IMPROVING SPEAKER RECOGNITION PERFORMANCE IN THE DOMAIN ADAPTATION CHALLENGE USING DEEP NEURAL NETWORKS

Daniel Garcia-Romero, Xiaohui Zhang, Alan McCree, Daniel Povey

Human Language Technology Center of Excellence & Center for Language and Speech Processing
The Johns Hopkins University, Baltimore, MD 21218, USA
{dgromero, xiaohui, alan.mccree}@jhu.edu, dpovey@gmail.com

ABSTRACT

Traditional i-vector speaker recognition systems use a Gaussian mixture model (GMM) to collect sufficient statistics (SS). Recently, replacing this GMM with a deep neural network (DNN) has shown promising results. In this paper, we explore the use of DNNs to collect SS for the unsupervised domain adaptation task of the Domain Adaptation Challenge (DAC). We show that collecting SS with a DNN trained on out-of-domain data boosts the speaker recognition performance of an out-of-domain system by more than 25%. Moreover, we integrate the DNN in an unsupervised adaptation framework, that uses agglomerative hierarchical clustering with a stopping criterion based on unsupervised calibration, and show that the initial gains of the out-of-domain system carry over to the final adapted system. Despite the fact that the DNN is trained on the out-of-domain data, the final adapted system produces a relative improvement of more than 30% with respect to the best published results on this task.

Index Terms— Unsupervised adaptation, speaker recognition, i-vectors, deep neural networks

1. INTRODUCTION

Current speaker recognition systems model i-vectors [1] with variants of Probabilistic Linear Discriminant Analysis (PLDA) [2, 3, 4, 5, 6, 7]. Given a large collection of labeled data (speaker labels), PLDA provides a powerful data-driven mechanism to separate speaker information from other sources of undesired variability. Typically, the PLDA systems are trained on tens of thousands of speech cuts from thousands of speakers with multiple cuts per speaker from different sessions. Assuming such a large amount of resources for every new domain of interest might be too expensive or even unrealistic. One way to alleviate this burden is to use domain adaptation to bootstrap an already available resource-rich out-of-domain system to produce good results in a new domain for which only unlabeled data is available.

To facilitate the study of domain adaptation techniques, MIT-LL[1] has designed a domain adaptation challenge (DAC) using Linguistic Data Consortium (LDC) telephone corpora. The DAC comprises an out-of-domain (OOD) set and and in-domain (IND) set that matches the evaluation data. The mismatch of the OOD set is attributed to the evolution of telephony systems over the years [8]. The DAC was extensively used during the CLSP 2013 summer workshop were two families of approaches were used to mitigate the mismatch problem. The first one [8, 9, 10, 11] focused on parameter adaptation, while the second one focused on compensation techniques [12, 13, 14]. In this paper, we focus on the unsupervised parameter adaptation family and select the approach in [10] due to its efficiency and performance. In particular, the approach uses the OOD set to build a PLDA system and then uses it to cluster the IND dataset. This produces an estimate of the IND speaker labels that are subsequently used to adapt the parameters of the PLDA system to the new domain. The clustering is based on agglomerative hierarchical clustering (AHC) using the metric induced by the OOD PLDA system. Recently, NIST has proposed an i-vector challenge (IVC) [15] in which an unlabeled IND dataset is provided but no OOD data is given. This setup was conductive to clustering and similar AHC approaches (with a variety of different metrics) were successful in that task [16, 17, 18, 19].

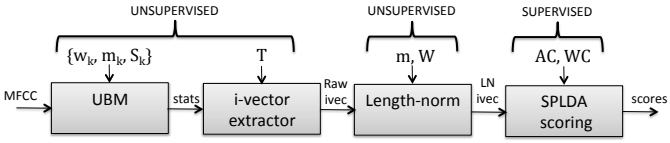


Fig. 1: Block diagram of speaker recognition system indicating which parameters are trained in supervised and unsupervised mode.

The traditional i-vector framework used in [10] uses a GMM to collect SS. The work in [20] has shown that replacing the GMM with a DNN to compute SS produces significant improvements (under controlled conditions that result in an in-domain setup). The DNN effectively leverages transcribed data and is able to produce soft classifications (in terms of posterior probabilities) of speech frames into sub-phonetic categories (senones). The alignment of speech frames to sub-phonetic categories facilitates the comparison of speakers when they are producing the same content. In the same spirit, the work in [21] proposed the use of a phonetically-aware UBM obtained from an ASR system. However, the performance improvements obtained by the more recent DNN approach are much larger. Also, the same concept of SS computation with DNN was explored in parallel by [22] and found it less promissing. The results in this paper are more in line with the optimistic findings in [20].

In this work, we integrate the DNN into our unsupervised domain adaptation approach and evaluate its merits in the DAC. We explore the effects of the DAC telephony mismatch and the influence of the amount of training data in the DNN. The reminder of the paper is organized as follows: Section 2 describes the system architecture and the role of the DNN. Section 3 summarizes the unsupervised adaptation technique. Section 4 describes our experimental setup and results. Finally, section 5 provides the conclusions.

[1]The authors thank MIT-LL for the domain adaptation challenge. A detailed description and resources (lists, i-vectors, and PLDA system) are available at: http://www.clsp.jhu.edu/workshops/archive/ws13-summer-workshop/groups/spk-13/

2. SPEAKER RECOGNITION SYSTEM

Figure 1 shows a block diagram of a state-of-the-art i-vector speaker recognition system. The first two blocks serve as a data-driven front-end that maps sequences of MFCCs into a low-dimensional vector denoted as i-vector [1]. The third block is a pre-processing stage that conditions the i-vectors so that they conform to the Gaussian modeling assumptions of the last block [6]. The goal of the final block is to produce a similarity score, based on the PLDA model [6], that is higher as the likelihood of an i-vector $\mathbf{x}_t$ belonging to speaker i increases. An efficient computation of this score is presented in [23].

On top of each block, Figure 1 shows the set of parameters that need to be trained. The term supervised/unsupervised indicates if the parameters require speaker labels or not. The parameters that do not require speaker labels are much easier to adapt since unlabeled in-domain data is much easier to acquire. In [9], we explored the impact of adapting all the parameters. Overall, it was observed that the largest improvement is obtained by adapting the PLDA parameters. Adapting the length-normalization is also important, whereas using an in-domain UBM and $\mathbf{T}$ matrix is not crucial. Therefore, in this paper we focus on adaptation of the length-normalization and PLDA parameters.

2.1. Role of the DNN

Traditional i-vector systems rely on a GMM-UBM to provide soft alignments of acoustic frames (i.e. MFCCs) to compute sufficient statistics [1]. Each mixture of the GMM represents a region/class and provides a context in which to characterize how speakers differ from each other. Ideally one would like these regions to correspond to phonetic content (i.e. to allow comparisons of how speakers differ in pronouncing the same content). However, the unsupervised nature of the GMM training does not guarantee this property. To enforce this property, the authors in [20, 22] have proposed to replace the GMM with a DNN that has been explicitly trained to discriminate between tied triphone states (senones). In this way, the DNN is in charge of providing the class/region alignments for the SS computation.

Figure 2 highlights the differences between the GMM and the DNN approaches. Notice that while the traditional GMM-based approach uses the same acoustic features (i.e. MFCCs designed for good speaker recognition performance) to compute the SS and to obtain the alignments (frame posteriors), the DNN-based approach uses ASR specific features to compute the alignments and then speaker features for the SS. Moreover, the DNN parameters Θ are trained using a transcribed training set. This extra piece of supervision is what allows the DNN to provide alignments that are phonetically-aware.

3. UNSUPERVISED DOMAIN ADAPTATION

3.1. Length-normalization

Length-normalization (LN) is a two step process that Gaussianizes i-vectors [6]. In the first step, the i-vectors are centered and whitened based on the sample mean and covariance of a training dataset. This produces the global mean $\mathbf{m}$, and the whitening transform $\mathbf{W}$. In the second step, the centered and whitened i-vectors are projected into the unit sphere. Since the estimation of $\mathbf{m}$ and $\mathbf{W}$ does not required labeled data, unsupervised adaptation of LN is straightforward. In [9] it was shown that a strategy with dataset-dependent centering (center each dataset around their sample mean) and common whitening (based on in-domain statistics) produces the best results. This is the strategy that will be used in this paper when presenting results with adapted LN.

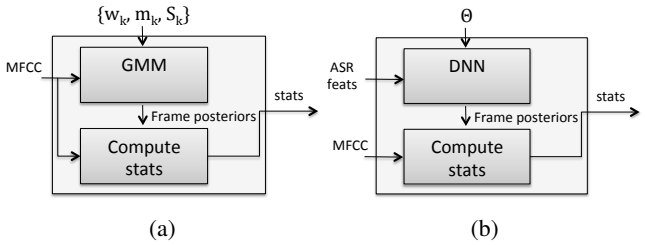


Fig. 2: Diagram of the (a) GMM-based and (b) DNN-based sufficient statistics computation.

3.2. PLDA parameters

In this section, we summarize the approach that we recently proposed in [10] to adapt the across-class and within-class covariances ($\mathbf{\Gamma}$, $\mathbf{\Lambda}$) of an already available PLDA system (which was trained on labeled out-of-domain data), to a new domain for which only unlabeled data is available. The approach uses the out-of-domain PLDA system to cluster the in-domain dataset. These clusters are treated as speakers and are subsequently used to adapt the parameters of the PLDA system to the new domain. We now describe the three key components of the approach: clustering technique, determination of number of clusters, and the adaptation mechanism.

3.2.1. Clustering

An exhaustive search over all possible partitions of a dataset is not scalable for large sets due to the combinatorial nature of the problem (e.g. for a set of size $N = 10$, there are already 115,975 partitions). Instead, to reduce the search space, we found in [8] that a greedy search based on *agglomerative hierarchical clustering* (AHC) is a good alternative. That is, starting with each i-vector as a separate cluster, at every step, we merge the two clusters that are closer based on a predefined metric. This merging schedule defines a path over the space of partitions and a final clustering is obtained based on a stopping criterion.

We use the out-of-domain PLDA system to define the metric by computing a pairwise similarity matrix between all i-vectors [8]. Then, the similarity between two clusters (i.e. linkage criterion) is defined as the average of the pairwise similarities between the elements of each cluster. Note that this approach only requires averaging scores from the precomputed pairwise similarity matrix. Therefore, AHC score averaging is computationally efficient.

3.2.2. Determination of number of clusters

To estimate the number of clusters, we define a threshold and stop the merging process when the similarity between the clusters to be merged goes below the threshold. A principled way of doing this is to calibrate the scores of the PLDA system so that we can use Bayesian decision theory to set a threshold analytically. We use an unsupervised calibration approach [24] where only unlabeled in-domain scores are required. This approach uses a generative model of scores [25] and fits a 2 component Gaussian mixture model (GMM) to a collection of unlabeled in-domain scores. The covariances of the GMM are tied and therefore the calibration mapping is affine. Once we learn a calibration mapping, we stop the AHC when

Table 1: Configuration of the two DNNs. WER is reported on the SWB subset of the Hub5 2000 evaluation set.

System name	SWB Train	Senones	Hidden layers	p-norm p/in/out	WER (%)
DNN-1	100h	4295	5	2/4000/400	19.7
DNN-2	300h	9006	5	2/5000/500	16.3

the calibrated similarity between the clusters to be merged goes below 0. That is, when the evidence in favor of the different-speaker hypothesis, $\mathcal{H}_d$, exceeds the evidence in favor of the same-speaker hypothesis, $\mathcal{H}_s$.

3.2.3. Adaptation

Once the in-domain dataset is clustered, it can be used as a labeled dataset to perform supervised adaptation of the PLDA parameters $\mathbf{\Gamma}$, and $\mathbf{\Lambda}$. Note that for the adaptation we only need the assignment of i-vectors to clusters, so there is no need to align clusters with real speaker identities. In [9], we studied four adaptation approaches and found them to perform very similarly. In this work, we use the PLDA parameter interpolation approach. That is, we use the estimated labels and the PLDA EM algorithm to obtain in-domain parameters, and then, we interpolate them with the out-of-domain parameters:

$$\begin{aligned} \mathbf{\Gamma}_{adapt} &= \alpha\, \mathbf{\Gamma}_{in} + (1-\alpha)\, \mathbf{\Gamma}_{out}, \\ \mathbf{\Lambda}_{adapt} &= \alpha\, \mathbf{\Lambda}_{in} + (1-\alpha)\, \mathbf{\Lambda}_{out}. \end{aligned} \tag{1}$$

The larger the interpolation parameter $\alpha \in [0, 1]$, the larger the contribution of the in-domain data. Note that this approach does not require access to the out-of-domain i-vectors.

4. EXPERIMENTS

4.1. Datasets

For our experiment we use the DAC created by MIT-LL. The setup uses the SRE10 telephone data [26] (condition 5 extended task) as enrollment (single cut) and test sets. This evaluation set provides 7,169 target and 408,950 non-target trials. For parameter training, two datasets were defined to expose the effects of domain mismatch. The in-domain SRE set comprises telephone calls from 3,790 speakers (male and female) and 36,470 speech cuts taken from pre-SRE10 collections. The out-of-domain SWB set comprises telephone calls from 3,114 speakers (male and female) and 33,039 speech cuts taken from Switchboard-I and II. Although the statistics of both datasets are quite similar, the SRE set matches the SRE10 evaluation data better than SWB. As exposed by the analysis presented in [8], the main mismatch cause is the evolution of telephony systems over time. For our unsupervised adaptation experiments we ignore the labels of the in-domain SRE set and refer to it as SRE-U (for unlabeled).

4.2. DNN setup

We trained two 5-hidden-layer p-norm neural networks [27] with power p = 2. The training recipe is basically the same as described in [27], except that fMLLR transform is removed both during training and testing. DNN-1 uses p-norm input/output dimensions set to 4000/400 for each hidden layer and it is trained on a 100h subset of the standard Switchboard 300 hour corpus (LDC97S62). DNN-2 uses dimensions 5000/500 for each hidden layer and it is trained on the whole 300 hour corpus. For word error rate (WER) evaluation we used a trigram language model (LM) trained on 3M words of Switchboard training transcripts and then interpolated it with another trigram LM trained on 11M words of the Fisher English Part 1 transcripts (LDC2004T19). The WER on the SWB subset of the Hub5 2000 evaluation set (LDC2002S09, also known as eval2000) is 16.3% for the larger network, and is 19.7% for the smaller network. The configuration for both nets is summarized in Table 1.

Table 2: Configuration for each task (OOD: out-of-domain; UA-LN: unsupervised adaptation of length-normalization; UA-LN-PLDA: unsupervised adaptation of length-normalization and PLDA; IND: in-domain).

Task	UBM, T	m,W	Γ, Λ
OOD	SWB	SWB	SWB
UA-LN	SWB	SRE-U	SWB
UA-LN-PLDA	SWB	SRE-U	SWB/SRE-U
IND	SWB	SRE	SRE

4.3. Speaker recognition system setup

4.3.1. GMM-based baseline

The baseline system in Figure 1 uses 40-dimensional MFCCs (20 base + deltas) with short-time mean and variance normalization. It is configured in a completely gender-independent way. It uses a 2048 mixture UBM with a 600 dimensional i-vector extractor, and a speaker subspace of 400 dimensions for PLDA. We report recognition performance in terms of equal error rate (EER) and/or normalized minimum detection cost function (DCF) [26] with probability of target trial set to either 10^{-2} or 10^{-3}, and the cost of misses and false alarms set to 1.

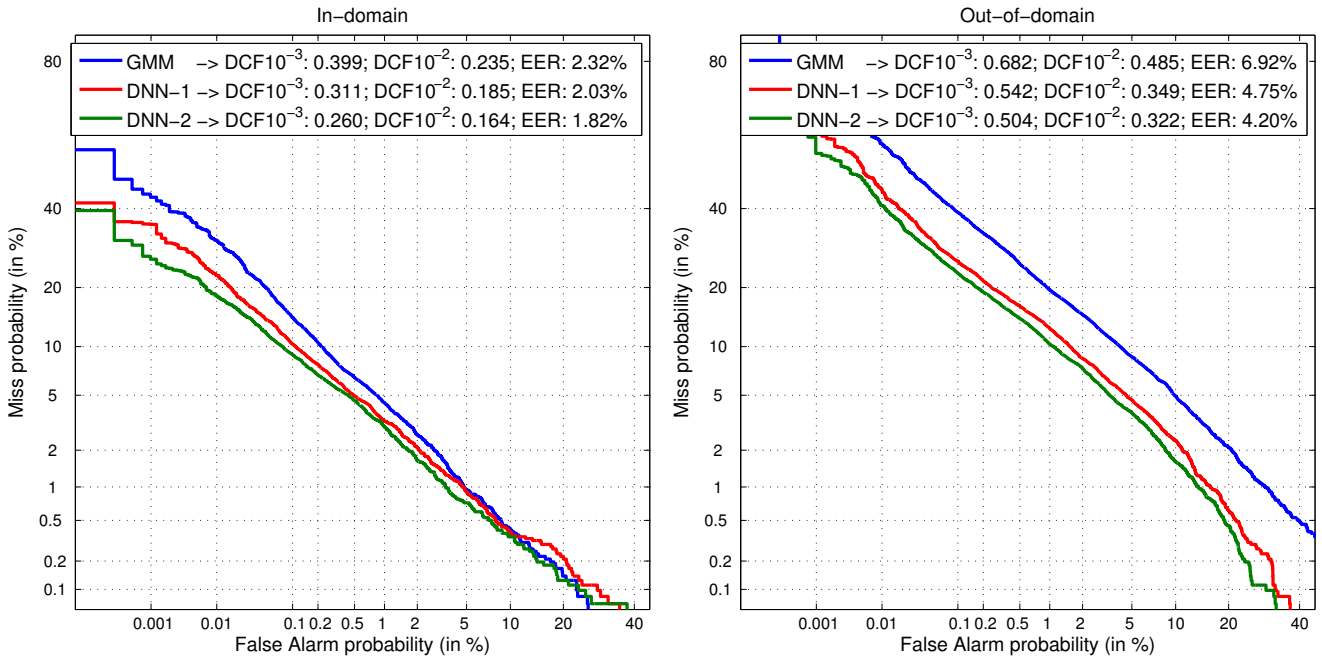
For the unsupervised adaptation of PLDA parameters (see eq. 1) we used a fixed value of $\alpha = 0.7$. This value was not tuned as the final speaker recognition performance is not very sensitive to it [9]. We explored running multiple iterations of the clustering and adaptation stages but the performance did not improve over just one iteration. The number of estimated clusters based on our stopping criterion was 3023 (3790 speakers in SRE).

4.3.2. DNN-based system

The only differences between the DNN and GMM-based configurations are due to the different way to compute the frame posteriors. The posteriors of the DNN-based system are used to compute the SS and to define an ancillary UBM needed for the i-vector computation [20, 22]. The number of mixtures of this UBM is given by the number of senones. We use full-covariance mixtures as in [20]. During the adaptation stage, the number of estimated clusters was 2957 (slightly smaller than for the GMM baseline). Only one iteration was used as in the baseline.

4.4. Results

Table 2 summarizes the tasks/configurations under which we have evaluated the systems. The OOD and IND configurations are used to quantify the performance gap due to the domain mismatch in the DAC. The unsupervised adaptation experiments (UA-LN and UA-LN-PLDA) indicate how much of that gap we are able to recover based on each technique.



(a) (b)

Fig. 3: Performance in-domain and out-of-domain

Figure 3 shows the DET curves for the three systems (baseline and the two DNN configurations) for the IND and OOD configurations. We can observe that the DNN-based systems greatly outperform the GMM-based baseline at all operating points in both tasks. However, the gap between the IND and OOD tasks still large for the DNN systems. We can also see that the DNN-2 system is better than the DNN-1 system. This indicates that a better senone classification accuracy (indirectly measured by the WER) has a positive impact in the speaker recognition performance. Hence, validating the hypothesis that a consistent partition of the acoustic space based on phonetic content facilitates speaker recognition. Also, the great performance of the DNN-based systems indicates that, despite the fact that the DNNs were only trained on SWB-I (small subset of the out-of-domain data), the estimation of the frame posteriors is robust to the telephony mismatch of the DAC.

As shown in Table 3, for the DNN-2 system, the unsupervised adaptation of LN and PLDA parameters recovers more than 90% of the OOD and IND gap (for the three operating points). For the GMM-based system the recovered proportion is around 85%. The fact that the DNN-based system closes a bigger portion of the gap is attributed to the better initial performance of the DNN-2 OOD system used to cluster the SRE-U set. Also, multiple iterations of clustering and PLDA adaptation did not help either the GMM-based nor the DNN-based systems. This suggest that for our adaptation approach, the starting point of the OOD system has a big impact in the potential of the adapted system to match the performance of a

Table 3: Performance comparison of GMM and DNN-2 systems for different tasks (see Table 2 for configuration details).

Task	System	DCF10^{-3}	DCF10^{-2}	EER(%)
OOD	GMM	0.682	0.485	6.92
	DNN-2	0.504	0.322	4.20
UA-LN	GMM	0.627	0.425	5.55
	DNN-2	0.431	0.273	3.31
UA-LN-PLDA	GMM	0.445	0.264	2.72
	DNN-2	0.271	0.172	2.09
IND	GMM	0.399	0.235	2.32
	DNN-2	0.260	0.164	1.82

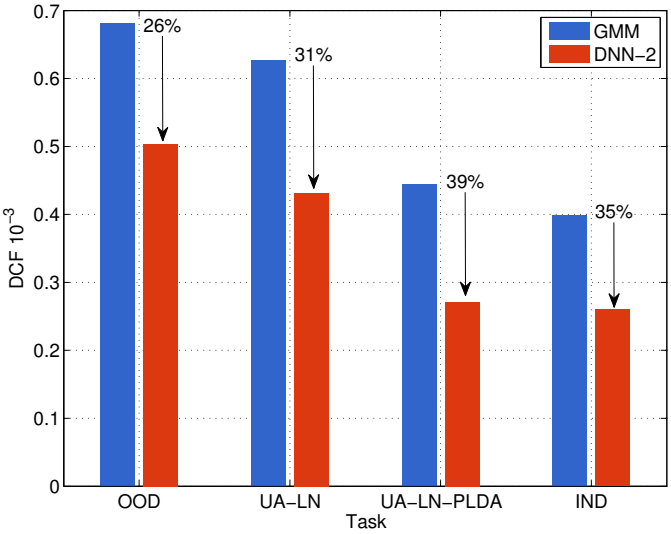


Fig. 4: Performance of the GMM and DNN-2 systems at the DCF-10^{-3} operating point for different configurations. The arrows indicate the relative improvement of the DNN-2 over the GMM system.

costly in-domain system.

In Figure 4 we highlight the relative improvements of DNN-2 system versus the GMM baseline for all configurations. Note that the initial gains of the DNN-2 OOD system carry over to the final adapted system, and they even get larger. Using the DNN-2 to compute frame posteriors results in an adapted (UA-LN-PLDA) system that significantly outperforms the oracle in-domain GMM baseline.

5. CONCLUSION

In this paper, we explored the use of DNNs to collect SS for a PLDA i-vector system in the DAC. We showed that the OOD performance of the DNN-based system is superior to the GMM baseline and that this advantage carries through the unsupervised adaptation process. The final DNN-based adapted system not only outperforms its GMM-based counterpart by almost 40% (relative improvement at DCF10^{-3}), but also outperforms the oracle IND baseline. We evaluated the influence of the amount of transcribed data to train the DNN in the final speaker recognition performance. We compared two DNNs trained with 100h and 300h of SWB-I data (small subset of our OOD set) and observed that the larger DNN achieved better WER performance that translated into better speaker recognition results. Even though the DNN was trained on OOD data, the excellent speaker recognition performance of the adapted system shows a consistent estimation of senone posteriors for the IND data. Since the main mismatch between the OOD and IND data is related to the evolution of telephony systems over time, this indicates robustness of the DNNs to this mismatch. Overall, using a DNN to collect SS produces a system that achieves the best published results on the unsupervised domain adaptation task of the DAC.

6. REFERENCES

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

[2] S. Prince and J. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *IEEE International Conference on Computer Vision (ICCV)*, Rio de Janeiro, 2007.

[3] N. Brümmer and E. De Villiers, "The speaker partitioning problem," in *Odyssey: The Speaker and Language Recognition Workshop*, Brno, Czech Republic, 2010.

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

[5] J. Villalba and N. Brümmer, "Towards fully Bayesian speaker recognition: Integrating out the between-speaker covariance," in *Interspeech*, Florence, Italy, August 2011.

[6] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Interspeech*, Florence, Italy, August 2011.

[7] D. Garcia-Romero and C. Y. Espy-Wilson, "Multicondition training of Gaussian PLDA models in i-vector space for noise and reverberation robust speaker recognition," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Kyoto, Japan, 2012.

[8] S. Shum, D. Reynolds, D. Garcia-Romero, and A. McCree, "Unsupervised clustering approaches for domain adaptation in speaker recognition systems," in *Odyssey: The Speaker and Language Recognition Workshop*, 2014.

[9] D. Garcia-Romero and A. McCree, "Supervised domain adaptation for i-vector based speaker recognition," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.

[10] D. Garcia-Romero, A. McCree, S. Shum, N. Brümmer, and C. Vaquero, "Unsupervised domain adaptation for i-vector speaker recognition," in *Odyssey: The Speaker and Language Recognition Workshop*, 2014.

[11] J. Villalba and E. Lleida, "Unsupervised adaptation of PLDA by using variational bayes methods," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.

[12] H. Aronowitz, "Inter dataset variability compensation for speaker recognition," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.

[13] H. Aronowitz, "Compensating inter-dataset variability in PLDA hyper-parameters for robust speaker recognition," in *Odyssey: The Speaker and Language Recognition Workshop*, 2014.

[14] O. Glembek, J. Ma, P. Matejka, B. Zhang, O. Plchot, L. Burget, and S. Matsoukas, "Domain adaptation via within-class covariance correction in i-vector based speaker recognition systems," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.

[15] C. Greenberg, D. Banse, G. Doddington, D. Garcia-Romero, J. Godfrey, T. Kinnunen, A. Martin, A. McCree, M. Przybocki, and D. Reynolds, "The NIST 2014 speaker recognition i-vector machine learning challenge," in *Odyssey: The Speaker and Language Recognition Workshop*, 2014.

[16] S. Novoselov, T. Pekhovsky, and K. Simonchik, "STC speaker recognition system for the NIST i-vector challenge," in *Odyssey: The Speaker and Language Recognition Workshop*, 2014.

[17] E. Khoury, L. El Shafey, M. Ferras, and S. Marcel, "Hierarchical speaker clustering methods for the NIST i-vector challenge," in *Odyssey: The Speaker and Language Recognition Workshop*, 2014.

[18] G. Liu, C. Yu, A. Misra, N. Shokouhi, and J. Hansen, "Investigating state-of-the-art speaker verification in the case of unlabeled development data," in *Odyssey: The Speaker and Language Recognition Workshop*, 2014.

[19] B. Vesnicer, J. Zganec-Gros, S. Dobrisek, and V. Struc, "Incorporating duration information into i-vector-based speaker recognition systems," in *Odyssey: The Speaker and Language Recognition Workshop*, 2014.

[20] Y. Lei, 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.

[21] M. Omar and J. Pelecanos, "Training universal background models for speaker recognition," in *Odyssey: The Speaker and Language Recognition Workshop*, 2010.

[22] P. Kenny, V. Gupta, T. Stafylakis, P. Ouellet, and J. Alam, "Deep neural networks for extracting Baum-Welch statistics for speaker recognition," in *Odyssey: The Speaker and Language Recognition Workshop*, 2014.

[23] D. Garcia-Romero and A. McCree, "Subspace-constrained supervector PLDA for speaker verification," in *Interspeech*, 2013.

[24] N. Brümmer and D. Garcia-Romero, "Generative modelling for unsupervised score calibration," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.

[25] D. van Leeuwen and N. Brümmer, "The distribution of calibrated likelihood ratios," in *Interspeech*, 2013.

[26] "The NIST year 2010 Speaker Recognition Evaluation plan.," (Available at `http://www.itl.nist.gov/iad/mig/tests/sre/2010/NIST_SRE10_evalplan.r6.pdf`), 2010.

[27] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur, "Improving deep neural network acoustic models using generalized maxout networks," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.