Full-covariance UBM and heavy-tailed PLDA in i-vector speaker verification

Conference Paper *in* Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on · May 2011

DOI: 10.1109/ICASSP.2011.5947436 · Source: DBLP

8 authors, including:

Pavel Matejka
Brno University of Technology
77 PUBLICATIONS **2,248** CITATIONS
SEE PROFILE

Ondrej Glembek
Brno University of Technology
53 PUBLICATIONS **3,047** CITATIONS
SEE PROFILE

Fabio Castaldo
Politecnico di Torino
17 PUBLICATIONS **561** CITATIONS
SEE PROFILE

Md Jahangir Alam
Centre de recherche informatique de ...
67 PUBLICATIONS **706** CITATIONS
SEE PROFILE

FULL-COVARIANCE UBM AND HEAVY-TAILED PLDA IN I-VECTOR SPEAKER VERIFICATION

Pavel Matějka [1], *Ondřej Glembek* [1], *Fabio Castaldo* [2], *M.J. Alam* [3,4], *Oldřich Plchot* [1], *Patrick Kenny* [3], *Lukáš Burget* [1], *and Jan "Honza" Černocky* [1]

(1) Brno University of Technology, Speech@FIT, Brno, Czech Republic, (2) Loquendo, Italy,
(3) Centre de Recherche Informatique de Montréal (CRIM), Montréal, Canada
(4) INRS-EMT, Montreal, Canada
{matejkap,glembek,burget,iplchot,cernocky}@fit.vutbr.cz,
fabio.castaldo@loquendo.it,{jahangir.alam,patrick.kenny}@crim.ca

ABSTRACT

In this paper, we describe recent progress in i-vector based speaker verification. The use of universal background models (UBM) with full-covariance matrices is suggested and thoroughly experimentally tested. The i-vectors are scored using a simple cosine distance and advanced techniques such as Probabilistic Linear Discriminant Analysis (PLDA) and heavy-tailed variant of PLDA (PLDA-HT). Finally, we investigate into dimensionality reduction of i-vectors before entering the PLDA-HT modeling. The results are very competitive: on NIST 2010 SRE task, the results of a single full-covariance LDA-PLDA-HT system approach those of complex fused system.

Index Terms— GMM, speaker recognition, PLDA, heavy-tailed PLDA, full-covariance UBM, i-vectors

1. INTRODUCTION

Total variability or "i-vector" systems have become the state-of-the-art technique in the speaker verification field [1]. They provide an elegant way of reducing the large-dimensional input data to a small-dimensional feature vector while retaining most of the relevant information. The technique was originally inspired by Joint Factor Analysis framework introduced in [2]. The basic principle is that the i-vector extractor converts sequence of feature frames to the single low dimensional vector representing the whole utterance.

A large UBM (typically with 2048 Gaussian components) with diagonal covariance matrices is used to collect statistics for the evaluation of i-vectors, which involves a lot of computation. Our idea was therefore to experiment with smaller UBMs with full-covariance matrices, in the hope of obtaining more compact representation for i-vector extraction.

When increasing the number of Gaussians of full-covariance UBM we have found that the full-covariance system has approximately the same performance as the system with diagonal covariance matrix but with 2 to 4 times less Gaussian components. We trained the full-covariance model till 2048 Gaussian components to mach it with diagonal system. We also investigated the use of new modeling techniques, such as PLDA [3] and heavy-tailed PLDA [4], that have been recently reported to overcome classical cosine-distance scoring of i-vectors with normalization [1]. Finally, a classical trick of pattern recognition — reducing the dimensionality of features before classification — was tested with encouraging results.

This paper is organized as follows: Section 2 describes thoroughly the full-covariance paradigm used and gives brief description and references of i-vector scoring. Section 3 describes our experimental setup and section 4 the results obtained on NIST 2010 SRE data. We conclude in section 5.

The work was partly supported by European project MOBIO (FP7-214324), Grant Agency of Czech Republic project No. 102/08/0707, Czech Ministry of Education project No. MSM0021630528 and by BUT FIT grant No. FIT-10-S-2. Great part of the work was done at the BOSARIS workshop held at BUT in July 2010.

2. THEORY

2.1. Universal background model

Similarly to classical speaker recognition systems, UBM is also the key element of an i-vector system, as it is necessary for collecting statistics from speech utterances. UBM contains C Gaussian components, and is defined by three sets of parameters: mean vectors $\boldsymbol{\mu}^{(c)}$, covariance matrices $\boldsymbol{\Sigma}^{(c)}$ and weights $\omega^{(c)}$. In our past work, as well as that of other labs, covariance matrices $\boldsymbol{\Sigma}^{(c)}$ are diagonal.

This work investigates into the use of full covariance matrices. These are however sensitive to (possibly very low) values of off-diagonal elements, therefore, variance flooring needs to be applied: we used floor $f\boldsymbol{\Sigma}^{avg}$ where $\boldsymbol{\Sigma}^{avg} = \sum_c^{\mathrm{C}} \boldsymbol{\Sigma}^{(c)}/\mathrm{C}$ is the average covariance matrix and $f = 0.1$ is reasonable setting. Then we set $\hat{\boldsymbol{\Sigma}}^{(c)} \leftarrow \mathrm{floor}(\boldsymbol{\Sigma}^{(c)}, f\boldsymbol{\Sigma}^{avg})$ implemented using floor function defined in [5]:

Function: $\tilde{\mathbf{S}} = \mathrm{floor}(\mathbf{S}, \mathbf{F})$

1. $\mathbf{F} = \mathbf{L}\mathbf{L}^T$ (Cholesky decomposition)
2. $\mathbf{T} \leftarrow \mathbf{L}^{-1}\mathbf{S}(\mathbf{L}^{-1})^T$ (normalization of target matrix)
3. $\mathbf{T} = \mathbf{U}\mathbf{D}\mathbf{U}^T$ (Eigenvalue Decomposition - diagonalization of target matrix)
4. Set diagonal matrix $\tilde{\mathbf{D}}$ to $\mathbf{D}$ floored to 1, i.e. $\tilde{d}_{ii} = max(d_{ii}, 1)$
5. $\tilde{\mathbf{T}} \leftarrow \mathbf{U}\tilde{\mathbf{D}}\mathbf{U}^T$ (making the matrix full again)
6. $\tilde{\mathbf{S}} \leftarrow \mathbf{L}\tilde{\mathbf{T}}\mathbf{L}^T$ (de-normalization)

2.2. I-vector extraction

I-vector system aims at modeling overall variability of the training data and compressing the information to a low-dimensional vector. The technique is closely related to JFA in the sense that each training segment acts as a separate speaker. Speaker (and/or channel) modeling techniques are then applied on these low-dimensional vectors. This way, an i-vector system can be viewed as a front-end for further modeling.

Let us first state the motivation for the i-vectors. The main idea is that the speaker- and channel-dependent GMM supervector $\mathbf{s}$ can be modeled as:

$$\mathbf{s} = \mathbf{m} + \mathbf{T}\mathbf{w} \tag{1}$$

where $\mathbf{m}$ is the UBM GMM mean supervector, $\mathbf{T}$ is a low-rank matrix representing M bases spanning subspace with important variability in the mean supervector space , and $\mathbf{w}$ is a standard normal distributed vector of size M.

For each observation $\mathcal{X}$, our aim is to estimate the parameters of the posterior probability of $\mathbf{w}$:

$$\mathrm{p}(\mathbf{w}|\mathcal{X}) = \mathcal{N}(\mathbf{w}; \mathbf{w}_\mathcal{X}, \mathbf{L}_\mathcal{X}^{-1}). \tag{2}$$

The i-vector is the MAP point estimate of the variable $\mathbf{w}$, i.e. the mean $\mathbf{w}_\mathcal{X}$ of the posterior distribution $\mathrm{p}(\mathbf{w}|\mathcal{X})$. It maps most of the relevant information from a variable-length observation $\mathcal{X}$ to a fixed- (small-) dimensional vector. $\mathbf{T}$ is referred to as the i-vector extractor.

The input data for the observation $\mathcal{X}$ is given as a set of *zero-* and *first-order statistics* — $\mathbf{n}_\mathcal{X}$ and $\mathbf{f}_\mathcal{X}$. These are extracted from F dimensional features using a GMM UBM with C mixture components, defined by a mean supervector $\mathbf{m}$, component covariance matrices $\mathbf{\Sigma}^{(c)}$, and a vector of mixture weights $\boldsymbol{\omega}$. For each Gaussian component c, the statistics are given respectively as:

$$N_\mathcal{X}^{(c)} = \sum_t \gamma_t^{(c)} \tag{3}$$

$$\mathbf{f}_\mathcal{X}^{(c)} = \sum_t \gamma_t^{(c)} \mathbf{o}_t \tag{4}$$

where $\mathbf{o}_t$ is the feature vector in time t, and $\gamma_t^{(c)}$ is its occupation probability. The complete zero- and first-order statistics supervectors are $\mathbf{f}_\mathcal{X} = \left(\mathbf{f}_\mathcal{X}^{(1)\prime}, \ldots, \mathbf{f}_\mathcal{X}^{(C)\prime}\right)'$, and $\mathbf{n}_\mathcal{X} = \left(N_\mathcal{X}^{(1)}, \ldots, N_\mathcal{X}^{(C)}\right)'$.

For convenience, we *center* the first order statistics around the UBM means, which allows us to treat the UBM means effectively as a vector of zeros

$$\begin{aligned} \mathbf{f}_\mathcal{X}^{(c)} &\leftarrow \mathbf{f}_\mathcal{X}^{(c)} - N_\mathcal{X}^{(c)} \mathbf{m}^{(c)} \\ \mathbf{m}^{(c)} &\leftarrow \mathbf{0}. \end{aligned}$$

Similarly, we "normalize" the first-order statistics and the matrix $\mathbf{T}$ by the UBM covariances, which again allows us to treat the UBM covariances as an identity matrix[1]:

$$\mathbf{f}_\mathcal{X}^{(c)} \leftarrow \mathbf{\Sigma}^{(c)-\frac{1}{2}} \mathbf{f}_\mathcal{X}^{(c)} \tag{5}$$

$$\mathbf{T}^{(c)} \leftarrow \mathbf{\Sigma}^{(c)-\frac{1}{2}} \mathbf{T}^{(c)} \tag{6}$$

$$\mathbf{\Sigma}^{(c)} \leftarrow \mathbf{I},$$

[1]Part of the factor estimation is a computation of $\mathbf{T}'\mathbf{\Sigma}^{-1}\mathbf{f}$, where the decomposed $\mathbf{\Sigma}^{-1}$ can be projected to the neighboring terms, see [6] for detailed formulae.

where $\mathbf{\Sigma}^{(c)-\frac{1}{2}}$ is a Cholesky decomposition of an inverse of $\mathbf{\Sigma}^{(c)}$, and $\mathbf{T}^{(c)}$ is a $F \times M$ sub-matrix of $\mathbf{T}$ corresponding to the c mixture component such that $\mathbf{T} = \left(\mathbf{T}^{(1)\prime}, \ldots, \mathbf{T}^{(C)\prime}\right)'$.

Note that in classical diagonal-covariance system, the normalization in (5) is done by a simple division by standard deviations.

For an observation $\mathcal{X}$, the corresponding i-vector is a point MAP estimate

$$\mathbf{w}_\mathcal{X} = \mathbf{L}_\mathcal{X}^{-1} \mathbf{T}' \mathbf{f}_\mathcal{X}, \tag{7}$$

where $\mathbf{L}$ is the precision matrix from Eq. 2 estimated as

$$\mathbf{L}_\mathcal{X} = \mathbf{I} + \sum_{c=1}^{C} N_\mathcal{X}^{(c)} \mathbf{T}^{(c)\prime} \mathbf{T}^{(c)} \tag{8}$$

Model hyper-parameters $\mathbf{T}$ are estimated using the same EM algorithm as in case of JFA [6]. For more detailed description of how to train and evaluate i-vector system see [1, 7].

2.3. Working with full covariance matrices

Full-covariance matrix in UBM plays two roles in an i-vector system:

1. in the generation of mixture component occupation probabilities $\gamma_t^{(c)}$ used in the collection of statistics (3,4).
2. in the normalization of the first order statistics (5).

The system using full covariance matrix for both is further denoted **FullCov**. This system works the best on our test set but which part of the system is the most important? Is it generation of occupation probabilities or normalization of first order statistics (see Eq. 5)? Furthermore the collection of statistics with full-covariance GMM is very computationally expensive. We were investigating and analyzing following possibilities:

FULL2DIAG_normDiag statistics were collected only with a diagonal extracted from the full covariance matrix, normalization was done using only the same diagonal.

FULL2DIAG_normFULL statistics were collected only with a diagonal extracted from the full covariance matrix, normalization was done using the full matrix.

These simplifications are however rather crude, as full-covariance model is trained and only the diagonal is extracted. With fixed means and mixture weights, this can cause quite a change to the model and frame alignment. Therefore, we have investigated following approach: in UBM training, first, diagonal covariance UBM is trained. Then, one iteration of the EM is run with fixed means and mixture weights, to obtain full-covariance model. In this way, we hope to get a tandem of coherent diagonal-covariance and full-covariance models. For utterance $\mathcal{X}$, statistics are collected using the diagonal model and the normalization is done by full-covariance model. This is denoted **Diag + FullCovNorm**.

2.4. Cosine distance

The same technique as in [1] was used. The extracted i-vectors were scaled down by an LDA matrix and further normalized i-vectors such that within-class covariance matrix is identity. Cosine distance of the two input vectors was used as the raw score:

$$\mathrm{score}\left(\mathbf{w}_\mathrm{target}, \mathbf{w}_\mathrm{test}\right) = \frac{\langle \mathbf{w}_\mathrm{target}, \mathbf{w}_\mathrm{test} \rangle}{\|\mathbf{w}_\mathrm{target}\| \|\mathbf{w}_\mathrm{test}\|}. \tag{9}$$

2.5. PLDA

The fixed-length i-vectors extracted per utterance can be used as input to standard pattern recognition algorithm. We use a Probabilistic Linear Discriminant Analysis Model (PLDA) [3] that provides a probabilistic framework applicable to fixed-length input vectors. PLDA can be seen as a special case of Joint Factor Analysis (JFA) [2] with a single Gaussian component. The i-vectors $\mathbf{w}_{\mathcal{X}}$ are assumed to be distributed according to the well-known form

$$\mathbf{w}_{\mathcal{X}} = \mathbf{m} + \mathbf{V}\mathbf{y} + \mathbf{U}\mathbf{x} + \epsilon \tag{10}$$

incorporating speaker $\mathbf{V}$ and channel $\mathbf{U}$ subspaces. Using the PLDA model, one can directly evaluate the log-likelihood ratio for the hypothesis test corresponding to "the two i-vectors were or were not generated by the same speaker." Note that the difference between *enrollment segment* (on which a model used to be created) and *test segment* (which is scored against the model) vanishes – i-vector scoring is completely symmetrical. The results for PLDA scoring are denoted **PLDA-Gaussian**.

2.6. Heavy tailed PLDA

PLDA assumes Gaussian priors of both channel and speaker factors $\mathbf{y}$ and $\mathbf{x}$ (Eq. 10). Heavy-tailed version of PLDA introduced in [4] replaces Gaussian distributions with Student's t distributions and was shown to substantially improve the SRE results compared to JFA. The results for heavy-tailed variant of PLDA are denoted **PLDA-HT**.

2.7. LDA dimensionality reduction

Our final contribution is the dimensionality reduction before PLDA modeling. Based on improved results of many classification techniques when the dimensionality of features is reduced, we decided to use standard Linear Discriminant Analysis (LDA) to process i-vectors before being scored by PLDA. The individual speakers in the development set are considered as classes when estimating LDA projection matrix. The optimum number of retained dimensions must be tuned on a development set.

3. EXPERIMENTAL SETUP

3.1. Test Set and Evaluation Metric

NIST SRE 2010 data extended core condition (telephone-telephone) was used as the evaluation data. The detection cost function (DCF) is used as a primary evaluation metric. We report two numbers: DCF_{Old} and DCF_{New} which correspond to the primary evaluation metric for the NIST speaker recognition evaluation in 2008 and 2010 respectively. The difference is that in 2010 NIST focus more on lower false alarm scenario. Third operating point - EER is also reported. For more details see evaluation plans of NIST SRE [2].

3.2. Voice Activity Detection

Speech/silence segmentation is performed by our Hungarian phoneme recognizer [8], where all phoneme classes are linked to the *speech* class. Heuristics based on short term energy are applied to discard segments with cross-talk for 2-channel files. The *interview data* we processed as 1-channel; we took ASR transcripts of the interviewer and removed his/her speech segments from our segmentation files based on time-stamps provided by NIST. Details of our VAD are provided in [9].

[2] www.itl.nist.gov/iad/mig/tests/sre/

3.3. Feature Extraction

We use MFCC 19 + energy augmented with their delta and double delta coefficients, making 60 dimensional feature vectors. The analysis window has 20 ms with shift of 10 ms. First we remove silence frames according to VAD and after that we apply short-time cepstral mean and variance normalization which uses a window of 300 frames. We have found similar performance with Short time gaussianization, but it is more efficient.

3.4. GMM UBM Training

One gender-independent UBM was represented as a full or diagonal covariance 2048-component GMM, if not stated otherwise. It was trained on the NIST SRE 2004 and 2005 telephone data (376 female speakers in 171 hours of speech, 294 male speakers in 138 hours of speech). The variance flooring was used in each iteration of EM algorithm during the UBM training.

3.5. I-vector Extractor Training

Gender-dependent i-vector extractors were trained on the following telephone data: NIST SRE 2004, 2005, 2006, Switchboard II Phases 2 and 3, Switchboard Cellular Parts 1 and 2, Fisher English Parts 1 and 2 giving 8396 female speakers in 1463 hours of speech, and 6168 male speakers in 1098 hours of speech (both after VAD). The results are reported with 400 dimensional i-vectors if not stated otherwise.

3.6. LDA/cosine distance, PLDA, PLDA-HT Training

All techniques are trained on the same data as the i-vector extractor, except for the Fisher data that was excluded, resulting in 1684 female speakers in 715 hours of speech and 1270 male speakers in 537 hours of speech.

4. RESULTS

The following experiments investigate in full-covariance UBMs. Obviously, this model has more parameters than the diagonal one. We have therefore tested also a classical diagonal-covariance UBM with more Gaussians and bigger size of i-vectors. The results in Table 1 present slight improvement by increasing the number of Gaussians and a little more by increasing the size of i-vector. However, the full-covariance UBM has superior results to all of them. It is likely that increasing the size of i-vectors will bring further improvement, this work is in progress and will be reported in the final paper.

Analysis of gathering statistics and normalization are also shown in Table 1. All three techniques including collecting statistics using the diagonal model and normalization by closely related full-covariance model (Diag + FullCovNorm) deteriorate the results, often to the level of diagonal covariance models. All results in Table 1 are reported for Female only with Cosine distance, but the results for Male and with PLDA modeling have the same trend. For cosine-distance scoring, LDA reduce the original 400-dimensional i-vectors to 200-dimensions. The cosine distance scores were normalized using gender-dependent s-norm with a cohort of 400 speakers having 2 utterances per speaker.

Table 2 reports results for different modeling techniques. Here, the results are reported for pooled genders on the NIST 2010 SRE tel-tel task. For the NIST evaluation we have experimented with the size of the model and found out that the optimal dimensions for PLDA-Gaussian is 90 for eigen-voices and full-rank (400) for eigen-channels. The same experiment was repeated with PLDA-HT with 120 dimensional eigen-voices and full-rank eigen-channel. Later,

we have also tried to reduce dimensionality of i-vectors using LDA to 90 and model these i-vectors using PLDA with 90 eigen-voices and 90 eigen-channels (both full-rank). The same was repeated with dimensionality reduction 120 for PLDA-HT. The last mentioned experiment yielded the best result on this task. The LDA reductions to 90 for PLDA-Gaussian and 120 for PLDA-HT was found to be optimal for the telephone condition and new operating point DCF_{new} on the development data. The disadvantage of using PLDA-HT is a factor of 2 or 3 times slower than the PLDA-Gaussian.

Table 1. Results on NIST2010 extended core condition 5 telephone-telephone, Cosine distance, Female only. Except for "4096 Diag" experiment, all UBMs had 2048 Gaussian components.

	DCF_{old}	DCF_{new}	EER[%]
Diag	0.1705	0.5395	3.59
4096 Diag	0.1673	0.5199	3.29
Diag - 800 i-vec	0.1520	0.4956	3.08
Diag + FullCovNorm	0.1729	0.5376	3.46
FullCov	**0.1480**	**0.4802**	**2.94**
FULL2DIAG_normFULL	0.1916	0.5617	4.02
FULL2DIAG_normDiag	0.1748	0.5443	3.73

Table 2. Different modeling for 400-dimensional i-vector extracted with full-covariance UBM with 2048 Gaussian components and LDA reduction. Results on NIST2010 extended core condition (tel-tel), pooled genders.

	DCF_{old}	DCF_{new}	EER[%]
Cosine Distance	0.1318	0.4601	2.70
PLDA-Gaussian	0.1408	0.4139	3.24
PLDA-HT	0.0973	0.3855	1.78
LDA90 PLDA-Gaussian	0.1337	0.4132	3.09
LDA120 PLDA-HT	**0.0956**	**0.3421**	**1.88**
ABC NIST 2010 system	0.0868	0.3221	1.90
LPT NIST 2010 system	0.1182	0.4020	2.44

Finally, it was compared to two NIST 2010 SRE systems:

- In **ABC system**[3] [9], the primary submission for telephone condition is a fusion of 8 different subsystems. The First group are acoustic systems with different front-ends (feature extraction, normalization, VAD) and two kinds of modeling - JFA and i-vector, and cosine distance or PLDA for scoring. The second group is based on the extraction of speaker adaptation matrices from LVCSR system (CMLLR and MLLR). The matrices are modeled by SVM. The last subsystem is JFA system which models prosodic information. We have experimented also with the different kind of quality measures mainly for the interview and microphone conditions.
- **LPT system** [10] is a fusion of different acoustic systems, based on two modeling approaches (JFA, LDA-WCCN i-vector) two set of features (MFCC, PLP) and different feature dimensions (60,25). Each single system is a combination of these three orthogonal, resulting in eight (2^3) systems in total.

The results of ABC and LPT are in the last two lines in Table 2. We see that the performance of a single system with PLDA-HT with LDA dimensionality reduction is close to the performance of our very complex fused systems.

[3]System description and presentation can be found on www.fit.vutbr.cz/research/view_pub.php?id=9346

5. CONCLUSIONS

The work we presented aims at the best performance of the single stand alone system. We have presented full-covariance UBM and i-vector extraction with different kind of modeling. Our analysis shows that for the best performance it is necessary to have full-covariance i-vector without any approximation. The heavy-tailed variant of PLDA with dimensionality reduction by LDA was shown to be superior to all previously studied approaches. However recent results show that unity length normalization of the ivector indicates that Gauss-PLDA is as effective as HT-PLDA. For more detail analysis see upcoming Interspeech paper of Daniel Garcia-Romero.

Our curent one system approach the accuracy of more complex fused systems which were submitted to NIST SRE 2010 evaluation.

In our future work, we will investigate into Gaussian pre-selection, and efficient covariance matrix modeling approaches such as semi-tight covariance model to overcome the need of full-covariance modeling because of the computational complexity of collecting GMM statistics with full-covariance UBM.

6. REFERENCES

[1] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. PP, no. 99, 2010.

[2] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of inter-speaker variability in speaker verification," *IEEE Trans. Audio, Speech and Language Processing*, vol. 16, no. 5, pp. 980–988, July 2008.

[3] S. J. D. Prince, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. International Conference on Computer Vision (ICCV)*, Rio de Janeiro, Brazil, 2007.

[4] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Proc. Odyssey 2010 - The Speaker and Language Recognition Workshop*, 2010.

[5] Dan Povey et al., "The subspace gaussian mixture model - a structured model for speech recognition," *Computer Speech and Language, in press*, 2010.

[6] P. Kenny, "Joint factor analysis of speaker and session variability : Theory and algorithms - technical report CRIM-06/08-13. Montreal, CRIM, 2005," 2005.

[7] O. Glembek, L. Burget, P. Matejka, M. Karafiat, and P. Kenny, "I-vector extraction simplified," in *submitted to ICASSP*, 2011.

[8] P. Schwarz, P. Matějka, and J. Černocký, "Hierarchical structures of neural networks for phoneme recognition," in *Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Toulouse, France, May 2006, pp. 325–328.

[9] N. Brummer, L. Burget, P. Kenny, P. Matejka, E. Villiers de, M. Karafiat, M. Kockmann, O. Glembek, O. Plchot, D. Baum, and M. Senoussauoi, "ABC system description for NIST SRE 2010," in *Proc. NIST 2010 Speaker Recognition Evaluation*. 2010, pp. 1–20, Brno University of Technology.

[10] F. Castaldo, D. Colibro, C. Vair, S. Cumani, and P. Laface, "Loquendo - Politecnico di Torino system description for NIST 2010 speaker recognition evaluation," in *Proc. NIST 2010 Speaker Recognition Evaluation*, 2010.