SUPPORT VECTOR MACHINES AND JOINT FACTOR ANALYSIS FOR SPEAKER VERIFICATION

Najim Dehak[1,2], Patrick Kenny[1], Réda Dehak[3], Ondrej Glembek[4] , Pierre Dumouchel[1,2], Lukas Burget[4], Valiantsina Hubeika[4] and Fabio Castaldo[5]

[1]Centre de recherche informatique de Montréal (CRIM), Montréal, Canada
[2]École de Technologie Supérieure (ETS), Montréal, Canada
[3]Laboratoire de Recherche et de Développement de l'EPITA (LRDE), Paris, France
[4]Speech@FIT group, Faculty of Information Technology, Brno University of Technology, Czech Republic
[5]Politecnico di Torino, Turin, Italy

{najim.dehak,patrick.kenny,pierre.dumouchel}@crim.ca,reda.dehak@lrde.epita.fr
{glembek,burget}@fit.vutbr.cz,v.hubeika@gmail.com,fabio.castaldo@polito.it

ABSTRACT

This article presents several techniques to combine between Support vector machines (SVM) and Joint Factor Analysis (JFA) model for speaker verification. In this combination, the SVMs are applied to different sources of information produced by the JFA. These informations are the Gaussian Mixture Model supervectors and speakers and Common factors. We found that using SVM in JFA factors gave the best results especially when within class covariance normalization method is applied in order to compensate for the channel effect. The new combination results are comparable to other classical JFA scoring techniques.

Index Terms— Joint Factor Analysis, Support Vector Machine, Speaker factors space, Within Class Covariance Normalization.

1. INTRODUCTION

Over the last three years, the Joint Factor Analysis (JFA) [1] approach has become the state of the art in the speaker verification field. This modeling was proposed in order to deal with the speaker and channel variabilities in the Gaussian Mixture Models (GMM) [2] framework.

During the same period, the application of the Support Vector Machine (SVM) in GMM supervector space [3] also led to interesting results, especially when the Nuisance Attribute Projection (NAP) was applied to deal with the channel effect. In this approach, the kernel is based on a linear approximation of the Kullback Leibler (KL) distance between two GMMs. The speaker GMM means supervectors were obtained by adapting the Universal Background Model (UBM) supervector to speaker frames using Maximum A Posterior (MAP) adaptation [2].

In this paper, we propose to combine the SVM with JFA. We tried two types of combinations; the first one uses the GMM supervector obtained with JFA as input to the SVM using the classical linear KL kernel between two supervectors. The second, rather than using the GMM supervectors as features for the SVM, directly uses the information given by the speaker and common factors components (see section 2) defined by the JFA model.

The outline of the paper is as follows. Section 2 describes the factor analysis model. In section 3, we present the JFA-SVM approach and we describe all the kernels used to implement it. The comparison between different results is presented in section 5. Section 6 concludes the paper.

2. JOINT FACTOR ANALYSIS

Joint factor analysis is a model used to treat the problem of speaker and session variability in GMM's. In this model, each speaker is represented by the means, covariances, and weights of a mixture of C multivariate diagonal-covariance Gaussian densities defined in some continuous feature space of dimension F. The GMM for a target speaker is obtained by adapting the UBM means parameters (UBM). In joint factor analysis [1, 4, 5], the basic assumption is that a speaker and channel-dependent supervector M can be decomposed into a sum of two supervectors: a speaker supervector s and a channel supervector c

$$M = s + c \tag{1}$$

where s and c are normally distributed.

In [1], Kenny et al. described how the speaker-dependent and channel-dependent supervector can be represented in low dimensional spaces. The first term in the right hand side of (1) is modeled by assuming that if s is the speaker supervector for a randomly chosen speaker then

$$s = m + dz + Vy \tag{2}$$

Where m is the speaker and channel independent supervector (UBM), d is a diagonal matrix, V is a rectangular matrix of low rank and y and z are independent random vectors having standard normal distributions. In other words, s is assumed to be normally distributed with mean m and covariance matrix $VV^t + d^2$. The components of y and z are respectively the speaker and common factors.

The channel-dependent supervector c, which represents the channel effect in an utterance, is assumed to be distributed according to

$$c = ux \tag{3}$$

Where u is a rectangular matrix of low rank and x is a standard normal distribution. This is equivalent to saying that c is normally distributed with zero mean and covariance uu^t. The components of x are the channel factors in factor analysis modeling.

3. SVM-JFA

The SVM is a classifier used to find a separator between two classes. The main idea of this classifier is to project the input vectors into a high dimension space called feature space in order to find linear separation. This projection is carried out using a mapping function. In practice, SVMs use kernel functions to perform the scalar product computation in the feature space. These functions allow us to compute directly the scalar product in the feature space without defining the mapping function.

In this section, we will present several ways to carry out the combination between the SVM and JFA. The first approach is similar to the classical SVM-GMM [3, 6] when the speaker GMM supervectors are used as input to SVM. We have tested a second set of methods based on a new kernel that uses i) spekaer factors or ii) speaker and common factors, depending on the configuration of the JFA model.

3.1. GMM Supervector space

In order to apply SVM with JFA using speaker supervector as input, we used the classical linear Kullback- Leibler kernel. This kernel applied in GMM supervector space is based on Kullback-Leibler divergence between two GMMs [3]. This distance corresponds to the Euclidean distance between scaled GMM supervectors s and s'.

$$\mathcal{D}_e^2\left(s, s'\right) = \sum_{i=1}^{C} w_i \left(s_i - s_i'\right) \Sigma_i^{-1} \left(s_i - s_i'\right)^t \tag{4}$$

where w_i and Σ_i are the i^{th} UBM mixture weights and diagonal covariance matrix, s_i corresponds to the mean of Gaussian i of the speaker GMM. The derived linear kernel is defined as the corresponding inner product of the preceding distance

$$K_{lin}\left(s, s'\right) = \sum_{i=1}^{C} \left(\sqrt{w_i}\Sigma_i^{-\frac{1}{2}} s_i\right) \left(\sqrt{w_i}\Sigma_i^{-\frac{1}{2}} s_i'\right)^t \tag{5}$$

This kernel was proposed by Campbell et al. [3].

3.2. Speaker factors space

In this section, we discuss the use of speaker factors as parameters input to SVM. The speaker factor coefficients correspond to speaker coordinates in the speaker space defined by the eigenvoices matrix. The advantage of using speaker factors is that these vectors are of low dimensions (typically dimension = 300), making the decision process faster. We tested these vectors with three classical kernels which are linear, Gaussian and cosine kernels. These kernels are respectively given by the following equations:

$$k(y_1, y_2) = \langle y_1, y_2 \rangle \tag{6}$$

$$k(y_1, y_2) = \exp\left(-\frac{1}{2.\sigma^2} \left\|y_1 - y_2\right\|^2\right) \tag{7}$$

$$k(y_1, y_2) = \frac{\langle y_1, y_2 \rangle}{\left\|y_1\right\| \left\|y_2\right\|} \tag{8}$$

The motivation of using the linear kernel is that the speaker factor vectors are normally distributed with zero mean and identity variance matrix. In order to obtain the speaker factors for this system, we used the JFA configuration which contains the speaker and channel factors only. There are no common factors z (see equation 2).

3.2.1. Within Class Covariance Normalization

In this new approach, we propose to apply another channel compensation step in the speaker factors space. The first step is carried out by estimating the channel factors in GMM supervector space. To achieve this compensation, two choices are possible. The first one is the NAP [3] algorithm and the second is the Within Class Covariance Normalization algorithm (WCCN) [7]. We decided to apply WCCN algorithm rather than NAP because NAP algorithm realizes channel compensation by removing the nuisance directions; however the speaker factors are vectors of low dimension so removing additional directions could be harmful.

The WCCN algorithm uses the Within Class Covariance (WCC) matrix to normalize the kernel functions in order to compensate for the channel factor without removing any directions in the space. WCC matrix is obtained by the following formula:

$$W = \frac{1}{S} \sum_{s=1}^{S} \frac{1}{n_s} \sum_{i=1}^{n_s} \left(y_i^s - y_s\right) \left(y_i^s - y_s\right)^t \tag{9}$$

where $y_s = \frac{1}{n_s} \sum_{i=1}^{n_s} y_i^s$ is mean of speaker factors vectors of each speaker, S is the number of speaker and n_s is number of utterances for each speaker s.

The WCCN algorithm was applied to the linear and cosine kernels. The new versions of the two previous kernels are given by the following equations:

$$k(y_1, y_2) = y_1^t W^{-1} y_2 \tag{10}$$

$$k(y_1, y_2) = \frac{y_1^t W^{-1} y_2}{y_1^t W^{-1} y_1 \; y_2^t W^{-1} y_2} \tag{11}$$

3.3. Speaker and Common factors space

When the speaker and common factors are available, we propose and compare two techniques that combine these two sources of information. The first approach applies SVM in each space (speaker factors space and common factors space). Thereafter we linearly combine these two SVMs scores. The fusion weights are obtained by using a logistic regression [8]. The second approach is to define a new kernel which is the linear combination of two kernels. The first kernel is applied in the speaker factors space while the second kernel is applied in the common factors space. The kernel combination weights are chosen to maximize the margin between target speaker and impostors utterances. This technique was already applied in speaker verification [9].

4. EXPERIMENTAL SETUP

4.1. Test set

The results of our experiments are reported in the core condition of the NIST 2006 speaker recognition evaluation (SRE) dataset [10]. For score fusion system case, the weights were trained on the NIST 2006 SRE dataset. The systems were tested on the telephone data of the core condition of the NIST 2008 SRE.

4.2. Acoustic features

In our experiments, we used cepstral features extracted using a $25ms$ Hamming window. 19 mel frequency cepstral coefficients together with log energy are calculated every $10ms$. This 20-dimensional feature vector was subjected to feature warping [11] using a $3s$ sliding window. Delta and double delta coefficients were then calculated

Table 1. Comparison results between SVM-JFA in GMM supervectors space and JFA frame by frame scoring. The results are given on EER in the core condition of the NIST 2006 SRE.

System	English	All trials
JFA: $s = m + Vy$	1.95%	3.01%
JFA: $s = m + Vy + dz$	1.80%	2.96%
SVM-JFA: $s = m + Vy$	4.24%	4.98%
SVM-JFA: $s = m + Vy + dz$	4.23%	4.92%

using a 5 frame window corresponding to a 60-dimensional feature vectors. These feature vectors were modeled using GMM and factor analysis was used to treat the problem of speaker and session variability.

4.3. Factor analysis training

We used gender independent Universal Background Models which contains 2048 Gaussians. This UBM was trained using LDC releases of Switchboard II, Phases 2 and 3; switchboard Cellular, Parts 1 and 2 and NIST 2004-2005 SRE. The (gender independent) factor analysis models were trained on the same quantities of data as the UBM.

The decision scores obtained with the factor analysis were normalized using zt-norm normalization. We used 148 male and 221 female t-norm models and 159 male and 201 female z-norm utterances.

We used two factor analysis configurations. The first JFA used restricted configuration composed only with 300 speaker factors and 100 channel factors and the second one was a full JFA configuration; we added the diagonal matrix (d) in order to have speaker and common factors.

4.4. SVM impostors

We used 1875 gender independent impostors to train the SVM model. These impostors are taken from LDC releases of Switchboard II, Phases 2 and 3; switchboard Cellular, Parts 1 and 2 and NIST 2004-2005 SRE.

4.5. Within Class Covariance

We used a gender independent within class covariance matrix which is trained in the same dataset as the JFA training.

5. RESULTS

5.1. SVM-JFA: GMM supervector space

We start with the results obtained by the combination SVM-JFA when the GMM supervectors are used as input to the SVM. We used GMM supervector obtained using both JFA configurations (with or without Common factors). The results are given in Table 1. These results are compared to the frame by frame JFA scoring techniques.

The results show that the performances of the application of the SVM in the GMM supervector space are significantly worse than these obtained by the conventional frame by frame JFA scoring. These results can be explained by the fact that the linear KL kernel is not appropriate for GMM supervectors obtained by the JFA model because the assumption of independence of GMM Gaussians in the case of MAP adaptation is not true for the adaptation based on eigenvoices. The results show also that the addition of common factors didn't improve the results in the case of SVM-JFA compared to the JFA scoring.

Table 2. Comparison results between SVM-JFA in speaker factor space and GMM supervectors space. The Results are given on EER in the core condition of the NIST 2006 SRE.

	English		All trials	
	No-norm	T-norm	No-norm	T-norm
KL-kernel: GMM supervectors	-	4.24%	-	4.98%
Linear kernel	3.47%	2.93%	4.64%	4.04%
Gaussian kernel	3.03%	2.98%	4.59%	4.46%
Cosine kernel	3.08%	2.92%	4.18%	4.15%

Table 3. Comparison results between SVM-JFA in speaker factor space (with and without WCCN) with two JFA scoring techniques. The results are given on EER in the core condition of the NIST 2006 SRE, English trials.

	Without WCCN		With WCCN	
	t-norm	zt-norm	t-norm	zt-norm
Linear kernel	2.93%	-	2.44%	-
Cosine kernel	2.92%	-	2.43%	-
JFA frame by frame scoring	2.81%	1.95%	-	-

5.2. SVM-JFA: speaker factors space

This section present the results obtained with the linear, Gaussian and cosine kernels applied in speaker factors space. We compare these new results with the last one using the SVM- JFA applied in GMM supervectors. Table 2 gives these results.

Three remarks are in order in Table 2. The first one is that the application of the SVM in speaker factors space gave better results than applied SVM in GMM supervectors space. The second is that there is marked linear separation between the speakers if we compare the results between cosine and Gaussian kernel. The last remark is that t-norm did not yield a large improvement in the case of the cosine and Gaussian kernels, however it helps in the case of the linear kernel.

5.2.1. Within Class Covariance Normalization

We will now discuss the performance achieved with or without the WCCN technique in the case of linear and cosine kernels. Table 3 compares the results obtained with and without WCCN to the results given by frame by frame joint factor analysis scoring.

The results given in Table 3 show that with WCCN, we achieve 17% relative improvements in both kernels. We can see also that the performances obtained with WCCN are very comparable to the JFA scoring. However the advantage of using this new SVM-JFA scoring is that it is more faster than the claasical JFA scoring in the context of NIST speaker recognition evaluation.

Table 4. Comparison results between score fusion and kernels combination for SVM-JFA system.

	NIST 2006 SRE		NIST 2008 SRE	
	English	All trials	English	All trials
Cosine kernel on y	2.34%	3.59%	3.86%	6.55%
Cosine kernel on z	6.26%	8.68%	10.34%	13.45%
Linear score fusion	2.11%	3.62%	3.23%	6.86%
Kernel combination	2.08%	3.62%	3.20%	6.60%
JFA frame by frame Scoring	1.80%	2.96%	-	-
JFA integrate over channel factors	1.78%	2.90%	-	-
JFA LPT assumption	2.70%	3.98%	-	-

5.3. SVM-JFA: speaker and common factors space

We present a comparison between results obtained with score fusion and kernel combination applied in the speaker and common factors. In both fusion techniques, we applied cosine kernel in speaker and common factors space. We used also WCCN in order to normalize the speaker factors cosine kernel. The results are given in Table 4.

By examining these results, we can conclude that both fusion methods yield equivalent results. However, the use of the kernel combination is more appropriate because we don't need development data to set the kernel weights. The results reported in Table 4 using score fusion on NIST 2006 SRE are not realistic because we trained and tested the score fusion weights on the same dataset.

We also note that the common factor components give complementary information to speaker factor components and the combination between them improves the performance. If we compare our results obtained by the kernel combination method and the other JFA scoring methods, we can seen that the SVM-JFA gives better results than the LPT assumption scoring defined in [12] and closer results to classical frame by frame JFA and integrating over channel factors [4] scorings.

6. CONCLUSION

In this paper, we tested several combinations between discriminative model which is Support vector machine and generative model which is Joint Factor analysis for speaker verification. We found that using linear or cosine kernel defined in speaker and Common factors which are the components of the JFA gave better results than using linear Kullback Leibler kernel applied in GMM supervectors obtained also with JFA model. We proved that using within class covariance normalization in speaker space in order to compensate for the channel effect gave the best performances. The results obtained with SVM-JFA using the speaker factors were comparable to the results obtained with classical JFA scoring. However the advantage of using the SVM in speaker factors space (usally dimension 300) makes the scoring faster than others classical techniques.

7. ACKNOWLEDGE

This work was carried out during the CLSP 2008 workshop at Johns Hopkins University, Baltimore, MD, USA. This work was funded in part by the Canadian Heritage Fund for New Media Research Networks.

8. REFERENCES

[1] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A Study of Inter-Speaker Variability in Speaker Verification," *IEEE Transactions on Audio, Speech and Language Processing*, July 2008.

[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. 10–41, 2000.

[3] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, "Svm Based Speaker Verification using a GMM SuperVector Kernel and NAP Variability Compensation," in *IEEE-ICASSP*, Toulouse, 2006.

[4] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Trans. Audio Speech and Language Processing*, May 2007.

[5] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker and Session Variability in GMM-Based Speaker Verification," *IEEE Trans. Audio Speech and Language Processing*, May 2007.

[6] 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, May 2006.

[7] A. Hatch, S. Kajarekar, and A. Stolcke, "Within-Class Covariance Normalization for SVM-Based Speaker Recognition," in *ICSLP*, 2006.

[8] Niko Brummer, Lukas Burget, Jan Honza Cernocky, Ondrej Glembek, Frantisek Grezl, Martin Karaat, David A. van Leeuwen, Pavel Matejka, Petr Schwarz, and Albert Strasheim, "Fusion of Heterogeneous Speaker Recognition Systems in the STBU Submission for the NIST Speaker Recognition Evaluation 2006," *IEEE Trans. On Audio, Speech and Language Processing*, September 2007.

[9] R. Dehak, N. Dehak, P. Kenny, and P. Dumouchel, "Kernel Combination for SVM Speaker Verification," in *Odyssey Speaker and Language Recognition Workshop 2008*, Stellenbosch, South Africa, 2008.

[10] `http://www.nist.gov/speech/tests/spk/index.htm`, ," .

[11] J. Pelecanos and S. Sridharan, "Feature Warping for Robust Speaker Verification," in *Speaker Odyssey*, Crete, Greece, 2001, pp. 213–218.

[12] C. Vair, D. Colibro, F. Castaldo, E. Dalmasso, and P. Laface, "Loquendo-Politecnico Di Torino's 2006 NIST Speaker Recognition Evaluation System," in *Interspeech 2007*, Antwarp, Belgium, 2007.