ISCA Archive
http://www.isca-speech.org/archive

Odyssey 2008: The Speaker and
Language Recognition Workshop
Stellenbosch, South Africa
January 21-24, 2008

Comparison Between Factor Analysis and GMM Support Vector Machines for Speaker Verification

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

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

Abstract

We present a comparison between speaker verification systems based on factor analysis modeling and support vector machines using GMM supervectors as features. All systems used the same acoustic features and they were trained and tested on the same data sets. We test two types of kernel (one linear, the other non-linear) for the GMM support vector machines. The results show that factor analysis using speaker factors gives the best results on the core condition of the NIST 2006 speaker recognition evaluation. The difference is particularly marked on the English language subset. Fusion of all systems gave an equal error rate of 4.2% (all trials) and 3.2% (English trials only),

1. Introduction

In speaker verification, the Gaussian mixture models and the support vector machines became the most widely used models. In the two last years, a combination of these two models was successfully applied by using a linear and non-linear GMM supervector kernels [6] [5]. A channel compensation algorithm named Nuisance Attribute Projection (NAP) was also proposed for this new model, this algorithm was tested on both types of kernels [6] [10]. Factor analysis model [1] [2] proposed a model which compensates for channel effect in the GMMs and a model based on the speaker factors to enroll the target speaker model.

In this paper, we carried out a comparison between two models: Factor analysis and GMM support vector machines (GMM-SVM's) with NAP using the same training and testing dataset. We also prove that zt-norm score normalization for the GMM-SVM systems did not give any improvement and t-norm alone was better for these systems. The best results are obtained with factor analysis using the speaker factor components. We show also the fusion results between the GMM-SVM systems with linear and non linear kernels and factors analysis model.

The outline of the paper is as follows. Section 2 describes the factor analysis model. In section 4.4, we present the GMM-SVM and we describe the linear and non linear kernels that we used to implement it. 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. Joint Factor Analysis

Joint factor analysis model is used to address the problem of speaker and session variability in GMM's. In this model, each speaker is represented by the means, covariance, 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 Universal Background Model parameters (UBM). The UBM is trained using a large amount of data. In Joint Factor Analysis [1] [2], the basic assumption is that a speaker and channel-dependent supervector[1] $\boldsymbol{M}$ can be decomposed into a sum of two supervectors: a speaker supervector $\boldsymbol{s}$ and a channel supervector $\boldsymbol{c}$

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

where $\boldsymbol{s}$ and $\boldsymbol{c}$ are normally distributed.

In [1], Kenny *et al.* described how the speaker dependent supervector 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 $\boldsymbol{s}$ is the speaker supervector for a randomly chosen speaker then

$$\boldsymbol{s} = \boldsymbol{m} + \boldsymbol{v}\boldsymbol{y} + \boldsymbol{d}\boldsymbol{z} \tag{2}$$

where $\boldsymbol{m}$ is the speaker and channel independent supervector from the UBM, $\boldsymbol{d}$ is diagonal matrix, $\boldsymbol{v}$ is a rectangular matrix of low rank and $\boldsymbol{y}$ and $\boldsymbol{z}$ are independent random vectors having standard normal distributions. In other words, $\boldsymbol{s}$ is assumed to be normally distributed with mean $\boldsymbol{m}$ and covariance matrix $\boldsymbol{v}\boldsymbol{v}^* + \boldsymbol{d}^2$. The components of $\boldsymbol{y}$ are the speaker factors. The channel-dependent supervector $\boldsymbol{c}$ which represents the channel effect in an utterance is assumed to be distributed according to

$$\boldsymbol{c} = \boldsymbol{u}\boldsymbol{x} \tag{3}$$

where $\boldsymbol{u}$ is a rectangular matrix of low rank, $\boldsymbol{x}$ is distributed with standard normal distribution. This is equivalent to saying that $\boldsymbol{c}$ is normally distributed with zero mean and covariance $\boldsymbol{u}\boldsymbol{u}^*$. The components of $\boldsymbol{x}$ are the channel factors in factor analysis modeling.

3. GMM-SVM's

This approach consists of the application of support vector machines with GMM supervectors as input features for the speaker verification task. We refer to the supervectors as *input* features because, in the case of a general kernel $K(s, s')$ (whose arguments are pairs of supervectors), it is necessary to distinguish between input features and expanded features defined by the kernel mapping function

$$s \rightarrow K(s, \cdot). \tag{4}$$

[1] A GMM supervector is the concatenation of the GMM mean vectors.

We will denote this mapping function by $\phi(s)$. Of course, in the case of a linear kernel defned by an inner product in the input feature space, the kernel mapping is just the identity mapping.

3.1. Linear Kernel

The linear kernel that we used on GMM supervector space is derived from the distance between two GMMs based on Kullback-Leibler (KL) divergence [8] [9]. In the case of MAP adaptation with diagonal covariance matrices and when only the means of GMMs were adapted from the UBM, the weighted Euclidean distance between scaled version of GMM supervectors $\boldsymbol{s}$ and $\boldsymbol{s'}$ was given as follow:

$$\mathcal{D}_e^2\left(\boldsymbol{s}, \boldsymbol{s'}\right) = \sum_{i=1}^{C} w_i \left(s_i - s_i'\right) \Sigma_i^{-1} \left(s_i - s_i'\right)^t \tag{5}$$

where w_i and Σ_i are the i^{th} UBM mixture weight and diagonal covariance matrix, s_i corresponds to the mean of the $i^t h$ Gaussian of the speaker GMM.

The linear kernel is defined as the corresponding inner product:

$$K_{lin}(\boldsymbol{s}, \boldsymbol{s'}) = \sum_{i=1}^{C} \left(\sqrt{w_i}\Sigma_i^{-\frac{1}{2}} s_i\right)\left(\sqrt{w_i}\Sigma_i^{-\frac{1}{2}} s_i'\right)^t \tag{6}$$

This kernel was proposed by Campbell *et. al.* [6].

3.2. Non Linear Kernel

The non linear kernel that we used is the exponential version of the distance between two GMMs $\mathcal{D}_e^2\left(\boldsymbol{s}, \boldsymbol{s'}\right)$ given in (5):

$$K_{nonlin}(\boldsymbol{s}, \boldsymbol{s'}) = e^{-\mathcal{D}_e^2\left(\boldsymbol{s}, \boldsymbol{s'}\right)} \tag{7}$$

This kernel was first proposed by Dehak and Chollet in [5]. The non linear kernel is equivalent to the Gaussian kernel defined on the GMMs supervector space. The corresponding expanded feature space is infinite-dimensional. The feature mapping function $\phi(.)$ is [11]:

$$\boldsymbol{s} \mapsto \phi(\boldsymbol{s}) = K(\boldsymbol{s}, \, .\,) = e^{-\frac{\|\boldsymbol{s} - \,.\,\|^2}{2\sigma^2}} \tag{8}$$

3.3. Input feature normalization : M-norm

When GMM supervectors are used as input features for a kernel machine, care has to be taken to normalize them properly. In [5] [10] the authors demonstrated the effectiveness of the model normalization (M-Norm) technique, especially for the non-linear kernel SVM. This normalization consist of modifying the GMM supervectors so that the distance between all normalized supervectors and the UBM supervector $\boldsymbol{m}$ is a constant that we can take to be 1. Let $\{m_k\}$ be the set of UBM mean vectors and, for a given speaker X, let $\{s_k\}$ be the set of mean vectors in the speaker GMM. Denote by $\mathcal{D}_e(X, \boldsymbol{m})$ the distance between the speaker GMM supervector and the UBM supervector. For a particular mean vector s_k, the normalization procedure is

$$s_k \quad \leftarrow \quad \frac{1}{\mathcal{D}_e\left(X, \boldsymbol{m}\right)} s_k \; + \; \left(1 - \frac{1}{\mathcal{D}_e\left(X, \boldsymbol{m}\right)}\right) m_k \tag{9}$$

3.4. Nuisance Attribute Projection

In [7][6], the authors proposed the Nuisance Attribute Projection (NAP) method to treat the session variability problem in the SVM framework. This method used an appropriate projection matrix P in the input feature space to remove unwanted variability (such as channel effects) in the input features. The new kernel obtained has the following form:

$$\begin{aligned} K(\boldsymbol{s}, \boldsymbol{s'}) &= \; < P\phi(\boldsymbol{s}) \;,\; P\phi(\boldsymbol{s'}) > \\ &= \; \phi(\boldsymbol{s})^t P \phi(\boldsymbol{s'}) \\ &= \; \phi(\boldsymbol{s})^t (I - VV^t) \phi(\boldsymbol{s'}). \end{aligned} \tag{10}$$

where V is a rectangular matrix of low rank whose columns are orthonormal. 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 input feature space.

The design criterion for P and the corresponding matrix V is

$$\tilde{P} = \arg\min_{P} \sum_{i,j} W_{i,j} \|P\phi(\boldsymbol{s_i}) - P\phi(\boldsymbol{s_j})\|^2 \tag{11}$$

where $\phi(.)$ is the kernel mapping function. The $\{\boldsymbol{s_i}\}$ are typically a background data set. The $W_{i,j}$ matrix contains the speaker weights. We pick $W_{i,j} = 1$ if $\boldsymbol{s_i}$ and $\boldsymbol{s_j}$ correspond to the same speaker, and $W_{i,j} = 0$ otherwise.

Campbell *et. al.* [6] noticed that with the linear kernel and session variability as nuisance variable, the NAP subspace is equivalent to the channel subspace modeled in the factor analysis [2]. 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. The solution of Equation (11) ($\tilde{P}$ and corresponding $\tilde{V}$) is obtained from the k eigenvectors having the k largest eigenvalues of the following covariance matrix:

$$\mathcal{C} = \frac{1}{S} \sum_{j=1}^{S} \frac{1}{n_j} \sum_{i=1}^{n_j} \left(\tilde{\mathbf{s}}_{\mathbf{i}}^{\mathbf{j}} - \bar{s}_j\right)\left(\tilde{\mathbf{s}}_{\mathbf{i}}^{\mathbf{j}} - \bar{s}_j\right)^t \tag{12}$$

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

$$\bar{s}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} \tilde{\mathbf{s}}_{\mathbf{i}}^{\mathbf{n_j}} \tag{13}$$

The channel covariance matrix $\mathcal{C}$ is equivalent to the matrix $\boldsymbol{uu}^*$ which is the covariance matrix of the channel supervector $\boldsymbol{c}$ in the factor analysis model. In [10], we applied the same projection matrix $\tilde{P}$ in GMM supervector space for both the linear and non-linear kernels. Our best results were obtained when we applied M-norm on the GMM supervectors after applying the projection $\tilde{P}$.

4. Experimental set up

4.1. Test set

All our experiments are carried out on the core condition of the NIST 2006 speaker recognition evaluation (SRE) dataset [4]. This condition evaluation set contains 350 males, 461 females, and 51,448 test utterances. For each target speaker, a five minute recording is available containing roughly two minutes of speech.

4.2. Acoustic features

In our experiments, we used cepstral features, extracted using a 25 ms Hamming window. 19 mel frequency cepstral coefficients together with log energy are calculated every 10 ms. This 20-dimensional feature vector was subjected to feature warping [3] using a 3 s sliding window. Delta coefficients were then calculated using a 5-frame window giving a 40-dimensional feature vector. These feature vectors were modeled using GMMs and factor analysis was used to address the problem of speaker and session variability.

4.3. Factor analysis training

We used two gender dependent Universal Background Models (UBM). Each UBM containing 1024 Gaussians was trained using LDC releases of Switchboard II, Phases 2 and 3; Switchboard Cellular, Parts 1 and 2; the Fisher English Corpus, Part 1 and Part2; NIST 2003 Language recognition evaluation dataset; and NIST 2004 evaluation data.

The (gender dependent) factor analysis models were trained on the LDC releases of Switchboard II, Phases 2 and 3; Switchboard Cellular, Parts 1 and 2; and the NIST 2004 evaluation data.

The decision scores obtained with factor analysis were normalized using zt-norm. We used 283 t-norm models for female trials and 227 t-norm models for male trials taken from Switchboard II, Phases 2 and 3; Switchboard Cellular, Parts 1 and 2; and NIST 2004 SRE. We used 1000 z-norm utterances for each gender taken from the same dataset as t-norm. The motivation of using large number of z-norm utterances was given in the companion paper [12].

We tested two factor analysis configurations both having 50 channel factors. In the first case, we did not use any speaker factors and in the second case we used 300 speaker factors. The latter configuration gave the best results on the core condition of the NIST 2006 SRE dataset [12].

4.4. UBM and NAP training

In GMM-SVM systems, we used two gender dependent UBMs composed of 1024 Gaussians. These UBMs were trained on the same dataset as used for factor analysis UBM training (Section 4.3).

The channel covariance matrix $\mathcal{C}$ given in Equation (12) was computed using the same dataset as for factor analysis. The first 40 eigenvectors of this matrix were used to build the NAP projection matrix. (This gave the best results in [10].)

4.5. GMM-SVM imposters

Imposters play two different roles in an SVM system: they are needed to train the SVM for each target speaker and they can be used to normalize the scores of verification trials which facilitates setting a verification decision threshold.

We experimented with two types of imposter modeling and score normalization for each SVM system. For the first experiment, we divided the 1000 z-norm utterances used in the factor analysis system into two equal parts. 500 impostors were used to train the SVM for each target speaker and the remaining 500 utterances were employed for z-norm score normalization. We used the same t-norm impostors as in factor analysis to carry out t-norm and zt-norm. It turned out that (contrary to our experience with factor analysis [12]), only t norm was effective.

So, in a second experiment, we used all 1000 z-norm utterances as impostors to train target speaker SVM's. We used the same t-norm imposters as in factor analysis. We did not test the zt-norm score normalization in this case.

5. Results

Score normalization is vitally important to the success of our factor analysis system [12] so our first experiments with the GMM-SVM systems were concerned with this question. We present these results before presenting the back to back comparison between the GMM-SVM and factor analysis systems.

5.1. GMM-SVM score normalization

The results presented in Tables 1 and 2 show that zt-norm does not improve the performance of the SVM systems compared with t-norm and the z-norm utterances are better used as imposters for training target speaker SVM's than for score normalization. Note that non linear kernel gives better EER than linear kernel for both experiments. However the DCF is quite similar for both kernels in each case. We obtained a similar result in [10].

Table 1: *GMM-SVM supervector score normalization results on English trials of the core condition of the NIST 2006 SRE.*

	linear kernel		non linear	
	EER	DCF	EER	DCF
500 imposters with zt-norm	5.0%	0.025	4.6%	0.024
1000 imposters with t-norm	5.0%	0.025	4.4%	0.024

Table 2: *GMM-SVM supervector score normalization results on all trials of the core condition of the NIST 2006 SRE.*

	linear kernel		non linear	
	EER	DCF	EER	DCF
500 imposters with zt-norm	5.8%	0.030	5.7%	0.030
1000 imposters with t-norm	5.7%	0.030	5.5%	0.030

5.2. Comparison of GMM-SVM and factor analysis

We compare in this section the best results obtained on the core condition of the NIST 2006 SRE using both linear and non linear kernels (as described in the previous section) with the best results obtained using factor analysis with and without speaker factors (as described in the companion paper [12]). The results are summarized in Tables 3 and 4. The Figures 1 and 2 show the DET curves of all four systems .

Table 3: *Results on the English language trials of the core condition of the NIST 2006 SRE.*

	EER	DCF
linear kernel	5.0%	0.025
non linear kernel	4.4%	0.024
factor analysis with 0 speaker factors	4.6%	0.022
factor analysis with 300 speaker factors	**3.5%**	**0.021**

These results show that the two factor analysis configurations gave better results than the linear kernel in the English trials of the NIST 2006 SRE (Table3). However the linear ker-

Table 4: *Results on all trials of the core condition of the NIST 2006 SRE.*

	EER	DCF
linear kernel	5.7%	0.030
non linear kernel	5.5%	0.030
0 speaker factors	6.2%	0.028
300 speaker factors	**5.0%**	**0.027**

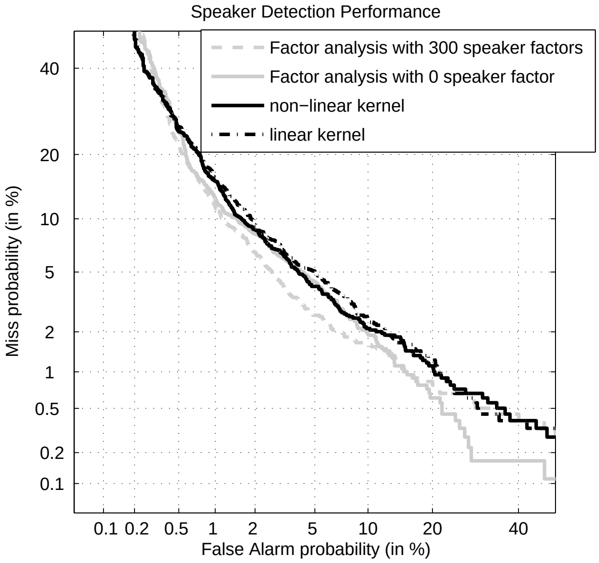

Figure 1: *DET curves showing the comparison results on English trials of the core condition of the NIST 2006 SRE.*

nel provided better EER than factor analysis without speaker factors on all trials of the core condition.

The non-linear kernel produced a better EER than factor analysis without speaker factors. In the absence of speaker factors, the procedure for enrolling a target speaker with a factor analysis model is similar to traditional MAP adaptation which is the first step in enrolling a target speaker in a GMM-SVM system. However the results obtained with 300 speaker factors are clearly better than those obtained with the other systems especially for the English language trials (see Fig. 1). More results on the effectiveness of speaker factors can be found in [12].

5.3. Fusion

It is widely recognized in this field that fusing systems leads to better results. Although it is suboptimal, naive Bayes fusion (where all systems are given equal weight) has the merit that it does not require any development data. Fusion using logistic regression fusion is also easy to implement [13] but, strictly speaking, held-out development data ought to be used to estimate the fusion weights.

We fused the two GMM-SVM systems and the factor analysis system with 300 speaker factors using both naive Bayes and logistic regression where the regression coefficients were estimated from the test data (with the help of the answer key). Thus we obtained upper and lower bounds on the performance improvements that can be expected from fusion. The results are summarized in Tables 5 and 6.

Naive Bayes fusion gave a slight improvement in EER and DCF on all trials of the core condition, but no improvement on the English language trials (presumably because the contribution of the factor analysis system was under-weighted).

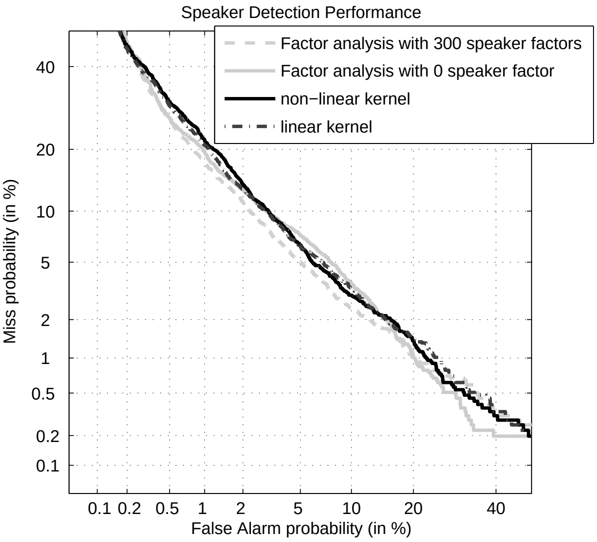

Figure 2: *DET curves showing the comparison results on all trials of the core condition of the NIST 2006 SRE.*

Table 5: *Fusion results on all trials of the core condition of the NIST 2006 SRE.*

	EER	DCF
naive Bayes	4.7%	0.026
logistic regression	4.2%	0.024

Logistic regression fusion on all trials gave an absolute improvement of 0.8% in EER and 0.003 in DCF in comparison to factor analysis with 300 speaker factors. However the improvement on the English language subset was smaller (presumably for the same reason).

6. Conclusions

This paper presents a comparison between two approaches to speaker verification, factor analysis and GMM support vector machines with linear and non linear kernels. We trained and tested these models on the same data sets using the same acoustic features. The results show that factor analysis without speaker factors gives similar results to the GMM-SVM systems with the non-linear kernel and that the non linear kernel outperforms the linear kernel.

However when speaker factors are used, factor analysis produced substantially improved results especially in the English language trials of the core condition of the NIST 2006 SRE. The key differnence here is in the way target speakers are enrolled: in the absence of speaker factors, the enrollment procedure is similar to classical MAP which is the first step in enrolling a speaker for a GMM-SVM system. This suggests that the GMM-SVM systems could be improved by using speaker factors at enrollment time.

We also found that in the GMM-SVM systems t-norm is more appropriate than zt-norm. Why zt-norm is so effective with factor analysis remains something of a mystery.

Table 6: *Fusion results on English trials of the core condition of the NIST 2006 SRE.*

	EER	DCF
naive Bayes	3.7%	0.021
logistic regression	3.2%	0.019

Naive Bayes fusion of the factor analysis and GMM-SVM systems gave small improvements in the performance; posterior fusion using logisitic regression gave larger improvements. But, for English language trials, the contribution of the factor analysis system was predominant.

7. References

[1] P. Kenny, G. Boulianne, P. Ouellet and P. Dumouchel, Joint Factor Analysis versus Eigenchannels in Speaker Recognition, IEEE Trans. Audio Speech and Language Processing, Vol 15,4, May, 2007.

[2] 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, Vol 15,4, May, 2007.

[3] J. Pelecanos and S. Sridharan, Feature Warping for Robust Speaker Verification, Proc. Speaker Odyssey, Crete, Greece, pp 213-218, jun 2001.

[4] http://www.nist.gov/speech/tests/spk/index.htm.

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

[6] 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.

[7] A. Solomonoff, W. Campbell, and I. Boardman, Advances in Channel Compensation for SVM Speaker Recognition, in ICASSP, vol. 1, 2005, pp. 629-632.

[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] 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.

[10] R. Dehak, N. Dehak, P. Kenny, P. Dumouchel, Linear and Non Linear Kernel GMM SuperVector Machines for Speaker Verification, in Interspeech 2007, Antwerp, Belgium, August 27-31, 2007.

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

[12] P. Kenny, N. Dehak, R. Dehak, V. Gupta and P. Dumouchel, The Role of Speaker Factor in the NIST Extended Data Task, Submited to IEEE Odyssey 2008.

[13] N. Brummer, L. Burget, J. Honza Cernocky, O. Glembek, F. Grezl, M. Karafiat, D. A. van Leeuwen, P. Matejka, P. Schwarz, A. Strasheim, Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006, To appear in IEEE transaction on audio, speech and language processing, September 2007.