Linear and Non Linear Kernel GMM SuperVector Machines for Speaker Verification

Réda Dehak[1]*, Najim Dehak*[2,3]*, Patrick Kenny*[2]*, Pierre Dumouchel* [2,3]

[1]Laboratoire de Recherche et de Développement de l'EPITA (LRDE), Paris, France
[2] Centre de Recherche en Informatique de Montréal (CRIM), Montréal, Canada
[3] École de Technologie Supérieure (ETS), Montréal, Canada
reda.dehak@lrde.epita.fr, {najim.dehak,patrick.kenny,pierre.dumouchel}@crim.ca

Abstract

This paper presents a comparison between Support Vector Machines (SVM) speaker verification systems based on linear and non linear kernels defined in GMM supervector space. We describe how these kernel functions are related and we show how the nuisance attribute projection (NAP) technique can be used with both of these kernels to deal with the session variability problem. We demonstrate the importance of GMM model normalization (M-Norm) especially for the non linear kernel. All our experiments were performed on the core condition of NIST 2006 speaker recognition evaluation (all trials). Our best results (an equal error rate of 6.3%) were obtained using NAP and GMM model normalization with the non linear kernel.

Index Terms: Speaker recognition, Gaussian mixture model, support vector machine, Linear kernel, Non linear kernel, Nuisance attribute projection, M-Norm.

1. Introduction

In text independent speaker verification applications, the principal state-of-the-art approach is based on cepstral Gaussian Mixture Models (GMM) [1]. Speaker detection is based on a likelihood ratio calculated using a Maximum A-Posteriori (MAP) adapted GMM from a Universal Background Model (UBM). In recent NIST Speaker Recognition Evaluation (NIST-SRE) campaigns, new discriminative approaches were combined with GMM systems to improve performance. An exciting development is the use of Support Vector Machines [2] for speaker verification.

In this paper, we present an extension of the work in [3] which consists in applying support vector machines in GMM supervector space using a non linear kernel. This kernel is based on Kullback-Leibler (KL) divergence between two GMMs. A similar approach based on a linear kernel was proposed by Campbell *et. al.* [4]. These two kernels exploit the same assumptions to build the kernel function. The non linear kernel is based on an exponential version of a KL distance approximation whereas the linear kernel is just the scalar product corresponding to this distance.

In this work, we compare the linear kernel and the non linear kernel methods for speaker verification. We have used the SVM Nuisance Attribute Projection (NAP) method proposed by [5] with both kernels to deal with the channel compensation problem. When session variability is regarded as the nuisance variability, this approach consists in finding a projection matrix that minimize the distance between the mapping vectors corresponding to the same speaker in the feature space. In our experiments, we used the same projection for both kernels. This projection is applied in the GMM supervectors space which corresponds to the feature space of linear kernel and the input space of non linear kernel. (The feature space of our non-linear kernel is actually infinite dimensional.) We also explore the effect of GMM normalization (M-Norm) [6] in the two cases. All our experiments are performed on the core condition (all trials) of the NIST 2006 Speaker Recognition Evaluation (SRE).

The outline of the paper is as follows. Section 2 describes the basic framework for SVMs. We define the two kernels and present the distance metrics technique used in the two approaches to derive the kernel functions. In section 3, we present the M-Norm method of normalizing GMM parameters. To deal with session variability, section 4 presents the nuisance attribute projection method. The comparison results on the core condition of NIST-SRE 2006 is presented in section 5. Section 6 concludes the paper and gives some perspectives.

2. GMM-SVM for Speaker Recognition

An SVM [7] is a two-class classifier based on a hyperplane separators. This separator is chosen in order to maximize the distance between the hyperplane and the closest training vectors. These training vectors are called support vectors. SVMs usually operate in a high dimensional feature space (potentially with infinite dimension), non linearly related with a mapping function ϕ to the original input feature space X. Given an observation $x \in X$ and a mapping function ϕ, an SVM discriminant function is given by

$$f(x) = \sum_{k=1}^{M} \lambda_k y_k < \phi(x) \ , \ \phi(x_k) > + b \qquad (1)$$

$$= \sum_{k=1}^{M} \lambda_k y_k K(x, x_k) + b \qquad (2)$$

Here, the function $K(x_i, x_j)$ is the kernel function; it is defined as an inner product $< \phi(x_i) \ , \ \phi(x_j) >$ (so that the Mercer condition is satisfied) although the 'kernel trick' avoids evaluating the mapping function $\phi(\cdot)$. The x_k's in (1) are the support vectors and the y_k's are the corresponding target class values ± 1. M is the number of support vectors and the λ_k's are obtained through a training process.

The distance in the feature space between the two corresponding mapping vectors $\phi(x_i)$ and $\phi(x_j)$ can be evaluated using the kernel function by:

$$\mathcal{D}\left(\phi\left(x_i\right), \phi\left(x_j\right)\right) = \sqrt{K(x_i, x_i) - 2K(x_i, x_j) + K(x_j, x_j)} \qquad (3)$$

2.1. GMM Supervector

We assume we are given a Gaussian mixture model universal background model (GMM-UBM)

$$p(x) = \sum_{i=1}^{N} w_i \mathcal{N}(x; \mu_i \Sigma_i) \tag{4}$$

where w_i are the mixture weights, $\mathcal{N}$ indicates a Gaussian density and μ_i and Σ_i are the corresponding mean and covariance.

From a speaker utterance, the GMM-UBM model is adapted by Maximum A-Posteriori (MAP) adaptation to provide the speaker GMM model. Generally, only the means μ_i of Gaussian components are adapted. In this case, all GMMs have the same covariance matrices Σ_i and differ only in means. As a consequence, for SVM classification, each model is represented only by the concatenation of all GMM Gaussians mean vectors, that is, a GMM *supervector*.

2.2. Distance Between GMMs

Given two probabilistic models p^a and p^b corresponding, respectively, to two speakers s_a and s_b, we can define a distance $\mathcal{D}$ between this two models using the KL divergence:

$$KL(p^a \| p^b) = \int_{\mathbb{R}^n} p^a(x) \log \left(\frac{p^a(x)}{p^b(x)} \right) \mathrm{d}x. \tag{5}$$

The KL distance ($KL2$) is a symmetrized version of this

$$KL2(p^a \| p^b) = KL(p^a \| p^b) + KL(p^b \| p^a) \tag{6}$$

As shown in [8], the KL divergence between two GMMs is upper bounded by

$$KL(p^a \| p^b) \leq KL\left(w^a \| w^b\right) + \sum_{i=1}^{N} w_i^a \, KL\left(\mathcal{N}\left(.; \mu_i^a, \Sigma_i^a\right) \| \mathcal{N}\left(.; \mu_i^b, \Sigma_i^b\right)\right) \tag{7}$$

In the case of MAP adaptation and when only the means of GMMs were adapted from the GMM-UBM (i.e. $w^a = w^b$ and $\Sigma_i^a = \Sigma_i^b, i = 1..N$), the KL distance ($KL2$) can be upper bounded as follow [8]:

$$KL2(p^a \| p^b) \leq \sum_{i=1}^{N} w_i \left(\mu_i^a - \mu_i^b\right) \Sigma_i^{-1} \left(\mu_i^a - \mu_i^b\right)^t \tag{8}$$

$$\leq \mathcal{D}_e^2\left(\boldsymbol{\mu^a}, \boldsymbol{\mu^b}\right) \tag{9}$$

where

$$\mathcal{D}_e^2\left(\boldsymbol{\mu^a}, \boldsymbol{\mu^b}\right) = \sum_{i=1}^{N} w_i \left(\mu_i^a - \mu_i^b\right) \Sigma_i^{-1} \left(\mu_i^a - \mu_i^b\right)^t \tag{10}$$

In the case of diagonal covariance matrices (which is the case of the state-of-the-art of speaker verification), the distance $\mathcal{D}_e^2$ between GMM supervectors $\boldsymbol{\mu^a}$ and $\boldsymbol{\mu^b}$ represents a weighted Euclidean distance between scaled version of $\boldsymbol{\mu^a}$ and $\boldsymbol{\mu^b}$. It is an upper bound of the KL distance. So if the distance between $\boldsymbol{\mu^a}$ and $\boldsymbol{\mu^b}$ is small, the corresponding KL distance is small. The authors in [6] succeed in applying this distance in speaker clustering applications.

2.3. Linear Kernel

The application of a kernel on sequential data using the KL distance was proposed first in [9], and was applied for speaker verification in [4, 3] to find a separator between the speaker models and impostor models.

The linear kernel was proposed by Campbell *et. al.* [4]. The main idea is to derive from the distance $\mathcal{D}_e^2\left(\mu^a, \mu^b\right)$ defined in (10) the corresponding inner product which is the kernel function:

$$K_{lin}(s_a, s_b) = \sum_{i=1}^{N} w_i \, \mu_i^a \, \Sigma_i^{-1} \left(\mu_i^b\right)^t = \sum_{i=1}^{N} \left(\sqrt{w_i}\Sigma_i^{-\frac{1}{2}}\mu_i^a\right)\left(\sqrt{w_i}\Sigma_i^{-\frac{1}{2}}\mu_i^b\right)^t \tag{11}$$

Note that if we apply the linear kernel K_{lin} in (3), we obtain the distance defined in (10).

The kernel K_{lin} is linear in the GMM supervectors. The feature space represents a simple diagonal scaling (using $\sqrt{w_i}\Sigma_i^{\frac{1}{2}}$) of the GMM supervectors space. As a consequence, it satisfies the Mercer condition [7].

2.4. Non Linear Kernel

Another way to obtain a kernel from a distance function $\mathcal{D}$, is to use the exponential function [7]:

$$K_{nonlin}(s_a, s_b) = e^{-\mathcal{D}_e^2\left(\boldsymbol{\mu^a}, \boldsymbol{\mu^b}\right)} \tag{12}$$

This non linear kernel was proposed by Dehak and Chollet in [3]. They used the distance defined by (10) as a distance between two speaker models.

The non linear kernel is different from the linear one. It represents the normalized version of the exponential of the linear kernel [7]. Consequently their derived distances in feature space are different. As we mentioned above, the distance in feature space corresponding to the linear kernel is the Euclidean distance defined by (10). For the non linear kernel, it can be derived using (3):

$$\mathcal{D}\left(\phi\left(s_a\right), \phi\left(s_b\right)\right) = \sqrt{2 - 2e^{-\mathcal{D}_e^2\left(\boldsymbol{\mu^a}, \boldsymbol{\mu^b}\right)}} \tag{13}$$

The non linear kernel is a Gaussian kernel defined on the GMMs supervector space. The corresponding feature space of Gaussian kernel has infinite-dimension [7]. The mapping function $\phi(.)$ is equivalent to Gaussian kernel mapping function [7]:

$$\boldsymbol{\mu} \mapsto \phi(\boldsymbol{\mu}) = K(\boldsymbol{\mu}, .) = e^{-\frac{\|\boldsymbol{\mu} - . \|^2}{2\sigma^2}} \tag{14}$$

3. Models Normalization : M-norm

The authors in [3] note that there is some improvement in the performance of non linear kernel system if the model normalization (M-Norm) technique [6] is used. The objective of this method is to modify the GMM mean vectors so that the distance between all normalized models and the UBM-GMM Ω is a constant, which we can take to be 1. Let $\{\mu_k^\Omega\}$ be the set of UBM mean vectors and, for a given speaker X, let $\{\mu_k^X\}$ be the set of mean vectors in the speaker GMM. Denote by $\mathcal{D}_e(X, \Omega)$ the distance between the speaker GMM and the UBM. Applied to a particular mean vector μ_k^X, the normalization procedure is

$$\mu_k^X \leftarrow \frac{1}{\mathcal{D}_e\left(X, \Omega\right)} \mu_k^X + \left(1 - \frac{1}{\mathcal{D}_e\left(X, \Omega\right)}\right) \mu_k^\Omega \tag{15}$$

4. SVM Nuisance Attribute Projection (NAP)

The principal objective of the SVM nuisance attribute projection (NAP) method, proposed in [5, 4], is to reduce the impact of channel and handset variations on system performances. It uses an appropriate projection matrix P in the feature space to remove subspaces that cause variability in kernel. NAP constructs a new kernel of the following form:

$$\begin{aligned} K(s_a, s_b) &= < P\phi(s_a)\ ,\ P\phi(s_b) > \\ &= \phi(s_a)^t P\phi(s_b) \\ &= \phi(s_a)^t (I - VV^t)\phi(s_b). \end{aligned} \tag{16}$$

where V is a rectangular matrix of low rank whose columns are orthornormal. If we express V in terms of its columns, $V = [v_1, v_2, ...v_k]$, the vectors v_i are the directions which are removed from the feature space.

The design criterion for P and the corresponding matrix V is

$$\tilde{P} = \arg\min_P \sum_{i,j} W_{i,j} \|P\phi(s_i) - P\phi(s_j)\|^2 \tag{17}$$

where $\phi(.)$ is identity mapping function for linear kernel and as given in equation 14 for non linear one. The $\{s_i\}$ are typically a background data set. The $W_{i,j}$ matrix can be defined in several different ways. If we want a projection based on session variability, we pick $W_{i,j} = 1$ if s_i and s_j correspond to the same speaker, and $W_{i,j} = 0$ otherwise (see [5] for other cases).

Campbell *et. al.* [4] noticed that with the linear kernel and session variability as nuisance variable, the NAP subspace is equivalent to the channel subspace of factor analysis [10]. In this case, the solution $\tilde{P}$ corresponds to the projection matrix which minimizes the Euclidean distance ($\|.\|^2$) between GMM supervectors belonging to the same speaker. In this case, the solution of equation (17) ($\tilde{P}$ and corresponding $\tilde{V}$) is obtained from the k eigenvectors having the k largest eigenvalues of the following covariance matrix:

$$\mathcal{C} = \frac{1}{S} \sum_{s=1}^{S} \frac{1}{n_s} \sum_{i=1}^{n_s} (\tilde{\boldsymbol{\mu}}_\mathbf{i}^\mathbf{s} - \bar{\mu}_s)(\tilde{\boldsymbol{\mu}}_\mathbf{i}^\mathbf{s} - \bar{\mu}_s)^t \tag{18}$$

where $\tilde{\boldsymbol{\mu}}_\mathbf{i}^\mathbf{s}$ represents the GMM supervector corresponding to the i^{th} session of the s^{th} speaker; S is the number of speaker in our background database; n_s represents the number of s^{th} speaker sessions; and $\bar{\mu}_s$ represents the mean of s^{th} speaker supervectors:

$$\bar{\mu}_s = \frac{1}{n_s} \sum_{i=1}^{n_s} \tilde{\boldsymbol{\mu}}_\mathbf{i}^{\mathbf{n_s}} \tag{19}$$

In the present work, we apply the same projection matrix $\tilde{P}$ in GMM supervector space for both the linear and non-linear kernels. In the case of the non linear kernel, another way of solving (17) would be to use Kernel Principal Component Analysis (Kernel PCA). Kernel PCA can solve this problem and avoids the mapping function evaluation. We will explore this in a future work.

5. Experiments

5.1. Test database

We performed our experiments on the core condition of NIST-SRE 2006 corpus (all trials)[1]. The train and test utterances contain 2.5 minutes of speech on average. The whole speaker detection task consists of 53966 tests (3612 target tests). We use equal error rate (EER) and the minimum decision cost value (minDCF) as metrics for performance evaluation.

[1] See `http://www.nist.gov/speech/tests/spk/2006/` for more details

5.2. Front End Processing

The speech parametrization is done as follows. A 16-dimensional Linear Frequency Cepstral Coefficients (LFCC) is extracted from speech signal every 10ms using a 20ms Hamming window. First order deltas and delta-energy are appended to the cepstral vector. Finally, mean and variance normalization are applied to each feature of the 33-dimensional final vector.

5.3. GMM-UBM Baseline System

Two gender-dependent background models are built. Each GMM-UBM consists of 512 mixture components. They was trained using the Baum Welch algorithm from a corpus extracted from Fisher English database Part 1 and 2 and NIST 2003. For each target speaker, a specific GMM with diagonal covariance matrices is trained via MAP adaptation of Gaussian means of the matching gender background model. T-norm score normalization was applied to this GMM system score. We used a corpus of 449 male and 486 female impostors extracted from NIST-SRE 2004 and Fisher databases. This baseline system was the LRDE submission in the NIST-SRE 2006 campaign [11].

5.4. SVM-GMM System

For the first experiment, our SVM systems used the same GMMs as in the Baseline GMM-UBM system. Each SVM is trained using the single available positive example adapted from the target speaker segment. The impostor models used during GMM-UBM baseline system score normalization are used as negative examples. This classifier is used to make decision during the test phase. SVM-GMM systems scores are not normalized.

For the second experiment, M-Norm Model normalization (Section 3) was used to normalize the speaker, test and impostor models. We have tested the influence of this step on the performance of each system.

5.5. NAP Channel Compensation

To deal with session variability problem, we use NAP to create a projection matrix $\tilde{P} = I - \tilde{V}\tilde{V}^t$ which projects points in the d-dimensional feature space (that is, the GMM supervector space so that $d = 512 \times 33$) onto a subspace that is hopefully more resistant to session variability effects. We project all target, test and impostor GMM supervectors onto this subspace.

A corpus extracted from NIST-SRE 2004 database was used to compute the covariance matrix $\mathcal{C}$ and obtain the $d \times k$ matrix $\tilde{V}$. We tested the performance for different values of k (the corank of the matrix $\tilde{P}$).

In our third experiment, the projected GMM supervectors are used instead of original ones and, in our final experiment, M-Norm was used with NAP to normalize the projected speaker, test and impostor models.

5.6. Results and Discussion

Systems performances (EER and MinDCF) of the first and second experiences are presented in Table 1. We note the effectiveness of model normalization in the case of non linear kernel —

	EER	MinDCF
GMM-UBM baseline system	9.11%	0.044
linear kernel	8.04%	0.038
non linear kernel	10.46%	0.048
linear kernel **with M-Norm**	8.08%	0.037
non linear kernel **with M-Norm**	8.50%	0.040

Table 1: Comparison between GMM-UBM and GMM-SVM systems performance with and without M-Norm. No NAP. NIST 2006 SRE core condition (all trials).

an improvement of $\cong 2\%$ absolute in EER. This operation has a little influence on the linear kernel performance. The performance of linear kernel is a little bit better than the non linear one with M-norm. Note that the SVM linear kernel system results are always better than the GMM-UBM baseline system results. However the SVM non linear kernel system are better than GMM-UBM ones only when M-norm is used.

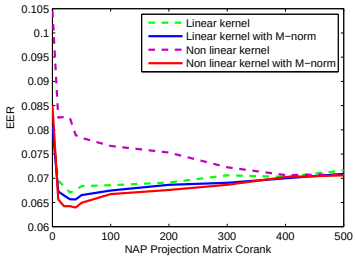

Figure 1: The influence of NAP corank on GMM-SVM supervector systems. NIST 2006 SRE core condition (all trials).

	EER	MinDCF
linear kernel	6.75%	0.031
non linear kernel	7.88%	0.036
linear kernel **with M-Norm**	6.56%	0.030
non linear kernel **with M-Norm**	**6.39%**	0.029

Table 2: Combination of M-Norm and NAP (corank=40) with GMM-SVM systems. NIST 2006 SRE core condition (all trials).

We plot on Figure 1 the influence of the corank of matrix P (i.e. the second dimension of the $d \times k$ matrix V) on the EER of both the linear and non linear kernel systems with and without M-Norm. With the linear kernel, NAP improves performance better when M-Norm is used. Our best improvement ($k = 40$, Fig. 1) is obtained with the non linear kernel with M-norm (Tab. 2).

In the final experiment, we obtained our best EER (6.39%) (Tab. 2) using the non linear kernel with $k = 40$ in nuisance attribute projection. The corresponding DET curves are plotted on Figure 2.

6. Conclusions and Perspectives

In this paper, we compared linear and non linear kernels for SVM-based speaker recognition. We obtained slightly better results with the non-linear kernel than with the linear kernel. We demonstrated the effectiveness of combining M-Norm model normalization with nuisance attribute projection for both kernels. We used the same implementation of NAP for both kernels, namely the natural implementation for the linear kernel, and we showed that this type of NAP was also effective with the non-linear kernel. This result suggests that it would be worthwhile to explore an alternative implementation of NAP based on kernel principal components analysis which would be more natural for the non-linear kernel. We will address this question in future work.

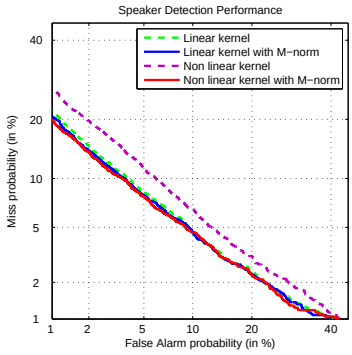

Figure 2: Using NAP and M-Norm with GMM-SVM systems. NIST 2006 SRE core condition (all trials).

7. References

[1] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker Verification using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[2] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer, 1995.

[3] N. Dehak and G. Chollet, "Support Vector GMMs for Speaker Verification," in *IEEE Odyssey*, San Juan, Puerto Rico, 2006.

[4] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, "SVM Based Speaker Verification using a GMM Supervector Kernel and NAP Variability Compensation," in *ICASSP*, vol. 1, 2006, pp. 97–100.

[5] A. Solomonoff, W. Campbell, and I. Boardman, "Advances in Channel Compensation For SVM Speaker Recognition," in *ICASSP*, vol. 1, 2005, pp. 629–632.

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

[7] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambrige, 2004.

[8] M. Do, "Fast Approximation of Kullback-Leibler Distance for Dependence Trees and Hidden Markov Models," *IEEE Signal Processing Letters*, pp. 115–118, 2003.

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

[10] P. Kenny and P. Dumouchel, "Experiments in Speaker Verification Using Factor Analysis Likelihood Ratios," in *Odyssey*, 2004, pp. 219–226.

[11] R. Dehak, C. Deledalle, and N. Dehak, "LRDE-EPITA System Description," in *NIST Speaker Recognition Evaluation*, 2006.