SUPPORT VECTOR GMMS FOR SPEAKER VERIFICATION

*Najim DEHAK , Gérard CHOLLET**

Centre de Recherche Informatique de Montréal (CRIM), École de Technologie Supérieure (ETS)
najim.dehak@crim.ca
*TSI Department, CNRS-LTCI ENST, PARIS, France
chollet@tsi.enst.fr

ABSTRACT

This article presents a new approach using the discrimination power of Support Vectors Machines (SVM) in combination with Gaussian Mixture Models (GMM) for Automatic Speaker Verification (ASV). In this combination SVMs are applied in the GMM model space. Each point of this space represents a GMM speaker model. The kernel which is used for the SVM allows the computation of a similarity between GMM models. It was calculated using the Kullback-Leibler (KL) divergence. The results of this new approach show a clear improvement compared to a simple GMM system on the NIST2005 Speaker Recognition Evaluation primary task.

1. INTRODUCTION

In the framework of text independent Automatic Speaker Verification, GMMs are the most largely used [1]. These models have the advantage of modeling well the complex distribution of the speech acoustic vectors [2]. However GMMs lack discrimination, a disadvantage in ASV. In the field of Machine Learning, SVMs are among the best discriminating models and offer good results in many applications [3].

In order to use SVMs for ASV, two classes of approaches have been experimented :

The first consists in making a combination between GMMs and SVMs. Several types of combination were proposed. We can quote the work presented in [4] which performs a discriminating training of GMMs by using a continuous density SVM. Another form of combination consists of using SVMs as a post treatment of the GMMs Models using Fisher mapping [5]. This mapping allows to obtain vectors of high dimensions where the number of dimensions is equal to the number of parameters of the GMM model. These vectors are then used by SVMs to achieve discrimination and decision. Finally, the work presented in [6] exploits the advantages of the GMM models and SVMs in a single system by deriving a probabilistic distance kernel which is computed using the divergence of KL between GMMs.

The second class of approaches consists of applying SVMs directly to the acoustic data. The method implemented in [7] trains SVMs directly on the acoustics vectors which characterize the client data and the impostor data. During testing, the segment score is obtained by averaging the scores of the SVM output for each frame. There are also others applications of SVM in ASV that used kernels sequences [8][9].

This work was initiated in the framework of the First Biosecure Residential Workshop (*http://www.tsi.enst.fr/biosecure/*). It was carried out when the first author worked at TSI departement in CNRS-LTCI ENST, Paris, France.

In this paper, we present a new approach to combine the GMM model and SVM. This approach is based on the use of a kernel calculated using a distance between GMM in the model space defined in [10]. This approach can be compared with two others methods presented in [6] and [11]. The difference between our approach and the method in [6] lies in the fact that the method given in [6] doesn't use any normalization of the GMMs models. whereas in our approach we have used a GMMs normalization based on the KL distance between GMMs. We remark that this normalization gives good improvement in the final results. Our approach is closer to that presented in [11]. The two methods exploit the same assumptions to build the kernel function. However in our approach the kernel is based on an exponential version of a distance where the approach in [11] uses a scalar product version. In [11] the authors deal with the problem of channel compensation and propose an algorithm called Nuissance Attribute Projection (NAP). In this paper, the problem of channel compensation is not dealt with.

The structure of this paper is as follow: Section 2 provides a definition of a distance between GMMs; Section 3 gives an overview of SVM and defines the probabilistic distance kernel used in our method; experimental evaluation and results are presented in section 4; Section 5 concludes the paper and gives some perspectives.

2. DISTANCE BETWEEN GMMS

In [10], the author succeeds in applying a distance between GMMs for ASV. In this approach, the decision score is not based on the mean of the log likelihood Ratio for each frame, but on the distance between GMMs.

The method consists of carrying out an estimation of test models in order to adapt the world model to the test data in the same way as the adaptation is done for the client models. A decision score is calculated using a Euclidean distance between a test model, a target model and a world model. The distance used is based on the KL divergence between two GMMs.

In the case of a Bayesian adaptation and according to a general property of the KL divergences [10] [12]. The KL divergence between GMM P and $\tilde{P}$ with $(\{w_k\},\{\mu_k\},\{\Sigma_k\})$, $\left(\{\tilde{w}_k\},\{\tilde{\mu}_k\},\left\{\tilde{\Sigma}_k\right\}\right)$ their respective parameters, is bounded by the following formula:

$$\mathrm{KL}\left(p \,\|\, \tilde{p}\right) \leq \mathrm{KL}\left(w \,\|\, \tilde{w}\right) + \sum_{k=1}^{K} \mathrm{KL}\left(N_k \,\|\, \tilde{N}_k\right) \qquad (1)$$

The term $\mathrm{KL}\left(w \,\|\, \tilde{w}\right)$ is the divergence between the masses of the weights w and $\tilde{w}$. $\mathrm{KL}\left(N_k \,\|\, \tilde{N}_k\right)$ is the divergence between the

k^{th} Gaussian of the model P and the k^{th} Gaussian of the model $\tilde{P}$.

The formula (1) is correct when there is an explicit alignment of the k^{th} Gaussians on both mixtures. This condition is implicit when we use a maximum a posteriori (MAP) adaptation technique because each Gaussian of index k of each speaker GMM is adapted from the same Gaussian k of the world model. When the means of GMMs are simply adapted from a world model and using symmetrical KL divergence [12], the formula (1) can be rewritten in the following way.

$$\begin{aligned} \mathrm{KL2}\left(p \,\|\, \tilde{p}\right) &= \mathrm{KL}\left(p \,\|\, \tilde{p}\right) + \mathrm{KL}\left(\tilde{p} \,\|\, p\right) \\ &\leq \sum_{k=1}^{K}\sum_{d=1}^{D} w_k \frac{\left(m_{k,d} - \tilde{m}_{k,d}\right)^2}{\sigma_{k,d}^2} \end{aligned} \quad (2)$$

$m_{k,d}$ and $\tilde{m}_{k,d}$ are the d's component of the means of the Gaussian N_k and $\tilde{N}_k$. $\sigma_{k,d}^2$ is the element of the common variance matrix.

In equation (2), the right term of this inequality gives a similarity measure between two GMMs, for which only the mean vectors are adapted. This term is homogeneous with the square of a Euclidean distance between points in a space defined by these parameters:

$$\lambda_{k,d} = \sqrt{w_k}\frac{m_{k,d} - m_{k,d}^{(\Omega)}}{\sigma_{k,d}} \quad (3)$$

In this new space called model space, the origin corresponds to the world model Ω and each point represents a speaker GMM. The Euclidean distance between two GMMs which can be defined in this space is given by the following formula:

$$\mathrm{D}\left(p, \tilde{p}\right) = \sqrt{\sum_{k=1}^{K}\sum_{d=1}^{D} w_k \frac{\left(m_{k,d} - \tilde{m}_{k,d}\right)^2}{\sigma_{k,d}^2}} \quad (4)$$

The square of this distance corresponds to the right term of the equation (2).

2.1. Scores in the model space

The score of decision in a system based on GMMs for the ASV is based on Log Likelihood Ratio given in the following equation:

$$S_X\left(Y\right) = \frac{1}{N}\sum_{n=1}^{N} \log \frac{\Pr\left(y_n|X\right)}{\Pr\left(y_n|\Omega\right)} \quad (5)$$

In [10], the author showed that this score can be calculated using the KL divergence.

$$S_X\left(Y\right) = \mathrm{E}\left[\log \frac{\Pr\left(y|Y_M\right)}{\Pr\left(y|\Omega\right)}\right] - \mathrm{E}\left[\log \frac{\Pr\left(y|Y_M\right)}{\Pr\left(y|X\right)}\right] \quad (6)$$

Where Y_M is a GMM model of the test segment Y.

In a similar way, the author could write the score by using the distance between the test and the world models and between the test and the target models:

$$\begin{aligned} S_X\left(Y\right) = &\mathrm{D}_E^2\Big(\Pr\left(y|Y_M\right) \,\|\, \Pr\left(y|\Omega\right)\Big) \\ &- \mathrm{D}_E^2\Big(\Pr\left(y|Y_M\right) \,\|\, \Pr\left(y|X\right)\Big) \end{aligned} \quad (7)$$

The decision is taken by comparing this new score to a threshold.

2.2. Normalization in the model space

The authors of [13][10] proposed a model-based normalization method. To carry out this normalization, a reference distance D_{ref} is set, to which the distances between the model of the world and all speakers $\mathrm{D}_E\left(X, \Omega\right)$, must be normalized. By using this reference distance the new space parameters $\lambda_{k,d}^{(X)\mathrm{norm}}$ are calculated using the initial parameters $\lambda_{k,d}^{(X)}$ as shown on the following formula:

$$\lambda_{k,d}^{(X)\mathrm{norm}} = \frac{\mathrm{D}_{ref}}{\mathrm{D}_E\left(X, \Omega\right)}\lambda_{k,d}^{(X)} \quad (8)$$

The new means of GMMs of the speakers normalized by the distance $\mathrm{D}_E\left(X, \Omega\right)$ is given by the following equation:

$$\begin{aligned} m_k^{(X)norm} = &\frac{\mathrm{D}_{ref}}{\mathrm{D}_E\left(X, \Omega\right)} m_k^{(X)} \\ &+ \left(1 - \frac{\mathrm{D}_{ref}}{\mathrm{D}_E\left(X, \Omega\right)}\right) m_k^{(\Omega)} \end{aligned} \quad (9)$$

This normalization is called M-norm by the author [10].

3. SUPPORT VECTOR MACHINES

The SVMs [14] are models used to find a separator between two classes. They are very useful to solve clustering problems that are not linearly separable. The fundamental idea is to project the input vectors using a function $\phi\left(x\right)$ to a new space called feature space. This space is of greater dimension (it can even be of infinite dimension).

The purpose of projecting the data into the feature space is to find a hyperplane which allows making a linear separation, with great generalisation properties.

In practice, SVMs use kernel functions to perform the computation of a scalar product in the feature space. These functions are also used to calculate the optimal hyperplane in the feature space without using directly the mapping function $\phi(x)$. The separating hyperplane is chosen in order to maximize the distance between the hyperplane and the training vectors closest to the border. These training vectors are called support vectors. An SVM classifier is given by the following formula:

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i K(x, x_i) + b \quad (10)$$

Where x is the vector to classify, x_i are the training examples, $k(x, x_i)$ is the kernel function and y_i correspond to the class label $y_i \in \{-1, +1\}$. α_i and b are the parameters of the model obtained in the training phase of the SVM.

3.1. Probabilistic Distance Kernels

The application of a kernel on sequential data using the KL distance was proposed first in [15], and was also applied for ASV in [6] to find a separator between two speaker models. The method consists of finding a correspondence between the data and the speaker models. The goal of this kernel is to find a similarity between densities of probability. In the ASV case, the probability densities correspond to the GMMs of the speakers. The kernel used is given by the following formula:

$$\mathrm{K}\left(\Pr\left(x|X\right),\Pr\left(x|Y\right)\right) = e^{-A\,\mathrm{D}\left(\Pr_{(x|X)},\Pr_{(x|Y)}\right)+B} \quad (11)$$

With $\mathrm{D}\Big(\Pr\left(x|X\right),\Pr\left(x|Y\right)\Big)$ corresponds to the symmetrical form of the KL divergence between the two models X and Y given segment X. A and B are factors used for numerical stability reason.

In the present work, a particular case of this kernel type is used. It is given by the following formula :

$$\mathrm{K}\left(\Pr\left(x|X\right),\Pr\left(x|Y\right)\right) = e^{-\mathrm{D}^2\left(\Pr_{(x|X)},\Pr_{(x|Y)}\right)} \quad (12)$$

With $\mathrm{D}\Big(\Pr\left(x|X\right),\Pr\left(x|Y\right)\Big)$ corresponds to the Euclidean distance between two GMM models in the model space. It was given by the equation (4).

4. EXPERIMENTS

4.1. Database

The experiments are carried out on the protocol defined in the first BioSecure Residential Workshop biosecure[1]. This protocol comprises :

- A subset from NIST2003 and NIST2004 data used for training the gender dependent background models for the GMMs,
- Another subset from NIST2004 for pseudo-imposters (77 male, 113 female) used for normalization,
- A last subset from NIST2004 for the development.

The systems were evaluated on the primary task data of the NIST2005 Speaker Recognition Evaluation campaign[2]

4.2. Front End processing

The speech parametrization is done with Mel Frequency Cepstral Coefficients (MFCC), calculated on 20 ms windows, with a 10 ms shift. For each frame, a 15-element cepstral vector is computed and appended with first order deltas and delta-energy. Cepstral mean substraction is applied to the 15 static coefficients. The mean used for the normalization is computed file by file on all the frames kept after applying the frame removal processing. Many silence frame removal processing have been tested during Biosecure residential workshop and the best results were obtained using the NIST automatic transcription files. Only bands in the 300-3400 Hz frequency range are used.

4.3. The baseline GMM system

Two gender-dependent background models are first built and then, for each target speaker, a specific GMM with diagonal covariance matrices is trained via maximum a posteriori (MAP) adaptation [16] of the Gaussian means of the matching gender background model using 25 iterations of the Expectation Maximization (EM) algorithm. Each of the two gender-dependent background model includes 512 Gaussians. This system is based on the BECARS package [17]. Two experiments were conducted with this GMM system. In the first one, the GMM system was used without any score normalization. In the second experiment, T-norm [18] score normalization was applied to this GMM system.

4.4. The distance between GMM system

In order to carry out a distance between GMMs, we used the same GMM models which were elaborate in the first two experiments. However we learned the tests segments models by making an adaptation in the same manner as the target models adaptation. We also realized a M-norm normalization of GMMs means corresponding to the targets and tests utterances as given by the equation (9). The decision score is based on the distance between GMMs. It is given by the formula (7).

4.5. The enhanced SVM/GMM system

To apply the SVM [14] in ASV, probabilistic distance kernel given by the formula (11) were used. GMMs are identical to previous experiments but we also computed GMMs for the pseudo-impostors that were used for T-norm. All GMMs are normalized by M-norm[10]. For each target speaker, SVM classifier is learned by finding a decision border between the target model and the whole of pseudo impostors corresponding to the same gender as the target speaker. This classifier is used to make a decision in the test phase. The SVM system is based on the libsvm package[3]

4.6. Results

The results of our experiments are given by DET-curves [19] in Fig. 1. We remark that the T-norm does not give significant improvement compared to the GMM system without scores normalization. However, when we used distances between GMMs with a M-norm, we obtain a slight improvement compared to GMM system without scores normalization. The best result is obtained with SVMs method. This approach gives an improvement of more than 3% of the Equal Error Rate (EER) compared to the other systems.

The application of SVM in the GMM space represents a speaker by a specific area in the model space. This area is defined by the supports vectors selected by the SVM in training step. These supports vectors correspond in reality to the speaker GMM and pseudo-impostor GMMs. These last ones can be seen as a reference model characterizing the speaker model in the model space. This new representation improves separation between speaker and impostors.

5. CONCLUSION AND PERSPECTIVES

We have proposed a new method for ASV based on a combination of SVM and GMM systems. In our approach, the input space is represented by GMM parameters. Our experiments on the NIST SRE database show that this technique gives good improvement compared to classical ones. This method can be used if the hypothesis of GMM gaussians alignment holds which is trivial in the case of MAP adaptation. Further work could use more specific GMM kernels such as Mixture Model Kernel [20] to avoid this restriction.

[1] http://www.tsi.enst.fr/biosecure/

[2] http://www.nist.gov/speech/tests/spk

[3] http://www.csie.ntu.edu.tw/~cjlin/libsvm

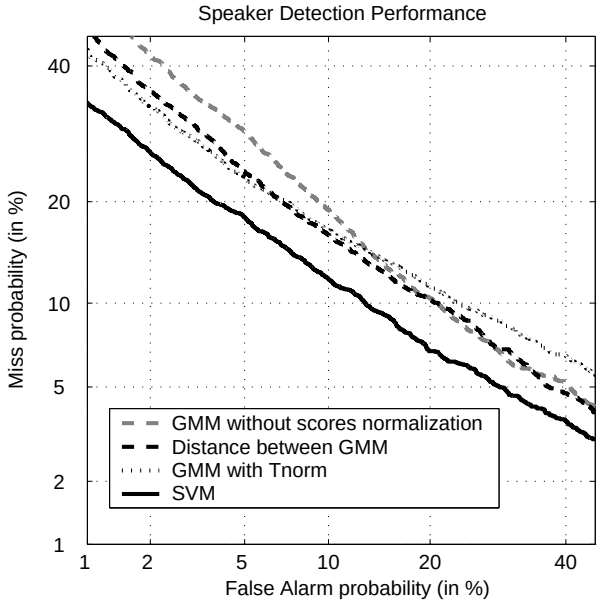

Fig. 1. DET Curves: Comparaison between SVM system, distance between GMM systems and basic GMM systems

6. ACKNOWLEDGMENTS

The authors would like to thank Asmaa ELHANNANI for her help to carry out the GMM system.

7. REFERENCES

[1] G.R. Doddington, M.A. Przybocki, A.F. Martin, and D.A. Reynolds, "The NIST Speaker Recognition Evaluation: Overview, Methodology, Systems, Results, Perspectives," in *Speech Communication*, 2000, vol. 31, pp. 225–254.

[2] 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.

[3] B. Schölkopf and A.J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*, MIT Press, 2001.

[4] X. Dong and W. Zhaohui, "Speaker Recognition using Continuous Density Support Vector Machines," *Electronics Letters*, vol. 37, no. 17, pp. 1099–1101, 2001.

[5] V. Wan and S. Renals, "SVMSVM: Support Vector Machine Speaker Verification Methodology," in *IEEE-ICASSP*, Hong Kong, 2003, vol. 2, pp. 221–224.

[6] P.J. Moreno and P.P. Ho, "A New SVM Approach to Speaker Identification and Verification Using Probabilistic Distance Kernels," in *EUROSPEECH*, 2003, pp. 2965–2968.

[7] M. Schmidt and H. Gish, "Speaker Identification via Support Vector Machines," in *IEEE-ICASSP*, 1996, pp. 105–108.

[8] V. Wan and S. Renals, "Speaker Verification Using Sequence Discriminant Support Vector Machines," *IEEE Transaction on Speech and Audio Processing*, vol. 13, no. 2, pp. 203–210, march 2005.

[9] J. Louradour and K. Daoudi, "SVM Speaker Verification Using a New Sequence Kernel," in *EUSIPCO*, 2005.

[10] M. Ben, *Approches Robustes pour la Vérification Automatique du Locuteur par Normalisation et Adaptation Hiarchique*, Ph.D. thesis, University of Rennes I, 2004.

[11] 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 *IEEE-ICASSP*, Toulouse, France, 2006.

[12] M.N. Do, "Fast Approximation of Kullback-Leibler Distance for Dependence Trees and Hidden Markov Models," *IEEE Signal Processing Letters*, vol. 10, no. 4, pp. 115–118, 2003.

[13] M. Ben and F. Bimbot, "D-MAP: a Distance-Normalized MAP Estimation of Speaker Models for Automatic Speaker Verification," in *IEEE-ICASSP*, 2003, vol. 2, pp. 69–72.

[14] V.N. Vapnick, *Statistical Learning Theory*, Wiley, 1998.

[15] P.J. Moreno, P.P. Ho, and N. Vasconcelos, "A Generative Model Based Kernel for SVM Classification in Multimedia Applications," in *NIPS*, 2003.

[16] J.L. Gauvain and C.H. Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Transaction on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, April 1994.

[17] R. Blouet, C. Mokbel, H. Mokbel, E. Sanchez Soto, G. Chollet, and H. Greige, "BECARS : a Free Software for Speaker Verification," in *ODYSSEY*, Spain, 2004, pp. 145–148.

[18] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score Normalization for Text-Independent Speaker Verification Systems," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 42–54, 2000.

[19] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET Curve in Assessment of Detection Task Performance," in *EUROSPEECH*, 1997, vol. 4, pp. 1895–1898.

[20] T. Jebara and R. Kondor, "Bhattacharyya and Expected Likelihood Kernels," in *COLT/KERNEL*, Washington DC, USA, 2003, pp. 57–71.