IDIAP RESEARCH REPORT

EMPLOYMENT OF SUBSPACE GAUSSIAN MIXTURE MODELS IN SPEAKER RECOGNITION

Petr Motlicek Subhadeep Dey Srikanth Madikeri
Lukas Burget

Idiap-RR-16-2015

JUNE 2015

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

EMPLOYMENT OF SUBSPACE GAUSSIAN MIXTURE MODELS IN SPEAKER RECOGNITION

Petr Motlicek[1], *Subhadeep Dey*[1,2], *Srikanth Madikeri*[1], *Lukas Burget*[3]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Federale de Lausanne, Switzerland
[3]Brno University of Technology, Czech Republic

{petr.motlicek,subhadeep.dey,srikanth.madikeri}@idiap.ch, burget@fit.vutbr.cz

ABSTRACT

This paper presents Subspace Gaussian Mixture Model (SGMM) approach employed as a probabilistic generative model to estimate speaker vector representations to be subsequently used in the speaker verification task. SGMMs have already been shown to significantly outperform traditional HMM/GMMs in Automatic Speech Recognition (ASR) applications. An extension to the basic SGMM framework allows to robustly estimate low-dimensional speaker vectors and exploit them for speaker adaptation. We propose a speaker verification framework based on low-dimensional speaker vectors estimated using SGMMs, trained in ASR manner using manual transcriptions. To test the robustness of the system, we evaluate the proposed approach with respect to the state-of-the-art i-vector extractor on the NIST SRE 2010 evaluation set and on four different length-utterance conditions: 3sec-10sec, 10 sec-30 sec, 30 sec-60 sec and full (untruncated) utterances. Experimental results reveal that while i-vector system performs better on truncated 3sec to 10sec and 10 sec to 30 sec utterances, noticeable improvements are observed with SGMMs especially on full length-utterance durations. Eventually, the proposed SGMM approach exhibits complementary properties and can thus be efficiently fused with i-vector based speaker verification system.

Index Terms— speaker recognition, i-vectors, subspace Gaussian mixture models, automatic speech recognition

1. INTRODUCTION

Current state-of-the-art speaker recognition is widely dominated by the use of i-vectors [1], modeled by a generative Probabilistic Linear Discriminant Analysis (PLDA). I-vector extractor represents a data-driven front-end which can map a sequence of acoustic feature vectors into a single point in a low-dimensional vector space. I-vector extractor training requires (though not manually transcribed) a large dataset with thousands of speakers.

Recently, novel frameworks for speaker recognition perform extraction of sufficient statistics for the i-vector extractor driven by an Automatic Speech Recognition (ASR) engine, such as a Deep Neural Network (DNN) [2]. In the field of speech recognition, DNNs achieve large improvements compared to standard Gaussian Mixture Models (GMMs) [3, 4]. In case of speaker recognition, DNN is able to substitute the role of the Universal Background Model (UBM), applied in the standard framework [2, 5, 6].

Another recent and successful ASR framework, especially in case of multilingual acoustic modeling and model adaptation is the Subspace Gaussian Mixture Model (SGMM) [7, 8]. SGMM has been proposed as an ASR acoustic modeling approach based on the GMM, where the parameters are represented by a more compact set and can be split into state-specific and globally-shared model parameters. Unlike DNN approach interlinked directly with the i-vector extractor, SGMMs allow for an efficient speaker-adaptation of the models using low-dimensional vectors in a "speaker subspace". These speaker vectors can therefore be exploited directly as an input for subsequent PLDA modeling in the NIST SRE 2010 speaker verification task, which is the goal of this paper. Unlike our SGMM approach for speaker verification, speaker vectors from SGMM have already been used as complementary features for language identification task [9].

The outline of the paper is as follows: Section 2 presents the SGMM modeling and speaker adaptation framework, while Section 3 summarizes the i-vector approach. The experimental protocol and corresponding results are given is Sections 4 and 5, respectively. Finally, Section 6 provides the conclusions.

2. SGMM

SGMM is able to compactly represent a large collection of mixture-of-Gaussian models and has been successfully applied in ASR tasks, especially for multilingual, or out-of-domain acoustic model adaptation [8]. Unlike conventional HMM/GMMs in which state model parameters are directly estimated from the data, subspace GMM model parameters are derived from a set of state-specific parameters, and from a set of globally-shared parameters which can capture phonetic and speaker variations.

More particularly, in the case of a conventional GMM, the likelihood is given as:

$$p(\mathbf{x} \mid j) = \sum_{i=1}^{M_j} w_{ji} \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{ji}, \boldsymbol{\Sigma}_{ji}), \quad (1)$$

where j is the state and parameters of the model are the weights w_{ji}, means $\boldsymbol{\mu}_{ji}$ and covariance matrices $\boldsymbol{\Sigma}_{ji}$. The SGMM in the basic case is given as:

$$p(\mathbf{x} \mid j, s) = \sum_{m=1}^{M_j} c_{jm} \sum_{i=1}^{I} w_{jmi} \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{jmi}^{(s)}, \boldsymbol{\Sigma}_i) \quad (2)$$

$$\boldsymbol{\mu}_{jmi}^{(s)} = \mathbf{M}_i \mathbf{v}_{jm} + \mathbf{N}_i \mathbf{v}^{(s)} \quad (3)$$

$$w_{jmi} = \frac{\exp \mathbf{w}_i^{\mathrm{T}} \mathbf{v}_{jm}}{\sum_{l=1}^{I} \exp \mathbf{w}_l^{\mathrm{T}} \mathbf{v}_{jm}}, \quad (4)$$

where $\mathbf{v}_{jm}$ are state specific vectors (with dimension similar to that of the speech features), $\mathbf{v}^{(s)}$ are speaker-specific vectors, and $\mathbf{w}_i$, $\mathbf{M}_i$, $\mathbf{N}_i$ and $\mathbf{\Sigma}_i$ are globally shared parameters. I is the number of Gaussians in the shared GMM structure, and M_j defines number of sub-states for each HMM state. This paper suggests to employ vectors $\mathbf{v}^{(s)}$ estimated for each individual speaker from enrollment and test sets as internal speaker representations to be subsequently modeled by a data-driven back-end such as PLDA. Since $\mathbf{v}_{jm}$ should correspond to a particular point in phonetic subspace, we presume that SGMM approach can factor out the phonetic variability from $\mathbf{v}^{(s)}$.

This paper exploits the symmetric version of SGMM (an extension of the original SGMM as described in [10]), which was shown to outperform the original model in ASR task. In this version, Equation 4 is modified to the following:

$$w_{jmi}^{(s)} = \frac{\exp \mathbf{w}_i^{\mathrm{T}} \mathbf{v}_{jm} + \mathbf{u}_i^{\mathrm{T}} \mathbf{v}^{(s)}}{\sum_{l=1}^{I} \exp \mathbf{w}_l^{\mathrm{T}} \mathbf{v}_{jm} + \mathbf{u}_i^{\mathrm{T}} \mathbf{v}^{(s)}}. \quad (5)$$

Vectors $\mathbf{u}_i \in \mathbb{R}^{S^{(S)}}$ now capture the effect of the speaker vectors on the weights ($S^{(S)}$ is the speaker subspace dimension). Our work is mainly interested in "speaker vectors" $\mathbf{v}^{(s)} \in \mathbb{R}^{S^{(S)}}$, which live in a "speaker-subspace" defined by matrices $\mathbf{N}_i$. Equation 3 is reminiscent of the Joint Factor Analysis approach in speaker identification [12]. The update for speaker vector estimation $\mathbf{v}^{(s)}$ is given by Equation 15 in [10]. SGMM speaker-vectors largely improve speech recognition accuracies, and it can to some extent be compared with Speaker Adaptive Training (SAT) approach [11].

3. I-VECTOR EXTRACTOR (BASELINE SYSTEM)

To compare performance of the SGMM speaker vectors, we employ the state-of-the-art i-vector extractor, implemented for speaker recognition in [1]. I-vectors represent a GMM supervector using a single total-variability subspace [13]. An i-vector $\mathbf{v}^{(s)}$ estimated from speaker and session dependent GMM supervector $\boldsymbol{\mu}^{(s)}$ can be represented by:

$$\boldsymbol{\mu}^{(s)} = \mathbf{m} + \mathbf{T}\mathbf{v}^{(s)}, \quad (6)$$

where $\mathbf{m}$ is the speaker and session independent Universal Background Model (UBM) supervector, and $\mathbf{T}$ is a low rank matrix representing the variations across a large collection of development data. $\mathbf{v}^{(s)}$ is the "i-vector" representation, normally distributed with parameters $\mathcal{N}(\mathbf{0}, \mathbf{I})$, used for speaker verification.

4. EXPERIMENTAL SETUP

In this section, we present an experimental setup and an evaluation methodology of SGMM approach in speaker verification.

4.1. Feature extraction

Throughout all the experiments, we used Mel-Frequency Cepstral Coefficients (MFCCs) as an input for acoustic modeling. More specifically, 19 cepstral coefficients were extracted using 25ms Hamming window, together with C0, calculated every 10 ms. The final 20-dimensional feature vector was subjected to the short-time mean and variance normalization using a 3 sec sliding window. MFCCs were then augmented by their delta and double-delta coefficients to provide the final 60-dimensional feature vectors.

4.2. Speech/Non-speech segmentation

First, the speech/silence segmentation is performed by a Hungarian phoneme recognizer[1]. In this approach, 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 are processed as single channel files. More details are provided in [14].

Then, this paper introduces four different length-utterance conditions (per individual speaker) for the purpose of evaluating the speaker verification performance under various utterance lengths:

- duration: full,
- truncated duration: 30 sec to 60 sec,
- truncated duration: 10 sec to 30 sec,
- truncated duration: 3 sec to 10 sec.

Throughout the following experiments, we evaluate only matched length-utterance conditions, i.e., the enrollment data is truncated into length similar to the evaluation data.

4.3. I-vector – implementation

For the i-vector extractor, we used an implementation provided by Kaldi open source software [15]. Although this may not lead to the best baseline, it allows a quick comparison with the proposed SGMM algorithm while exploiting exactly same front-end. As a development data, LDC releases of Fisher English Parts 1 and 2 were used. This gives roughly 1076 hours of speech.

First, a gender-dependent UBM with 1024 mixture components was trained on the development data. More particularly, a single diagonal GMM was first initialized and then iteratively trained using the Expectation-Maximization (E-M) algorithm. This served as an initialization to estimate a full covariance UBM. Further, a gender-dependent i-vector extractor was trained on the same data as the UBM. 400 dimensional i-vectors were extracted to represent each of the utterances. The dimensionality of the vectors was further reduced to 150 using Linear Discriminant Analysis (LDA) projection. The dimensionality-reduced vectors were length-normalized so that they conform to the Gaussian modeling assumption of the last block.

For comparison of these length-normalized i-vectors in a verification trial, we model the distribution of i-vectors using probabilistic LDA (PLDA) model. We consider implementation based on [16]. LDA and PLDA were trained on female telephone data from NIST SRE 2004, 2005, 2006, 2008, Switchboard II Phase 2 and 3 and Switchboard Cellular Parts 1 and 2. This gives in total about 850 hours of segmented speech. LDA and PLDA models trained on full length-utterance condition data were exploited throughout all the experiments.

4.4. SGMM – implementation

Similar to the i-vector extractor, SGMM acoustic model, developed in Kaldi, employs the same front-end – 60-dimensional MFCCs and speech/non-speech detection. Unlike the UBM applied in i-vector extractor, the UBM used in SGMM is trained by clustering the Gaussians from all speech classes pooled together from an HMM/GMM ASR system trained on female utterances of Fisher English Parts 1 and 2 (i.e., exploiting manual transcriptions of the dataset). Similar to i-vector system, a gender-dependent UBM with 1024 mixture components is estimated on Fisher data. The same data is used

[1] http://speech.fit.vutbr.cz/software/phoneme-recognizer-based-long-temporal-context

Table 1. *ASR results on Fisher development set in Word Error Rates (WER) [%].*

system	WER [%]
HMM/GMM	43.2
+SAT (fMLLR)	33.3
SGMM (+ speaker vectors)	31.1
HMM/DNN (+fMLLR)	23.3

to train the SGMM parameters, exploiting manual transcriptions. SGMM is initialized from the UBM with a low-dimensional phonetic subspace dimension $S = 40$, while speaker dimension $S^{(S)}$ is varied. The model has 4.3k HMM states provided by a standard decision tree approach.

4.5. Evaluation methodology

The proposed method employing SGMM was evaluated using the NIST SRE 2010 data, on the conditions 1 to 5 (cond1 – Interview - Same Microphone; cond2 – Interview - Different Microphone; cond3 – Interview - Telephone; cond4 – Interview - Microphone; cond5 – Telephone - telephone), as given by the evaluation plan [18]. Throughout the experiments we refer to these sets as the evaluation data. Speaker verification performance is reported in terms of Equal Error Rate (EER), and later on also in terms of Detection Error Tradeoff (DET) curves [19].

Since the proposed SGMM framework has been first developed for speech recognition task, we also perform intermediate ASR evaluations of the proposed technique. A Fisher development set represented by 2 hours of speech data is used. The ASR system employs a CMU dictionary [20] with 42k words and a 3gram Language Model (LM) for the decoding.

5. EXPERIMENTAL RESULTS

5.1. ASR results

Through all the experiments in this section, acoustic models were always trained on Fisher English Parts 1 and 2 (similar to SGMMs), as reported in Section 4.4. Similar to SAT approach in ASR [11], the SGMM with speaker vectors is built as a several-pass approach. At the beginning, a speaker-independent acoustic model is applied to automatically estimate speaker vectors from initial alignments. Then the speaker-adapted acoustic model is used for subsequent decoding (or lattice-rescoring).

Table 1 shows ASR results for various acoustic models, as well as the SGMM system. More particularly, first, a conventional HMM/GMM system is trained, with 4.3k HMM states (obtained by a standard decision tree approach) and 100k Gaussians. Then, SAT is performed through fMLLR (Feature space Maximum Likelihood Linear Regression). Second, SGMM is trained with the same number of HMM states. Number of sub-states is roughly equal to the number of Gaussians in the HMM/GMM model (i.e., $I = 1024, J = 4.3k, M = \sum_j M_j = 100k$, $S = S^{(S)} = 60$). Results demonstrate that the SGMM slightly outperforms the SAT HMM/GMM system on the Fisher development set. We also developed state-of-the-art HMM/DNN (hybrid) throughout training Deep Neural Network (DNN). In this approach, the DNN replaces the GMM to compute the frame-based phone posteriors. Phone classes are represented by context-dependent phones obtained by the decision tree approach from the HMM/GMM system. A six-layer DNN with 540 input nodes, 2500 nodes in each hidden layer and 3.4k output nodes was trained with cross entropy using the alignment provided by the HMM/GMM. The input layer of the DNN is composed of the context of 9 frames, where each frame corresponds to 60-dimensional MFCCs transformed using speaker-specific fMLLR (obtained also from the HMM/GMM(. Achieved results indicate that HMM/DNN largely outperforms HMM/GMM as well as SGMM systems.

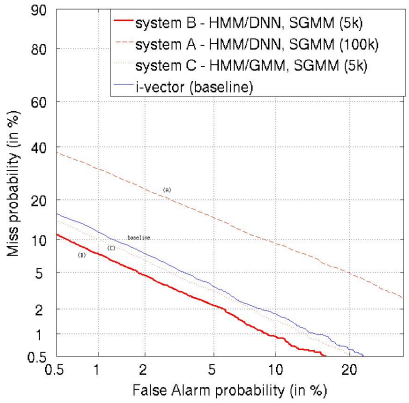

Fig. 1. DET curve – speaker verification results with scores pooled from five NIST SRE 2010 conditions (cond1 - cond5). We plot i-vector (baseline) and SGMM systems with different configurations as described in Section 5.2.1.

5.2. Speaker verification results

5.2.1. Full length-utterance condition

Figure 1 summarizes the results for multiple versions of SGMM system and compared with the i-vector baseline for the "full" length-utterance condition. The DET curve is plotted by pooling scores from all the first five (cond1 - cond5) NIST SRE 2010 conditions. The generation of speaker vectors in SGMM framework can be split into two passes. The first pass provides alignments for input speech utterances and can be easily replaced by another ASR system capable of automatically generating state-level alignments. The second pass loads initial alignments and employs the SGMM to estimate speaker vectors. More precisely, we experimented with these combination of system passes:

- system A - (pass 1) HMM/DNN+fMLLR, (pass 2) SGMM (100k sub-states, $S = S^{(S)} = 60$). This system applies LDA without dimensionality reduction;
- system B - (pass 1) HMM/DNN+fMLLR, (pass 2) SGMM (5k sub-states $\sim$ 1 sub-state per HMM state, $S = 60$, $S^{(S)} = 400$). According to [7], introducing sub-states in SGMM continuously improves ASR performance. In an usual setting, M is roughly equal to the number of Gaussians of an HMM/GMM system. Nevertheless, in our paper SGMM is exploited as a "speaker vector" extractor. We therefore hypothesize that more emphasis should be given to the speaker subspace by reducing the phonetic subspace. We also increase subspace dimension $S^{(S)}$ to be equal to the dimension of i-vectors, although we are aware that this leads to an increase in the number of parameters in $\mathbf{N}_i$ (given in Equa-

Table 2. *Comparison of the i-vector extractor baseline with the proposed SGMM system in terms of EERs for NIST SRE 2010 evaluation task.*

NIST SRE 2010 (female), EER[%]					
system	cond1	cond2	cond3	cond4	cond5
length-utterance: 3 sec to 10 sec					
i-vector	**17.0**	**21.4**	20.1	19.9	21.1
SGMM	17.3	21.7	**19.7**	**19.7**	**20.8**
fusion	14.7	18.6	16.2	17.5	18.8
length-utterance: 10 sec to 30 sec					
i-vector	**5.7**	**8.9**	9.0	8.9	9.6
SGMM	6.6	9.9	**7.9**	**7.4**	**7.6**
fusion	5.1	8.2	6.6	6.4	6.5
length-utterance: 30 sec to 60 sec					
i-vector	**2.8**	**4.6**	4.4	3.4	5.9
SGMM	3.1	4.7	**3.3**	**2.7**	**4.2**
fusion	2.8	4.1	2.9	2.6	4.0
length-utterance: full					
i-vector	1.6	2.7	2.2	1.8	2.5
SGMM	**1.3**	**2.4**	**2.1**	**1.2**	**2.0**
fusion	1.3	2.3	1.5	1.2	1.7

tion 3) which may lead to parameter estimation problems on modestly-sized systems;

- system C - (pass 1) HMM/GMM, (pass 2) SGMM (5k sub-states $\sim$ 1 sub-state per HMM state, $S = 60$, $S^{(S)} = 400$));

Similar to the i-vector extractor, SGMM speaker vectors are projected by LDA (dimensional reduction from 400 to 150), length-normalized and finally modeled by PLDA. Achieved results visualized in terms of DET curves indicate that (i) dimension expansion of speaker subspace ($S^{(S)} = 400$) together with reducing phonetic sub-states bring significant improvement (system A $\rightarrow$ B), (ii) an alignment obtained by the best ASR system (i.e., the HMM/DNN) further improves speaker verification results (system C $\rightarrow$ B).

5.2.2. Full and truncated length-utterance conditions

Further experiments take into account the best SGMM system (system B) described in the previous section. The SGMM system is applied to other three truncated length-utterance conditions (as defined in Section 4.2). Similar to Figure 1, Figure 2 plots DET curves by pooling scores from NIST SRE 2010 conditions 1 to 5. In addition, Table 2 demonstrates speaker verification results in terms of Equal Error Rates (EERs) individually for each length-utterance and NIST SRE 2010 condition.

In addition to SGMM and i-vector systems, Figure 2 and Table 2 show results for system fusion. More particularly, scores are linearly combined with weights equal to 0.3 and 0.7 for i-vector and SGMM systems respectively. We did not perform any calibration.

As expected, the speaker verification performance for both systems significantly degrades for short durations. The largest degradation can be observed for length-utterances 3sec to 10sec. In general, i-vector and SGMM systems provide complementary results. I-vector system performs better for NIST SRE 2010 conditions 1 and 2 (interview speech), while SGMM outperforms the i-vector system for conditions 3 to 5 (conversational speech). Furthermore, as clearly seen from DET curve in Figure 2 for the case of pooling scores from conditions 1 to 5, i-vector system achieves better performance for 3 sec to 10 sec and 10 sec to 30 sec long utterances. In case of 30 sec to 60 sec, performances are similar. If the utterances were not truncated (duration full), SGMM performs noticeably better.

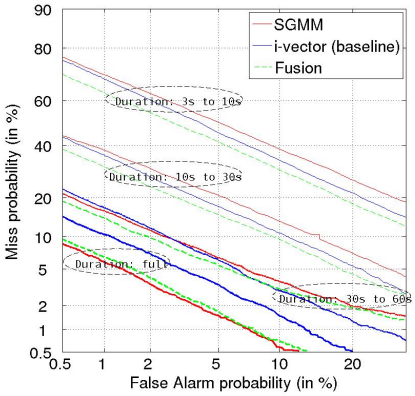

Fig. 2. DET curve – speaker verification results with scores pooled from five NIST SRE 2010 conditions (cond1 - cond5) for four different length-utterance conditions (3 sec-10 sec, 10 sec-30 sec, 30 sec-60 sec, untruncated (full)).

As mentioned before, LDA and PLDA models were always trained on full length-utterance condition using development data. We also experimented with LDA+PLDA models trained on truncated utterances, as partially motivated by [21], but without observing any improvement.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed an alternative approach for speaker recognition based on employment of speaker vectors estimated using the SGMM framework. The proposed approach integrates speech recognition in the speaker modeling process by using SGMM trained in the ASR manner. The proposed system operates in two passes. We showed that the first pass, which provides an alignment for a speech utterance, can be efficiently replaced by state-of-the-art HMM/DNN ASR system. The second (SGMM) pass exploits initial alignments and estimates speaker vectors. Eventually, SGMM-based speaker vectors are modeled by a Bayesian back-end represented by PLDA.

Experimental results suggest that SGMM system outperforms i-vector for NIST SRE 2010 conditions 3 to 5 (conversational speech), while the baseline system is generally better for conditions 1 and 2 (interview speech). In terms of different length-utterance experiments (matched conditions), the proposed SGMM system outperforms the baseline for untruncated (full) length-utterance condition. In case of very short (3 sec-10 sec and 10 sec-30 sec) utterances, the i-vector system gave overall better performance. The proposed SGMM approach exhibits complementary properties and can thus be efficiently fused with i-vector based speaker verification system. A natural next step for this work includes testing on mismatched length-utterance conditions.

7. ACKNOWLEDGMENTS

This work was supported by Speaker Identification Integrated Project (SIIP), funded by the European Unions Seventh Framework Programme for research, technological development and demonstration under grant agreement no 607784.

8. REFERENCES

[1] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 19, pp. 788-798, May 2010.

[2] Y Lei, N. Scheffer, L. Ferrer and M. McLaren"A novel scheme for speaker recognition using phonetically-aware deep neural net," *in Proc. of ICASSP*, Florence, Italy, 2014.

[3] G. Hinton, Li Deng, Dong Yu, G.E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine*, IEEE, vol. 29, no. 6, pp. 82-97, 2012.

[4] G.E. Dahl, Dong Yu, Li Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 20, pp. 30-42, 2012.

[5] D. Garcia-Romero, X. Zhang, A. McCree, and D. Povey, "Improving Speaker Recognition Performance in the Domain Adaptation Challenge using Deep Neural Networks", *in Proc. of SLT*, USA, 2014.

[6] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," *in Proc. of ICASSP*, 2014.

[7] D. Povey, L. Burget, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. K. Goel, M. Karafiat, A. Rastrow, R. C. Rose, P. Schwarz and S. Thomas, "The Subspace Gaussian mixture model - A structured model for speech recognition," In Computer Speech & Language, vol. 25, no. 2, pp. 404-439, 2011.

[8] L. Burget, P. Schwarz, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. K. Goel, M. Karafiat, D. Povey, A. Rastrow, R. C. Rose and S. Thomas, "Multilingual Acoustic Modeling For Speech Recognition Based On Subspace Gaussian Mixture Models", *in Proc. of ICASSP*, pp. 4334-4337, Dallas, USA, 2010.

[9] O. Plchot, M. Karafiat, N. Brummer, O. Glembek, P. Matejka, E. de Villiers and J. Cernocky, "Speaker vectors from Subspace Gaussian Mixture Model as complementary features for Language Identification," *in Proc. of Odyssey 2012*, pp. 330-333. Singapore, 2012.

[10] D. Povey. M. Karafiat, A. Ghoshal and P. Schwarz, "A symmetrization of the Subspace Gaussian Mixture Model," *in Proc. of ICASSP*, pp. 4504 - 4507, Prague, Czech Republic, 2011

[11] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A Compact Model for Speaker-Adaptive Training," *in Proc. of ICSLP*, pp. 1137 - 1140, Philadelphia, USA, 1996.

[12] P. Kenny, P. Ouellet, N. Dehak, and V. Gupta, "A study of interspeaker variability in speaker verification," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 16, no. 5, pp. 980-987, 2008.

[13] N. Dehak, R. Dehak, P. Kenny, N. Brummer, P. Ouellet and P. Dumouchel "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," *in proc. of INTERSPEECH*, pp. 1559-1562, 2009.

[14] N. Brummer, L. Burget, P. Kenny, P. Matejka, E. Villiers, M. Karafiat, M. Kockmann, O. Glembek, O. Plchot, D. Baum, and M. Senoussauoi, "ABC system description for NIST SRE 201, N. Scheffer, L. Ferrer, and M. McLaren, A novel scheme for speaker recognition using a phonetically-aware deep neural network, in International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2014.0," *in Proc. NIST 2010 Speaker Recognition Evaluation*, pp. 1-20, Brno University of Technology. 2010.

[15] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, " The Kaldi Speech Recognition Toolkit", *in Proc. of ASRU*, pp. 4, Hawaii, USA, December 2011.

[16] S. Ioffe, "Probabilistic Linear Discriminant Analysis," *in Computer Vision - ECCV*, Lecture Notes in Computer Science, Vol. 3954, pp. 531-542, 2006.

[17] D. Garcia-Romero and A. McCree, "Supervised domain adaptation for I-vector based speaker recognition," *in Proc. of ICASSP*, pp. 4047 - 4051, Florence, Italy, 2014.

[18] "2010 NIST Speaker Recognition Evaluation," http://www.itl.nist.gov/iad/mig/tests/sre/2010.

[19] A. Martin, G. Doddington, T. Kamm, M. Ordowski and M. Przybocki, "The DET Curve in Assessment of Detection Task Performance", *in Proc. of Eurospeech*, pp. 1895-1898, Greece, 1997.

[20] ArpaBet, "CMU pronouncing dictionary," http://www.speech.cs.cmu.edu/cgi-bin/cmudict, 2013.

[21] A. Kanagasundaram, D. Dean, J. Gonzalez-Dominguez, S. Sridharan, D. Ramos and J. Gonzalez-Rodriguez, "Improving Short Utterance based I-vector Speaker Recognition using Source and Utterance-Duration Normalization Techniques," *in Proc. of INTERSPEECH*, pp. 2465 - 2469, Lyon, France, 2013.