Odyssey 2014:
The Speaker and Language Recognition Workshop
16-19 June 2014, Joensuu, Finland

GMM Weights Adaptation Based on Subspace Approaches for Speaker Verification

Najim Dehak[1], Oldrich Plchot[2], Mohamad Hasan Bahari[3],
Lukas Burget[2], Hugo Van hamme[3] and Réda Dehak[4]

[1] MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, USA
[2] Brno University of Technology, Speech@FIT, Brno, Czech Republic
[3] KU Leuven, Department of Electrical Engineering, Leuven, Belgium
[4] Laboratoire de Recherche et de Développement de l'EPITA (LRDE), Paris, France

Abstract

In this paper, we explored the use of Gaussian Mixture Model (GMM) weights adaptation for speaker verification. We compared two different subspace weight adaptation approaches: Subspace Multinomial Model (SMM) and Non-Negative factor Analysis (NFA). Both techniques achieved similar results and seemed to outperform the retraining maximum likelihood (ML) weight adaptation. However, the training process for the NFA approach is substantially faster than the SMM technique. The i-vector fusion between each weight adaptation approach and the classical i-vector yielded slight improvements on the telephone part of the NIST 2010 Speaker Recognition Evaluation dataset.

1. Introduction

The i-vector approach has been proven to be a powerful speech representation for audio classification problems. It was first introduced for speaker verification [1] and was then applied successfully in several other applications such as language recognition [2] and speaker diarization [3]. This technique was proposed in the context of the Gaussian Mixture Model (GMM) framework in order to model all the variability between several GMMs corresponding to different speech recordings into a low dimensional representation space.

Since the early work on speaker recognition based on the GMM-Universal Background Model (UBM) approach and until the i-vector representation was introduced, it was shown that only adapting the means is largely enough for speaker recognition [4]. However, in other application such as language identification, updating both the GMM means and variances seems to be helpful [5]. In more recent studies, adapting the weight seems to provide some complementarity information for many applications such as speech recognition [6] and age estimation [7, 6].

In this paper, we applied two different subspace adaptation techniques to update the GMM weights for speaker recognition, the first approach named Subspace Multinomial Model was first introduced to model prosodic features [8] and later it was successfully applied to phonotactic systems for the language identification task [9]. The second approach, named Non-Negative factor Analysis, is a variant of a factor analysis modeling [10]. Both subspace approaches were recently compared and applied for adapting the GMM weights for language identification [10] [11].

The remainder of the paper is organized as follows. Section 2 describes the i-vector approach. We present both GMM weights adaptation techniques in Section 3. Section 4 presents the experimental setup and the results. Section 5 includes conclusions and avenues for future work.

2. The i-vector framework

The i-vector approach [1] is a very powerful technique that summarizes all the updates happening during the adaptation of the UBM mean components (named also GMM supervector) to a given utterance sequence of frames. All this information is modeled in a low dimensional space named the total variability space. In the i-vector framework, each speech utterance has a corresponding GMM supervector that is assumed to be generated as follows:

$$M = m + Tw \tag{1}$$

where m is the speaker independent and channel independent supervector (which can be taken to be the UBM

This work was supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. D10PC20015, by Technology Agency of the Czech Republic grant No. TA01011328 and by IT4Innovations Centre of Excellence CZ.1.05/1.1.00/02.0070. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

Mohamad Hasan Bahari was supported by the European Commission as a Marie-Curie ITN-project, namely Bayesian Biometrics for Forensics, under Grant Agreement number 238803 and the Fonds Wetenschappelijk Onderzoek (FWO) as a travel grant for a long stay abroad.

supervector), T is a rectangular matrix of low rank, and w is a random vector having a prior standard normal distribution $\mathcal{N}(0, \mathrm{I})$. The i-vectors are Maximum A Posteriori (MAP) point estimates of the latent variable w adapting the corresponding GMM (supervector m) to a given utterance. The process of extracting the i-vector serves as a features extraction step and i-vectors extracted from individual speech segments are used as an input to a classifier to obtain the final speaker verification scores.

3. Weight adaptation approaches

Applying subspace approaches for weight adaptation is different from means adaptation because of the constraints imposed to the weights. For each GMM the weights should be always positive and sum to one. These two constraints make weight adaptation harder. Several approaches have been proposed such as Non-negative Matrix Factorization [6], Subspace Multinomial Model [8] and more recently a factor analysis approach named Non-negative factor Analysis (NFA) [10]. The auxiliary function of data for the GMM weight adaptation is given as follows.

$$\Omega\left(\omega, \widehat{\omega}\right) = \sum_{i=1}^{N} \sum_{c=1}^{C} \gamma_{c,i} \log \omega_{c,i} \tag{2}$$

where $\gamma_{c,i}$ is the occupation count for Gaussian c and segment i. N denotes the total number of observations and $\omega_{c,i}$ are the probabilities of multinomial distribution for a segment i. In the next two subsections, we will present two different subspace approaches for GMM weight adaptation allowing us to adapt the weights in a ML sense to a specific segment.

3.1. Subspace Multinomial Model

The SMM approach assumes that the parameters of the corresponding multinomial distributions ω_c for a given speech utterance obtained in equation 2 can be represented as

$$\omega_c = \frac{e^{(\mu_c + L_c r)}}{\sum_{j=1}^{C} e^{(\mu_j + L_j r)}} \tag{3}$$

where μ_c is the c^{th} element of the origin of the supervector subspace, L_c is the c^{th} row of the subspace matrix and r is a low dimensional vector representing speaker and channel.

In this method, L and r are estimated using an Expectation Maximization (EM) algorithm. In each E- and M-step, an iterative optimization approach similar to the Newton-Raphson paradigm is applied to maximize the objective function (2). Details of parameter re-estimation can be found in [8]. In our experiments, we used the new r vector as a representation for a speaker verification system. After the model is trained, it can be used to extract ML estimates of the low-dimensional vector r for any segment and use it just like i-vectors (i.e. as features for the following classifier).

3.2. Non-Negative factor Analysis

The basic assumption of this framework is that the c^{th} Gaussian weight of the adapted GMM can be decomposed as follows:

$$\omega_c = b_c + L_c r \tag{4}$$

where b_c is the UBM weight of the corresponding component. L_c denotes the c^{th} row of the matrix L, which is a matrix of dimension $C \times \rho$ spanning a low-dimensional subspace. r is a low dimensional vector (rho dimension) that best describes the utterance-dependent weight offset Lr. This r vector will be used as input to a speaker verification system similar to the i-vector approach. In [10], we also find that imposing a standard normal distribution prior on the r vector similar to the classical i-vector did not help to improve the performances for language identification. The subspace matrix L is estimated via factor analysis to represent the directions that best model different speech recordings in a large training data set.

In this method, L_c and r are estimated using an EM algorithm. Each E-step and M-step of the EM algorithm, a gradient-ascend optimization scheme is applied to maximize the auxiliary function of Equation (2). The training process consists of optimizing the following problem:

$$\begin{aligned} \max \quad & \Omega\left(\omega, \widehat{\omega}\right) \\ \text{Subject to} \quad & g\left(b + Lr\right) = 1 \\ & b + Lr > 0 \end{aligned} \tag{5}$$

where g is a row vector of ones. Details of parameter re-estimation can be found in [10].

3.3. Comparison between SMM and NFA

In NFA, the adapted weights are assumed to be the UBM weights b offset by Lr, maximizing the likelihood of the data. The SMM replaces this simple and linear relation by the non-linear relation of (3), where the offset $\exp\left(Tw\right)$ is multiplied to the UBM weights b and the result of multiplication is normalized so that the adapted weights sum up to one.

Figures 1, 2 and 3 demonstrates the GMM weights modeling using the subspace techniques SMM and NFA. Figures 1 shows weights of a UBM with 3 Gaussian components, which is ML adapted to different speech segments. In this figure, each dot represents weights for one segment. Since the adapted weights are constrained to be positive and sum up to one, they are constrained to live on a two-dimensional simplex. Note that the weights shown in the figure are derived from the occupation counts $\gamma_{c,i}$, which were artificially designed in such a way that SMM can fit this data well.

This is demonstrated on Figure 2 showing one dimensional manifold (the red curve) learned using SMM from the same occupation counts. Once the SMM subspace L is learned, different vectors r (in this simple case one dimensional) corresponds to different multinomial distributions (GMM weights), which are in our example constrained to live on the red curve. Our task is to derive such vector r for each speech segment that produces GMM weights fitting the occupation counts the best in the ML sense.

Figure 3 demonstrates the subspace learned using NFA. In this case, the GMM weights adapted to the individual speech segments are constrained to live in a linear subspace (red line in our example) of the simplex. Since the data in our example were handcrafted specifically for SMM, NFA cannot fit the data that well (red line does not approximate well the blue dots in the simplex corners). We believe that this simple example reflect well what is happening in the real high dimensional cases. We have observed that SMM generally lead to higher improvements of the objective function (2) compared to NFA. This was caused by the inability of NFA to represent multinominal distributions with many small (close to zero) probabilities. On the other hand, avoiding very small probabilities can be seen as a natural smoothing property of NFA, which might help to avoid over-fitting. It is known that SMM can suffer from over-fitting that has to be addressed by adding a regularization term as in [9] . However, the regularization parameter requires fine-tuning over a development data set. Our experiments show that using either NFA or SMM results in a comparable speaker recognition performance.

The procedure of updating subspace matrix and subspace vectors is also different between the SMM and the NFA frameworks. In our implementation, NFA uses a simple and very fast gradient ascend technique to estimate the subspace matrix and the subspace vectors. The gradient descent optimization was not found effective for SMM. An optimization resembling Newton-Raphson technique is applied in the SMM case [9], which requires to calculate costly approximations to the Hessian matrix making the optimization significantly slower compared to NFA.

4. Experiments and Results for Speaker Recognition

4.1. Experiment Setup

Our experiments operate on cepstral features, extracted using a 25ms Hamming window. 19 mel frequency cepstral coefficients together with log energy are calculated every 10ms. Delta and double delta coefficients were then calculated using a 5 frame window to produce 60-dimensional feature vectors. This 60-dimensional feature vector was subjected to feature warping using a 3s sliding window. We used a gender independent UBM containing 2048 Gaussians. It was trained on Switchboard data and NIST 2004,2005,2006 and 2008 SRE.The extractor for classical 600-dimensional i-vectors was trained on the same data as the one used for training the UBM. The obtained i-vectors were projected first by Linear Discriminant Analysis (LDA). Probabilistic LDA [12, 13, 14] is then used to compute the final verification score. The results are reported on both genders of the core and eight conversations conditions of the telephone part of the NIST 2010 SRE dataset.

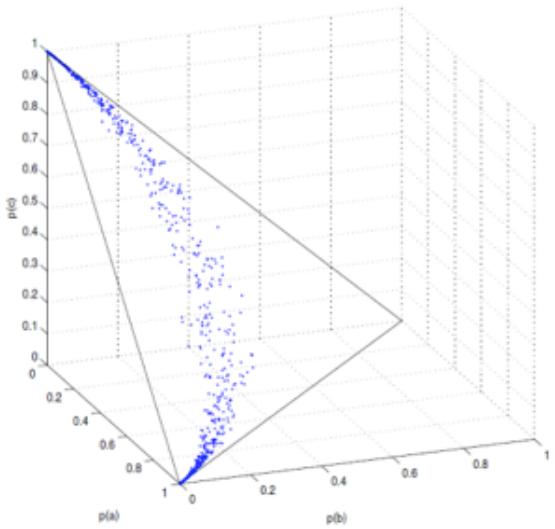

Figure 1: Adapted weights using the ML approach.

4.2. Varying the dimensionality of weight subspace approaches

The first set of experiments that we carried out is to verify the effect of different subspace dimension on both NFA and SMM for speaker verification task. Figure 4 and 5 show the performance obtained with dimension 500,1000 and 1500. These results are only with the best LDA dimensionality reduction for each dimensionality space. From both figures, we notice than varying the dimensionality of the subspace does not affect the results obtained by both weights adaptation approaches.

4.3. Comparison results

A comparison of results between three different techniques of weights adaptations for the speaker verification task is shown in Figure 6. The results of both subspace techniques outperform the performance achieved by the classical maximum likelihood estimation. However the results obtained by the NFA approach are very close to the ones obtained with SMM technique. The results are given on Equal Error rate (EER), old (SMM08, NFA08,

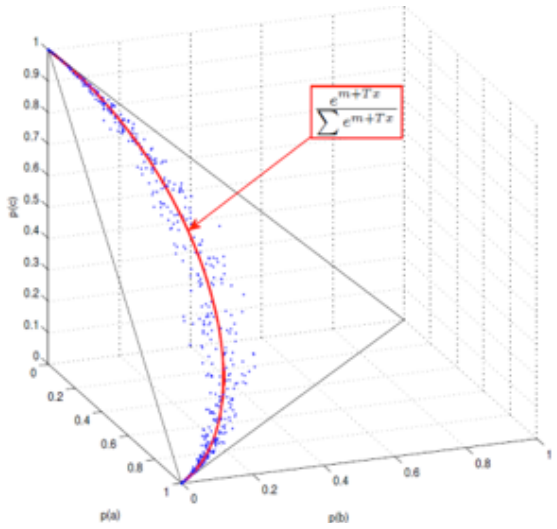

Figure 2: Adapted weights using the SMM approach.

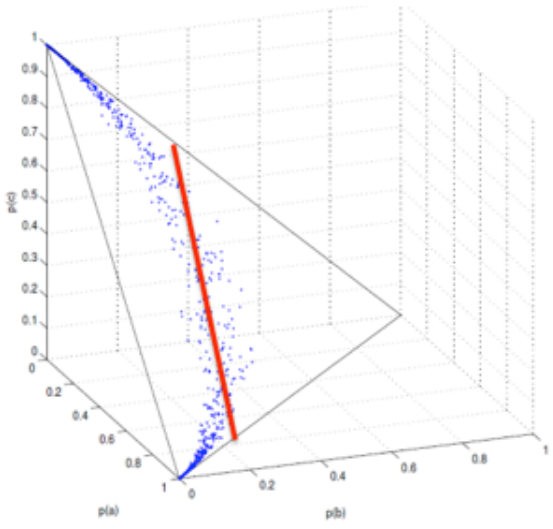

Figure 3: Adapted weights using the NFA approach.

...) and new minDCF (SMM10, NFA10, ...) operating point for both gender of the core and 8 conversations conditions. The new minDCF is the cost function adopted by NIST-SRE since 2010, the old minDCF is the previous one [15]. Another remark from Figure 6 is that the performances for both subspace approaches on 8 conversations task are very promising with EER less than 2%. Although the SMM and NFA achieved very comparable results, the NFA approach however has the advantage to have a training process that is very fast compared to the SMM technique.

Figure 7 shows a comparison between the i-vector system and both GMM weight subspaces adaptation methods. We can notice that both weight adaptation approaches achieved a comparable results with I-vector method especially for 8 conversation task.

4.4. I-vector fusion Results

In this section, we present the fusion of the classical i-vector system and each of the two subspace approaches. We carried out the fusion on the i-vector level because the score fusion did not achieve any improvements. The obtained results are reported in Table 1 and 2.

Table 1: Fusion results on Female part core condition of the NIST 2010 SRE

Cond 5 Female	New minDCF	Old minDCF	EER (%)
Baseline	0.4467	0.1269	2.42
+ NFA	0.4251	0.1227	2.50
+ SMM	**0.4197**	**0.1218**	**2.24**

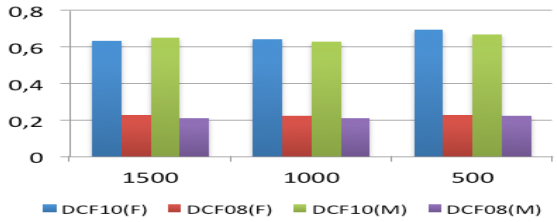

Figure 4: Results obtained with different NFA subspace dimensions on the telephone part of the core condition of the NIST 2010 SRE for both genders (M,F).

We notice from both Tables that SMM gives nice improvement in the new minDCF for Female trials when fused with the i-vector. However the NFA approach obtained better results when combined with the classical i-vector for Male.

4.5. Varying the number of UBM Gaussian components

The complexity of GMM means adaptation is related to both the number of Gaussians in the mixture and the dimensionality of the feature frames. This complexity can

Table 2: Fusion results on Male part core condition of the NIST 2010 SRE

Cond 5 Male	New minDCF	Old minDCF	EER (%)
Baseline	**0.3869**	0.1114	2.20
+ NFA	0.3896	**0.1023**	**2.07**
+ SMM	0.3890	0.1033	2.16

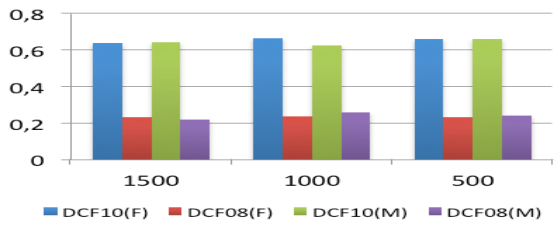

Figure 5: Results obtained with different SMM subspace dimensions on the telephone part of the core condition of the NIST 2010 SRE for both genders (M,F)

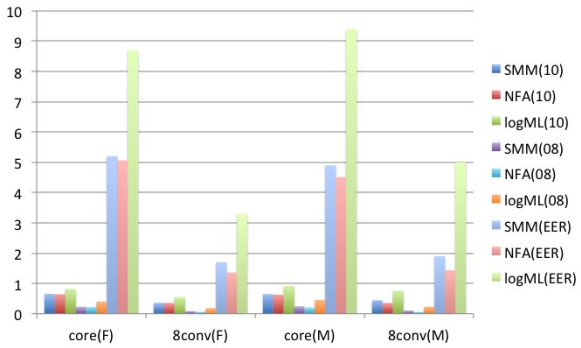

Figure 6: Comparison results obtained for both genders (M, F) between three different approaches for the GMM weights adaptation.

increase dramatically if the number of components is increased a lot and therefore the majority of GMM systems for speaker verification does not exceed 2048 Gaussians. However the GMM weights adaptation is only related to the number of Gaussian in the mixture. For this reason, we can experiment using a UBM of size 4096 Gaussians. In this section, we varied the number of Gaussians from 1024 to 4096 components. These UBMs were tried with NFA approach because as we stated earlier it is much faster to train NFA than SMM.

Table 3: NFA results with different UBM sizes. Results obtained on Female part for telephone data core condition of the NIST 2010 SRE

Cond 5 Female	New minDCF	Old minDCF	EER (%)
1024 Gaussians	0.7169	0.2465	5.29
2048 Gaussians	0.6320	0.2287	5.24
4096 Gaussians	**0.6021**	**0.2120**	**5.07**

The results reported in Table 3 and 4 show a nice improvement when increasing the number of UBM Gaussians to 4096 compared to a smaller number.

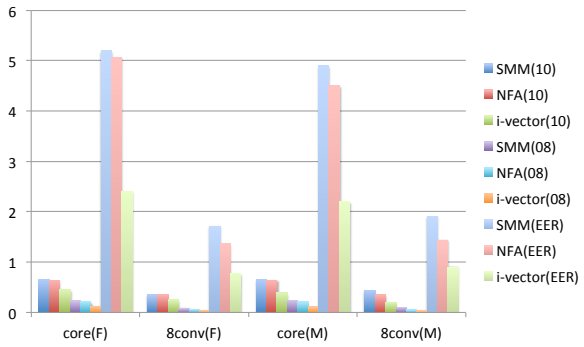

Figure 7: Comparison results obtained for both genders (M, F) between SMM, NFA and IVector approaches.

Table 4: NFA results with different UBM sizes. Results obtained on Male part for telephone data core condition of the NIST 2010 SRE

Cond 5 Male	New minDCF	Old minDCF	EER (%)
1024 Gaussians	0.6806	0.2350	4.704
2048 Gaussians	0.6493	0.2112	4.44
4096 Gaussians	**0.550**	**0.1818**	**3.81**

Table 5: Fusion results between a classical i-vector and NFA approach with different UBM sizes. Experiments carried out on Female part core condition of the NIST 2010 SRE

Cond 5 Female	New minDCF	Old minDCF	EER (%)
Baseline	0.4467	0.1269	2.42
+ NFA (4096)	**0.4375**	0.1325	2.65
+ NFA (1024)	0.4592	**0.1255**	**2.37**

Table 6: Fusion results between a classical i-vector and NFA approach with different UBM sizes. Experiments carried out on Male part core condition of the NIST 2010 SRE

Cond 5 Male	New minDCF	Old minDCF	EER (%)
Baseline	0.3869	0.1114	2.20
+ NFA (4096)	0.4090	0.1109	2.01
+ NFA (1024)	**0.3856**	**0.1031**	**1.98**

From both tables 5 and 6, we can notice that varying the number of Gaussians in the UBM improve in some cases the i-vector baseline with a UBM of size 2048. If we look to the new operating point (newDCF), 4096 Gaussians obtained the best improvement for Female and 1024 Gaussians for Male. However if we compared those

results with the one obtained in Tables 1 and 2, we can see that there is no improvement by varying the size of the UBM. It seems that using the same number of Gaussians in the UBM for both i-vector and weight adaptation is enough to carry information fusion between the GMM mean and weight components.

5. Conclusions

In this paper, we experimented with weights adaptation for speaker recognition. Two subspace techniques were compared. Both approaches obtained similar performances on NIST 2010 SRE. The performances of both GMM weight subspaces are not as good as the classical i-vector which operate on the GMM means. However, we also show a slight improvements by combining both the i-vector approach and each weights adaption approach. Similar conclusions were also found in [16][17]. As future work we would like to find a better way to combine both GMM weights and means information.

6. References

[1] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transaction on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788—798, 2011.

[2] N. Dehak, P. Torres-Carrasquillo, D. Reynolds, and R. Dehak, "Language recognition via i-vectors and dimensionality reduction," in *INTERSPEECH*, Florence, Italy, Aug 2011, pp. 857–860.

[3] S. Shum, N. Dehak, R. Dehak, and J. Glass, "Unsupervised methods for speaker diarization: An integrated and iterative approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2015–2028, Oct 2013.

[4] D. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[5] W. M. Campbell, "A covariance kernel for svm language recognition," in *ICASSP*, Las Vegas, NV, 2008, pp. 4141–4144.

[6] X. Zhang, K. Demuynck, and H. Van hamme, "Rapid speaker adaptation in latent speaker space with non-negative matrix factorization," *Speech Communication*, vol. 55, no. 9, pp. 893—908, oct 213.

[7] M. Li, K. J. Han, and S. Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," *Computer Speech and Language*, vol. 27, no. 1, pp. 151—167, 2013.

[8] M. Kockmann, L. Burget, O. Glembek, L. Ferrer, and J. Cernocky, "Prosodic speaker verification using subspace multinomial models with intersession compensation," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[9] M. M. Soufifar, L. Burget, O. Plchot, S. Cumani, and J. Cernocky, "Regularized subspace n-gram model for phonotactic ivector extraction," in *Proceedings of Interspeech*, Lyon, 2013, pp. 74–78.

[10] M. H. Bahari, N. Dehak, and H. Van hamme, "Gaussian mixture model weight supervector decomposition and adaptation," Tech. Rep., june 2013.

[11] M. H. Bahari, N. Dehak, H. Van hamme, L. Burget, and A.M. Ali, "Non-negative factor analysis for gaussian mixture model weights adaptation," *submitted IEEE/ACM Transaction on Audio, Speech and Language Processing*, 2013.

[12] D. Garcia-Romero and C. Y. Espy-Wilso, "Analysis of i-vector length normalization in speaker recognition systems," in *INTERSPEECH*, Florence, Italy, Aug 2011.

[13] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Odyssey Speaker and Language Recognition Workshop*, Brno, Czech Republic, June 2010.

[14] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *11th International Conference on Computer Vision*, Rio de Janeiro, Brazil, Oct 2007, pp. 1–8.

[15] "The nist year 2010 speaker recognition evaluation plan," 2010.

[16] N. Scheffer and J.F. Bonastre, "Ubm-gmm driven discriminative approach for speaker verification," in *IEEE Odyssey*, San Juan, Jun 2006.

[17] H. Aronowitz, "Speaker recognition using kernel-pca and intersession variability modeling," in *INTERSPEECH*, Antwerp, Belgium, AUG 2007.