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

UNSUPERVISED DOMAIN ADAPTATION FOR I-VECTOR SPEAKER RECOGNITION

Daniel Garcia-Romero[1], *Alan McCree*[1], *Stephen Shum*[2], *Niko Brümmer*[3], *and Carlos Vaquero*[4]

[1]Human Language Technology Center of Excellence, Johns Hopkins University, Baltimore, MD
[2]MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA
[3]AGNITIO Research, Somerset West, South Africa
[4]AGNITIO, Madrid, Spain

ABSTRACT

In this paper, we present a framework for unsupervised domain adaptation of PLDA based i-vector speaker recognition systems. Given an existing out-of-domain PLDA system, we use it to cluster unlabeled in-domain data, and then use this data to adapt the parameters of the PLDA system. We explore two versions of agglomerative hierarchical clustering that use the PLDA system. We also study two automatic ways to determine the number of clusters in the in-domain dataset. The proposed techniques are experimentally validated in the recently introduced *domain adaptation challenge*. This challenge provides a very useful setup to explore domain adaptation since it illustrates a significant performance gap between an in-domain and out-of-domain system. Using agglomerative hierarchical clustering with a stopping criterion based on unsupervised calibration we are able to recover 85% of this gap.

Index Terms— speaker recognition, unsupervised domain adaptation, clustering, unsupervised calibration, PLDA, i-vectors

1. INTRODUCTION

State-of-the-art 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. Specifically, it learns the within-class variability, that characterizes distortions, and the between-class variability, which characterizes speaker information. This knowledge is leveraged to obtain robustness against these distortions. To achieve this, the PLDA training set must contain multiple recordings of a speaker under different distortions (channel distortions, noise, reverberation). Typically, the PLDA systems used for NIST speaker recognition evaluations [8] 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 domain of interest might be too expensive or even unrealistic. In [9], a supervised adaptation approach is used to tackle this problem. An existing, resource-rich, out-of-domain system is bootstrapped and is able to produce good performance with smaller amounts of labeled in-domain data than a cold-start system would require. In this work, we rely on the same adaptation mechanisms explored in [9], but only require unlabeled in-domain data. This opens the door for using larger in-domain datasets since the cost of labeling is eliminated. Our approach uses the out-of-domain PLDA system to cluster a large in-domain dataset. This produces an estimate of the in-domain speaker labels that are subsequently used to adapt the parameters of the PLDA system to the new domain.

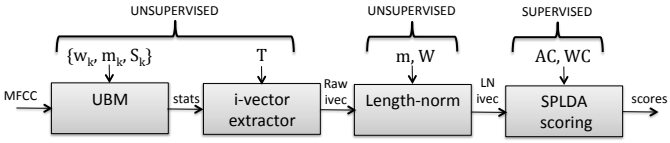

Fig. 1: Block diagram of speaker recognition system indicating which parameters are trained in supervised and unsupervised mode.

The rest of the paper is organized as follows: Section 2 describes the system architecture. Section 3 introduces the unsupervised adaptation framework. Section 4 describes our experimental setup and results. Finally, Section 5 provides the conclusions.

2. SPEAKER RECOGNITION SYSTEM

Figure 1 shows a block diagram of our state-of-the-art i-vector speaker recognition system. The first two blocks can be considered 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 determine whether an i-vector $\mathbf{x}_t$ belongs to speaker i or not. In the PLDA framework [6], this is equivalent to asking whether $\mathbf{x}_t$ was generated from the same latent speaker variable, $\mathbf{h}_i$, as the collection of J_i i-vectors from speaker i, $\mathcal{D}_i = \{\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iJ_i}\}$, or not. This corresponds to a model selection problem between two alternative generative models. Under the same-speaker hypothesis, $\mathcal{H}_s$, the generative model assumes that $\mathbf{h}_i = \mathbf{h}_t$. Under the different-speaker hypothesis, $\mathcal{H}_d$, the generative model assumes that $\mathbf{h}_i$ and $\mathbf{h}_t$ are independently drawn from a standard Gaussian. Note that this paradigm can also be applied to the case where, instead of a single test i-vector $\mathbf{x}_t$, we are presented with a collection of $\mathcal{D}_t = \{\mathbf{x}_{t1}, \ldots, \mathbf{x}_{tJ_t}\}$ test i-vectors from a unique speaker. The more likely the same-speaker hypothesis $\mathcal{H}_s$ is, the larger the score produced by the PLDA system. An efficient computation of this score is presented in [10].

On top of each block, Figure 1 shows the set of parameters that need to be trained. The terms supervised/unsupervised indicate 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, the largest improvement is obtained by adapting the PLDA parameters, which requires labeled data. This is not surprising, since the labeled data provides the information to learn a within-class variability matrix $\mathbf{\Lambda}$, that characterizes distortions, and an across-class variability matrix $\mathbf{\Gamma}$, that character-

izes speaker information. PLDA leverages this information to obtain robustness against the observed distortions. In the next section, we use clustering as a mechanism to obtain in-domain labeled data.

3. UNSUPERVISED DOMAIN ADAPTATION

In this section, we propose an approach 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. This produces an estimate of the in-domain speaker labels that are subsequently used to adapt the parameters of the PLDA system to the new domain. We now discuss the three key components of the approach: clustering technique, determination of number of clusters, and the adaptation mechanism.

3.1. Clustering

In [3], following a Bayesian approach, clustering was cast as a model selection problem. This requires the evaluation of the PLDA marginal likelihood (model evidence) for all possible partitions of a dataset. Unfortunately, an exhaustive search over all partitions is not scalable for large sets due to a combinatorial explosion (e.g. for a set of size $N = 10$, there are already 115,975 partitions). Instead, to reduce the search space, we propose a greedy search based on *agglomerative hierarchical clustering* (AHC). 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 now present two different metrics and two stopping criteria.

3.1.1. AHC i-vector averaging

As indicated in Section 2, the PLDA model provides a mechanism to compute the similarity between two collections/clusters of i-vectors $\mathcal{D}_i$, and $\mathcal{D}_t$:

$$s(\mathcal{D}_i, \mathcal{D}_t) = \log \frac{p(\mathcal{D}_i, \mathcal{D}_t|\mathcal{H}_s)}{p(\mathcal{D}_i, \mathcal{D}_t|\mathcal{H}_d)} = \log \frac{\mathcal{N}(\mathcal{D}_i, \mathcal{D}_t|\mathbf{0}, \mathbf{C}_{it})}{\mathcal{N}(\mathcal{D}_i|\mathbf{0}, \mathbf{C}_i)\mathcal{N}(\mathcal{D}_t|\mathbf{0}, \mathbf{C}_t)}, \quad (1)$$

which corresponds to the log likelihood ratio between the $\mathcal{H}_s$ and $\mathcal{H}_d$ hypothesis. This similarity function requires the evaluation of Gaussians with zero mean and covariance matrices, $\{\mathbf{C}_{it}, \mathbf{C}_i, \mathbf{C}_t\}$, that are functions of $\mathbf{\Gamma}$, and $\mathbf{\Lambda}$ (see [9, 10] for details). However, (1) assumes that all the i-vectors within a cluster, $\mathcal{D}$, are conditionally independent given their common latent speaker variable $\mathbf{h}$. This assumption is not consistent with most real data, and in practice, (1) is evaluated by averaging the i-vectors within each cluster $\mathcal{D}$ and pretending that there is only one i-vector per cluster. For this reason, we refer to this approach as AHC i-vector averaging. Moreover, every time we merge two clusters, we recompute the average based on all the individual (length-normalized) i-vectors in the two clusters and scale the averaged i-vector to unit length. Then, we recompute the similarity between the newly merged cluster and the remaining clusters. This updates the cluster similarity matrix that is used to select which two clusters to merge in the next step.

3.1.2. AHC score averaging

An alternative to AHC i-vector averaging is to use the out-of-domain PLDA system to compute a pairwise similarity matrix between all i-vectors [11]. 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 does not require using (1) every time we merge two clusters. Instead, it only requires averaging scores from the precomputed pairwise similarity matrix. Therefore, AHC score averaging is computationally cheaper than AHC i-vector averaging.

3.2. Determination of number of clusters

Let Θ be the space of all possible partitions of a dataset, and θ be one particular partition. Then, AHC defines a search path Θ_{AHC} such that, $\Theta_{AHC} \subset \Theta$ with $|\Theta_{AHC}| \ll |\Theta|$. In this section, we propose two criteria to select which partition, $\theta \in \Theta_{AHC}$, is optimal.

3.2.1. Evaluation of the marginal likelihood

As proposed in [3], we can use the PLDA marginal likelihood, $\mathcal{L}(\theta)$, as a selection mechanism. This is now feasible since we only need to evaluate the partitions given by the AHC search path Θ_{AHC}. That is, for a set with N i-vectors we need to compute $N - 1$ marginal likelihoods. However, due to the lack of a closed-form solution for $\mathcal{L}(\theta)$ (i.e. integrating over hidden variables and model parameters is not tractable), it is customary to only marginalize the hidden variables and use maximum likelihood (ML) plug-in estimates for the $\mathbf{\Gamma}$, and $\mathbf{\Lambda}$ matrices. This produces a reasonable approximation to $\mathcal{L}(\theta)$, but unfortunately, due to the ML point estimates of the parameters, it is not immune to overfitting. Alternatively, it is possible to use a variational approximation [5] to $\mathcal{L}(\theta)$ that is less prone to overfitting. However, the computational complexity of that solution is much larger. Therefore, due to scalability reasons, we only explore the plug-in approximation in this work.

3.2.2. Unsupervised calibration

Alternatively, to estimate the number of clusters, we can 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 from (1) so that we can use Bayesian decision theory to set a threshold analytically. However, to date, most calibration techniques make use of in-domain labeled data (which based on the premises of this work is not available to us). Fortunately, an unsupervised calibration approach, where only unlabeled in-domain scores are required, has been recently proposed in [12]. This approach uses a generative model of scores [13] 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 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, $\mathcal{H}_s$.

3.3. 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 $\mathbf{\Lambda}$. In [9], four adaptation approaches were studied and found 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

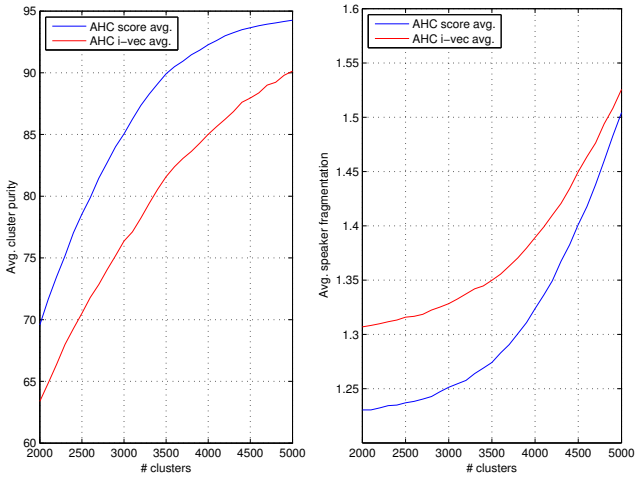

Fig. 2: Average purity and fragmentation for both AHC approaches when sweeping over the partitions in their search paths.

interpolate them with the out-of-domain parameters:

$$\begin{aligned} \mathbf{\Gamma}_{adapt} &= \alpha\, \mathbf{\Gamma}_{in} + (1-\alpha)\, \mathbf{\Gamma}_{out}, \\ \mathbf{\Lambda}_{adapt} &= \alpha\, \mathbf{\Lambda}_{in} + (1-\alpha)\, \mathbf{\Lambda}_{out}. \end{aligned} \tag{2}$$

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, the SRE10 telephone data [8] (condition 5 extended task) is used as enroll (single cut) and test sets. This evaluation set provides 7,169 target and 408,950 non-target trials. For parameter training, using Linguistic Data Consortium (LDC) telephone corpora, MIT-LL[1] has designed a *domain adaptation challenge* that exposes the effects of domain mismatch in recognition performance. Two datasets were defined for the challenge. 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. We use this domain adaptation challenge to explore unsupervised adaptation by ignoring the labels of the in-domain SRE set.

4.2. System setup

The 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) [8] with probability of target

[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/

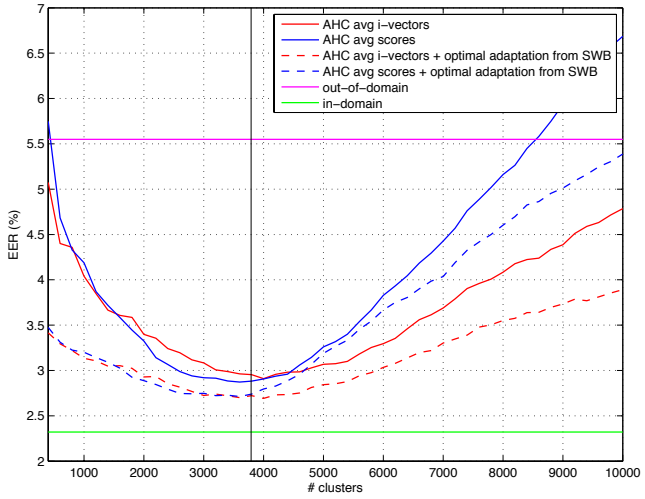

Fig. 3: EER over a set of partitions for out-of-domain, in-domain, and multiple adapted systems (see Section 4.3.2 for details). The black vertical line indicates the actual number of speakers.

trial set to either 10^{-2} or 10^{-3}. For the clustering, we report results in terms of average cluster purity and speaker fragmentation. The purity of a cluster is defined in terms of the dominant speaker in the cluster. The fragmentation corresponds to the number of clusters in which the cuts of a speaker are scattered apart.

4.3. Results

4.3.1. Performance gap

To focus on the effects of unsupervised PLDA adaptation, the in-domain and out-of-domain i-vectors are computed using a UBM and $\mathbf{T}$ trained on SWB. This setup is desirable since it does not require recomputing out-of-domain i-vectors. Moreover, in [9] it was shown that using an in-domain UBM and $\mathbf{T}$ does not produce significant gains in our domain adaptation challenge. Also, the whitening transformation of the length-normalization is based on the unlabeled in-domain SRE. As shown in Table 1, there is a considerable performance gap between a system with PLDA parameters trained on the out-of-domain SWB set or the in-domain SRE set. This validates the setup of the domain adaptation challenge and provides a significant gap to explore the effect of unsupervised adaptation.

4.3.2. AHC i-vector vs score averaging

Figure 2 compares the average purity and fragmentation of both AHC approaches as a function of the number of clusters. This is done by sweeping over a range of partitions of the in-domain SRE set given by the AHC search paths. It is clear that the clustering performance of the score averaging approach is better than i-vector averaging, both in purity and fragmentation. To understand how clustering performance translates into speaker recognition accuracy, Figure 3 shows the EER over an even larger set of partitions. The

Table 1: Performance as a function of in-domain SRE and out-of-domain SWB parameters. SPLDA system with rank 400.

UBM, $\mathbf{T}$	$\mathbf{W}$	$\mathbf{\Gamma}, \mathbf{\Lambda}$	DCF(10^{-3})	DCF(10^{-2})	EER(%)
SWB	SRE	SWB	0.627	0.425	5.55
SWB	SRE	SRE	**0.399**	**0.235**	**2.32**

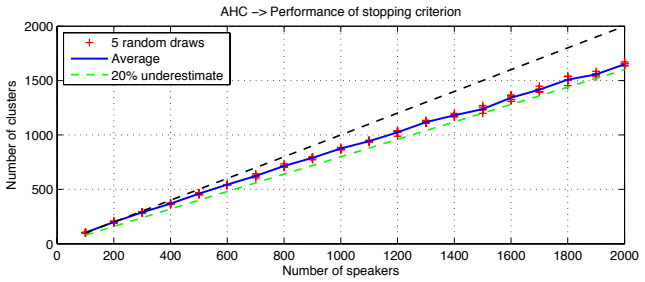

Fig. 4: Performance analysis of the unsupervised calibration stopping criterion for in-domain sets of different sizes. The dashed black line indicates perfect stopping.

horizontal lines correspond to out-of-domain and in-domain systems and show the performance gap. The vertical line marks the actual number of speakers in the SRE set. The solid curves are obtained by only using the in-domain SRE set and the estimated labels to train the PLDA parameters ($\alpha = 1$). The dashed curves correspond to the optimally adapted system that is allowed to use the out-of-domain parameters (i.e. optimal α). We make three key observations. First, the speaker recognition accuracy is not very sensitive to detecting the number of actual clusters. This sensitivity is even smaller when we adapt the PLDA parameters instead of just training them with the in-domain clustered data. Second, even though AHC score averaging outperforms i-vector averaging in terms of clustering performance, the recognition accuracies are very similar. That is, the clustering performance measures we are using are loosely correlated with recognition accuracy. Third, it is possible to recover 85% of the gap using AHC score averaging and domain adaptation.

Based on the better clustering performance, similar recognition accuracy, and lower computational cost than i-vector averaging, we select AHC score averaging as the recommended approach and focus on it from now on. Also, in the complementary work in [11], AHC score averaging is compared to two graph-based clustering methods and shown to outperform them.

4.3.3. Stopping criterion

Until now we have not dealt with the automatic selection of the number of clusters. However, our final solution requires an explicit stopping criterion. From the two approaches described in Section 3.2, the evaluation of the marginal likelihood is not successful. It grossly overestimates the number of clusters to be 9,300 (instead of 3,790). This can be attributed to the following: unrealistic conditional independence assumptions (see Section 3.1.1), ML plug-in approximation, and that the parameters are from out-of-domain.

We have validated the unsupervised calibration approach using subsets of SRE containing all speech cuts from 100 to 2000 speakers (5 random draws in each case), and in all cases the estimates were within the 20% of the actual number of speakers (see Figure 4). The estimated number of speakers for the full SRE set is 2,911, that is a relative error of 23%. This error is small enough to perform the adaptation task successfully, as shown in Figure 3.

4.3.4. Comparison with supervised adaptation

Putting together AHC score averaging, unsupervised calibration stopping criterion, and the interpolation of PLDA parameters, we obtain a fully functional unsupervised adaptation framework. In Figure 5 we compare this framework (dark blue) with a supervised

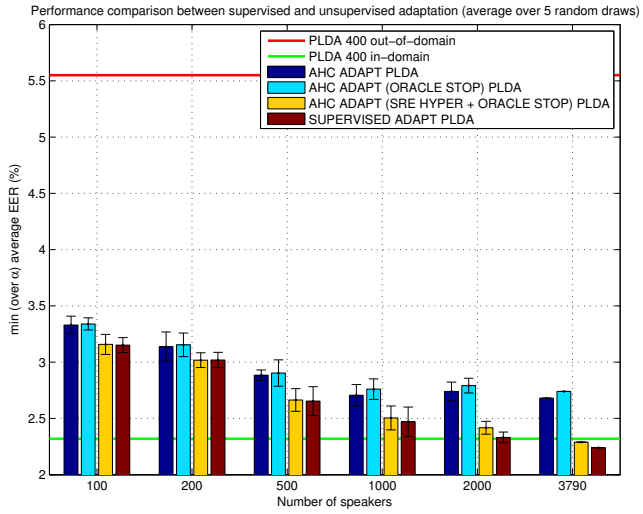

Fig. 5: Comparison of supervised and unsupervised domain adaptation using AHC score averaging and unsupervised calibration stopping for in-domain sets of different sizes. (See Section 4.3.4 for details).

adaptation (red) version that makes use of the SRE labels. We show the EER (averaged over 5 random draws) for subsets of SRE that contain different numbers of speakers (on average 10 cuts per speaker). The horizontal lines indicate the performance gap between the in-domain and out-of-domain systems. The supervised adaptation improves monotonically with the amount of data, but with diminishing returns. The unsupervised adaptation recovers 85% of the gap; however, it plateaus for datasets larger than a 1000 speakers.

To understand the cause of this behavior, we present two other results. In cyan, we show AHC with oracle stopping. In yellow, we additionally use the in-domain PLDA parameters to perform the clustering. Based on the results, we can rule out the stopping criterion as the cause. Instead, it is the use of out-of-domain parameters for clustering that causes the performance to plateau. This is a strong indication that multiple iterations of clustering and adaptation should help in closing the remaining 15% of the gap. At the moment, we have not explored this avenue yet, but it is our priority for future work. Also, we have used optimal values of α for our results; however, this is not a pressing issue since, as shown in [9], the performance sensitivity around optimal values of α is quite small.

5. CONCLUSION

In this paper, we presented an unsupervised domain adaptation framework that only requires an existing out-of-domain PLDA system and unlabeled in-domain data. The PLDA system was used to cluster the in-domain data and then adapted using the estimated labels. We explored two versions of AHC, and showed that AHC score averaging provides better clustering performance, similar recognition accuracy, and lower computational cost than i-vector averaging. We also studied two automatic ways to determine the number of clusters in the in-domain dataset. Stopping AHC based on unsupervised calibration was successful and provided estimates within 23% of the actual number of clusters. However, stopping AHC based on the marginal likelihood plug-in approximation was unsuccessful. All the experiments were conducted on the recently introduced domain adaptation challenge. This challenge provides a significant

performance gap between an in-domain and out-of-domain system. We recovered 85% of the gap using AHC score averaging with a stopping criterion based on unsupervised calibration.

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] "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.

[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 and A. McCree, "Subspace-constrained supervector PLDA for speaker verification," in *Interspeech*, 2013.

[11] S. Shum, D. Reynolds, D. Garcia-Romero, and A. McCree, "Unsupervised clustering approaches for domain adaptation in speaker recognition systems," *Odyssey*, 2014 (submitted).

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

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