speaker recognition is explored. The inherent advantage of this approach is to make it possible to infer the system's complexity from the data without having to make prior assumptions about it.

As mentioned above, the number of the underlying components for IGMM modeling is below 512, which means that the advocated value for the GMM modeling is an overestimation of the system complexity. However, this last conclusion should be handled with care and not be generalized to situations too different from the present context. The model complexity may change e.g. drastically if the speech language is different. All we can conclude is that with IGMM modeling it seems that only the necessary number of components is used which does increase the model accuracy. This is well shown by the results which are in their majority better than those obtained with the GMM approach.

REFERENCES

1. W. M. Campbell, D. E. Sturim, D. A. Reynolds, A. Solomonoff. SVM based speaker verification using a GMM supervector kernel and NAP variability compensation. *Acoustics, Speech and Signal Processing* 2006 ; 97-100.
2. Friha, S.; Mansouri. N. Application of GMMs to speech recognition using very short time series. *Intelligent Systems and Automation: 1st Mediterranean Conference on Intelligent Systems and Automation (CISA 08). AIP Conference Proceedings* 2008; 1019: 450-453.
3. Tomi Kinnunen, Haizhou Li. An overview of text-independent speaker recognition: from features to supervectors 2010; *Speech Communication*, **52**: 12-40.
4. W. M. Campbell, D. E. Sturim, D.A. Reynolds. Support vector machines using GMM supervectors for speaker verification. *IEEE Signal Processing Letters* 2006: 13(5) .308 -311.
5. Ganapathiraju A., Hamaker J., Picone J. Applications of support vector machines to speech recognition. *IEEE Transactions on Signal Processing* 2004; **52**(8) 2348-2355.
6. Jérôme Louradour, Khalid Daoudi and Francis Bach. Feature space Mahalanobis sequence kernels: application to SVM speaker verification. *IEEE Transactions on Audio Speech and Langage Processing* 2007;15(8):2465-2475.
7. M. F. Abu El yazid, M. A. El Gamal,, M. M. H. El Ayadi. On the determination of optimal model order for GMM-based text-independent speaker identification. *EURASIP Journal on Applied Signal Processing*, 2004.
8. Carl Edward Rasmussen. The infinite Gaussian mixture model. *Advances in Neural Information Processing Systems* 2000; **12:** 554-560.
9. Radford M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* 2000; 9: 249-265.
10. Tao Chen, Julian Morris and Elaine Martin. Probability density estimation via an infinite Gaussian mixture model: application to statistical process monitoring. *Appl. Statist.* 2006; **55**(5) : 699-715.
11. Chang Huai You, Kong-Aik Lee, Haizhou Li. GMM-SVM kernel with a Bhattacharyya-Based distance for speaker recognition. *IEEE Transactions on Audio, Speech & Language Processing* 2010 ;18(6): 1300-1312.
12. Ronan Collobert and Samy Bengio. SVMTorch: support vector machines for large-scale regression problems *The Journal of Machine Learning Research.* 2001; **1**: 143-160. DOI : 10.1162/15324430152733142.
13. The NIST Year 2006 Speaker Recognition Evaluation Plan. *http://www.itl.nist.gov/iad/mig/tests/spk/2006/sre-06_evalplan-v9.pdf. sre-06_evalplan-v9.doc, March 8, 2006*
14. Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei, Hierarchical Dirichlet processes in Journal of the American Statistical Association. 2006; 101(476):1566–1581.
15. Harati, A., Picone, J., & Sobel, M. Applications of Dirichlet Process Mixtures to Speaker Adaptation. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. 2012; 4321 – 4324.
16. Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, Sharing clusters among related groups: Hierarchical Dirichlet processes in Advances in Neural Information Processing Systems. 2005; Vol.17.