SUPERVISED DOMAIN ADAPTATION FOR I-VECTOR BASED SPEAKER RECOGNITION

Daniel Garcia-Romero and Alan McCree

Human Language Technology Center of Excellence, Johns Hopkins University, Baltimore, MD

dgromero@jhu.edu, alan.mccree@jhu.edu

ABSTRACT

In this paper, we present a comprehensive study on supervised domain adaptation of PLDA based i-vector speaker recognition systems. After describing the system parameters subject to adaptation, we study the impact of their adaptation on recognition performance. Using the recently designed *domain adaptation challenge*, we observe that the adaptation of the PLDA parameters (i.e. across-class and within-class covariances) produces the largest gains. Nonetheless, length-normalization is also important; whereas using an in-domain UBM and $\mathbf{T}$ matrix is not crucial. For the PLDA adaptation, we compare four approaches. Three of them are proposed in this work, and a fourth one was previously published. Overall, the four techniques are successful at leveraging varying amounts of labeled in-domain data and their performance is quite similar. However, our approaches are less involved, and two of them are applicable to a larger class of models (low-rank across-class).

***Index Terms*—** speaker recognition, supervised domain adaptation, PLDA, i-vectors

1. INTRODUCTION

The state-of-the-art in speaker recognition is widely dominated by the use of i-vectors [1] modeled by variants of Probabilistic Linear Discriminant Analysis (PLDA) [2, 3, 4, 5, 6, 7]. An i-vector extractor (in combination with a UBM) is a data-driven front-end that maps temporal sequences of feature vectors (e.g. MFCCs) into a single point in a low-dimensional vector space. To achieve optimal performance, it is normally trained on tens of thousands of speech cuts from thousands of speakers in multiple sessions. Although training an i-vector extractor does not require a labeled dataset, it is customary to use datasets that contain speaker labels and other kinds of metadata (language, gender) to obtain a balanced training set.

PLDA provides a powerful data-driven mechanism to separate speaker information from other sources of undesired variability. Given a large collection of labeled data (speaker labels), PLDA learns the within-class variability, that characterizes distortions, and the between-class variability, which characterizes speaker information. Then, this knowledge is leveraged to obtain robustness against the observed distortions, when answering the question of whether a collection of i-vectors are from the same speaker or not (i.e. speaker recognition). 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 prohibitory expensive or even unrealistic. Therefore, a cold-start strategy for building systems in new domains is quite limiting. Alternatively, one could try to bootstrap an existing resource-rich out-of-domain system, and then require a smaller amount of in-domain data to adapt it. In [9], using Bayesian adaptation, the parameters of an out-of-domain system were successfully adapted to a domain with low resources. A fully Bayesian approach was used, and a variational approximation to the intractable posterior was computed using conjugate priors. However, due to computational complexity, the verification scores were computed using only point estimates of the parameters (expected values).

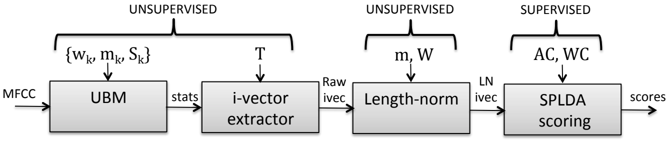

Fig. 1. Block diagram of speaker recognition system indicating which parameters are trained in supervised and unsupervised mode.

In this work, we present three alternative adaptation approaches that directly target point estimates of the parameters, and therefore, are more straightforward. Also, unlike the approach in [9], two of them work for models in which the across-class variability is not full rank. Moreover, our experimental setup allows for a resource-rich cold-start in-domain system that is used to asses the performance gap with respect to the out-of-domain system. This facilitates the study of how fast this gap gets closed in terms of the amount of in-domain data used. Additionally, we study the optimal amount of adaptation as a function of the amount of in-domain data.

Looking at Figure 1, we can see that there is opportunity to also adapt other system parameters. In [10], the authors explored the impact of: UBM, subspace used for i-vector extraction, length-normalization, score normalization, and calibration. Here, we also explore the first three, but, unlike in [10], we make a distinction between the parameters that require labeled or unlabeled data (speaker labels). Also, as part of the i-vector length-normalization, we explore the effects of the required whitening transformation [6]. Overall, the largest improvement is obtained by adapting PLDA and length-normalization parameters.

The rest of the paper is organized as follows: Section 2 describes the system architecture. Section 3 introduces the four adaptation techniques under study. 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. On top of each block we are showing the set of parameters that need to be trained. The terms supervised/unsupervised indicate if the parameters need to be trained using a dataset with speaker labels or not. The parameters that do

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

not require speaker labels are much easier to adapt since unlabeled in-domain data is much easier to acquire. In the following, we briefly describe each block.

2.1. UBM and i-vector extractor

The first two blocks of Figure 1 can be considered as a data-driven front-end that maps sequences of MFCCs into a low-dimensional vector [1]. This is accomplished by first training the parameters $\{w_k, \mathbf{m}_k, \mathbf{S}_k\}$ of a Gaussian mixture model, denoted as Universal Background model (UBM). The UBM is used to map the sequence of MFCCs into a single point in a high-dimensional space (typically around $100K$ dimensions). The i-vector extractor then uses factor analysis to perform dimensionality reduction to a low-dimensional subspace defined by the matrix $\mathbf{T}$. This subspace is learned in an unsupervised way on a large dataset (normally the same used for the UBM) and attempts to capture most of the speaker information.

2.2. Length-normalization

The third block of Figure 1 is a pre-processing stage that conditions the i-vectors so that they conform to the Gaussian modeling assumptions of the last block. Length-normalization 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 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. This produces length-normalized i-vectors.

2.3. Simplified PLDA (SPLDA)

2.3.1. Modeling

The SPLDA model [6] is a simplified version of PLDA introduced in [2], where, given a collection of J_i i-vectors from speaker i, $\mathcal{D}_i = \{\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iJ_i}\}$, we prescribe a generative model of the form:

$$\begin{bmatrix} \mathbf{x}_{i1} \\ \vdots \\ \mathbf{x}_{iJ_i} \end{bmatrix} = \begin{bmatrix} \mathbf{F} \\ \vdots \\ \mathbf{F} \end{bmatrix} \mathbf{h}_i + \begin{bmatrix} \boldsymbol{\epsilon}_{i1} \\ \vdots \\ \boldsymbol{\epsilon}_{iJ_i} \end{bmatrix}, \tag{1}$$

with the latent speaker variable $\mathbf{h}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and residual $\boldsymbol{\epsilon}_{ij} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Lambda})$ assumed independent. Moreover, the speaker subspace matrix $\mathbf{F} \in \mathbb{R}^{D \times P}$ is of rank $P < D$. Under these assumptions, an i-vector $x_{ij} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Gamma} + \mathbf{\Lambda})$ with across-class variability matrix $\mathbf{\Gamma} = \mathbf{F}\mathbf{F}^T$, and within-class variability matrix $\mathbf{\Lambda}$. Note that the rank of $\mathbf{\Gamma}$ corresponds to the number of columns of $\mathbf{F}$ (P). Also, the joint distribution of the i-vectors $\mathcal{D}_i$ is

$$p(\mathcal{D}_i|\mathbf{\Gamma}, \mathbf{\Lambda}) = \mathcal{N}(\mathcal{D}_i; \mathbf{0}, \widetilde{\mathbf{F}}\widetilde{\mathbf{F}}^T + \widetilde{\mathbf{\Lambda}}), \tag{2}$$

where, $\widetilde{\mathbf{F}}$ corresponds to the first matrix in the right hand side of (1), and $\widetilde{\mathbf{\Lambda}}$ is a block diagonal matrix with blocks set to $\mathbf{\Lambda}$.

The ability to control the rank of the across-class variability matrix $\mathbf{\Gamma}$ has significant impact in recognition performance. If we allow $P = D$, then, $\mathbf{\Gamma}$ becomes full-rank, and SPLDA becomes the two-covariance model (2-cov) introduced in [3] (i.e., no need to define $\mathbf{\Gamma}$ in terms of $\mathbf{F}$). Learning the SPLDA (or 2-cov) parameters $\mathbf{\Gamma}$, and $\mathbf{\Lambda}$, requires a large labeled dataset. Note that the 2-cov model requires a dataset with a number of speakers (and therefore total number of cuts) larger than the i-vector dimension D. For SPLDA, the number of speakers only needs to be larger than P, and the total number of cuts larger than D.

2.3.2. Scoring

The goal of the final block of Figure 1 is to determine whether an i-vector $\mathbf{x}_t$ belongs to speaker i or not. In the SPLDA framework, this is equivalent to asking whether $\mathbf{x}_t$ was generated from the same latent speaker variable, $\mathbf{h}_i$, as $\mathcal{D}_i$ 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. Since we are interested in a probabilistic answer, we compute a log likelihood ratio (LLR) between the two competing hypothesis:

$$\mathcal{R}(\mathcal{D}_i, \mathbf{x}_t) = \log \frac{p(\mathcal{D}_i, \mathbf{x}_t|\mathcal{H}_s)}{p(\mathcal{D}_i, \mathbf{x}_t|\mathcal{H}_d)} = \log \frac{p(\mathcal{D}_i, \mathbf{x}_t|\mathbf{\Gamma}, \mathbf{\Lambda})}{p(\mathcal{D}_i|\mathbf{\Gamma}, \mathbf{\Lambda})p(\mathbf{x}_t|\mathbf{\Gamma}, \mathbf{\Lambda})}. \tag{3}$$

An efficient computation of this LLR can be found in [11].

3. ADAPTATION APPROACHES

In this section, we present four approaches to adapt the across-class and within-class covariances ($\mathbf{\Gamma}$, $\mathbf{\Lambda}$) of a system trained on out-of-domain data, using a new set of in-domain data. The first two focus on to the 2-cov model, and the last two also apply to SPLDA.

3.1. Fully Bayesian adaptation

In [9], following a Bayesian treatment, the 2-cov model parameters are treated as random variables and a variational approximation to the intractable posterior is computed using conjugate priors[1]. The intractability of the model stems from the fact that the across-class $\mathbf{\Gamma}$ and within-class $\mathbf{\Lambda}$ matrices are not proportional to each other. The out-of-domain data is used to define a conjugate prior and, given in-domain data, an approximate posterior distribution of the 2-cov model parameters is obtained. This results in a posterior distribution that combines the information of the out-of-domain prior with the information provided by the in-domain data. The strength of the prior is controlled by ignoring the actual counts of the out-of-domain set an instead prescribing new hyper-parameters to the prior (a more detailed explanation is given in the next section). Although the adaptation mechanism provides an approximate posterior distribution, due to computational cost, the LLRs are based on point estimates (expected values) of the 2-cov model parameters.

3.2. Approximate MAP adaptation

For the 2-cov model, when there are a lot of cuts per speaker, the uncertainty of the latent speaker identity variable $\boldsymbol{\mu}$ is very small and we can treat it as an observed variable (i.e. the sample mean of $\mathcal{D}_i$). Under this assumption, the estimation of the covariance matrices decouples (see the graphical model in Figure 1 of [9]). Also, since we are only using point estimates for the computation of LLRs, MAP point estimates of the covariances are the quantities of interest. Using an inverse Wishart distribution $\mathcal{IW}(\mathbf{\Sigma}|\mathbf{S}_{out}^{-1}, \nu_{out})$ (i.e., conjugate prior) with hyper-parameters based on the out-of-domain covariance, we obtain an adapted covariance as:

$$\mathbf{\Sigma}_{map} = \frac{\mathbf{S}_{in} + \mathbf{S}_{out}}{N_{in} + N_{out}} = \alpha\, \mathbf{\Sigma}_{in} + (1 - \alpha)\, \mathbf{\Sigma}_{out} \tag{4}$$

[1] The authors thank Jesus Villalba for sharing a Matlab implementation of VB adaptation

with $N_{out} = \nu_{out} + D + 1$, and $\alpha = \frac{N_{in}}{N_{out}+N_{in}}$. Therefore, N_{out} represents pseudo-counts and controls the strength of the out-of-domain prior. The corresponding degrees of freedom ν_{out} can be obtained from it. For our experiments, we parameterize the strength of the prior through α, which can be mapped back to ν_{out}. This implies that we are ignoring the actual number of counts of the out-of-domain set. We can particularize the generic equation (4) to the within-class $\mathbf{\Lambda}$ and across-class $\mathbf{\Gamma}$ covariances using the corresponding in-domain scatter matrix $\mathbf{S}_{in}$ and setting N_{in} to either the number of cuts (for $\mathbf{\Lambda}$) or the number of speakers (for $\mathbf{\Gamma}$). Note that we are not modeling the mean since we use centered i-vectors due to length-normalization. Also, this approach does not require multiple iterations over the data, as is the case for all of the other approaches that we explore.

3.3. Weighted Likelihood

In the previous sections we have presented adaptation approaches for the 2-cov model. However, when the across-class matrix $\mathbf{\Gamma}$ is not full-rank, an inverse Wishart prior cannot be used. For the SPLDA model, a fully Bayesian approach could be used by placing a Normal prior per row of $\mathbf{F}$. However, since due to computational cost, we are using point estimates of the parameters to compute the LLRs, we instead propose to use maximum-likelihood (ML) point estimates of the SPLDA parameters based on a weighted log-likelihood objective

$$\mathcal{L}(\mathbf{\Gamma}, \mathbf{\Lambda}) = \alpha\, \mathcal{L}_{in}(\mathbf{\Gamma}, \mathbf{\Lambda}) + (1-\alpha)\, \mathcal{L}_{out}(\mathbf{\Gamma}, \mathbf{\Lambda}), \tag{5}$$

where $\mathcal{L}_{\bullet}(\mathbf{\Gamma}, \mathbf{\Lambda}) = \frac{1}{N_{\bullet}} \sum_{i=1}^{M_{\bullet}} \log p(\mathcal{D}_i | \mathbf{\Gamma}, \mathbf{\Lambda})$, $N_{\bullet}$ refers to cuts, and $M_{\bullet}$ to speakers, for either the in-domain or out-of-domain sets. Note that this approach requires the i-vectors of the out-of-domain set and not just the parameters learned from it. However, this is not a big issue due to the small dimensionality of the i-vectors. The SPLDA parameters are then learned using a modified EM algorithm to maximize (5). In particular, letting $k = 1$ index the in-domain set, and $k = 2$ the out-of-domain set, in the E-step, we compute the posterior mean $\langle \mathbf{h}_{ik} \rangle$ and correlation $\langle \mathbf{h}_{ik} \mathbf{h}_{ik}^T \rangle$ of the hidden speaker variables using the previous values of $\mathbf{F}_{old}$ and $\mathbf{\Lambda}_{old}$. Then, the M-step results in

$$\begin{aligned} \mathbf{F} &= \Big(\sum_{k=1}^{2} \dot{\alpha}_k \sum_{i=1}^{M_k} \bar{\mathbf{x}}_{ik} \langle \mathbf{h}_{ik}^T \rangle \Big) \Big(\sum_{k=1}^{2} \dot{\alpha}_k \sum_{i=1}^{M_k} N_{ik} \langle \mathbf{h}_{ik} \mathbf{h}_{ik}^T \rangle \Big)^{-1}, \\ \mathbf{\Lambda} &= \Big(\sum_{k=1}^{2} \dot{\alpha}_k \sum_{ij} \mathbf{x}_{ijk} \mathbf{x}_{ijk}^T \Big) - \mathbf{F} \Big(\sum_{k=1}^{2} \dot{\alpha}_k \sum_{i=1}^{M_k} \bar{\mathbf{x}}_{ik} \langle \mathbf{h}_{ik}^T \rangle \Big)^T, \end{aligned} \tag{6}$$

where $\bar{\mathbf{x}}_{ik} = \sum_j \mathbf{x}_{ijk}$, $\dot{\alpha}_1 = \frac{\alpha}{N_{in}}$, and $\dot{\alpha}_2 = \frac{1-\alpha}{N_{out}}$. After each M-step, we perform a minimum divergence step [12] to accelerate the convergence.

3.4. SPLDA parameter interpolation

When the in-domain set contains enough data (i.e., total number of cuts is larger than the i-vector dimension D), a good practical approximation to the weighted likelihood is to use weighted SPLDA parameters. That is, use the standard EM algorithm twice to obtain in-domain and out-of-domain PLDA parameters and then interpolate between them. Note that this approach does not require keeping the out-of-domain i-vectors. In practice, it is possible to use an in-domain dataset with less than D cuts to estimate SPLDA parameters if we regularize $\mathbf{\Lambda}_{in}$ to make it positive definite. Note however, that the rank of $\mathbf{\Gamma}_{in}$ will be equal to the number of speakers.

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(%)
1	SWB	SWB	SWB	0.682	0.485	6.92
2	SWB	SRE	SWB	0.627	0.425	5.55
3	SWB	SRE	SRE	**0.399**	**0.235**	**2.32**
4	SRE	SRE	SRE	0.425	0.255	2.43

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 [2] has designed a *domain adaptation challenge* that exposes the effects of data mismatch in recognition performance. Two datasets were defined for the challenge: the in-domain SRE set comprises telephone calls from all speakers taken from the SRE 04, 05, 06, and 08 collections. The out-of-domain SWB set comprises telephone calls from all speakers taken from the Switchboard-I and Switchboard-II (all phases) corpora. The SRE set consist of 3,790 speakers (male and female) and 36,470 speech cuts. The distribution of number of cuts per speaker is not homogeneous and has a mean of 9.6 and standard deviation of 7.7. On average each speaker made calls from 2.8 different phone numbers. The SWB set consist of 3,114 speakers (male and female) and 33,039 speech cuts. The distribution of number of cuts per speaker is not homogeneous and has a mean of 10.6 and standard deviation of 7.9. On average each speaker made calls from 3.8 different phone numbers. Although the statistics of both datasets are quite similar, the SRE set matches the SRE10 evaluation set better than SWB. This is mostly attributed to the evolution of telephone systems, as most of the SWB data was collected in the 90s and the SRE collection is more recent.

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 SPLDA. We report performance in terms of equal error rate (EER) and/or normalized minimum detection cost function (DCF) [8] with probability of target trial set to either 10^{-2} or 10^{-3}.

4.3. Results

4.3.1. Performance gap

As shown by the first and last rows of Table 1, there is a considerable gap in performance between a system trained on the out-of-domain SWB set (row 1), and a system trained on the matched in-domain SRE set (last row). For the EER, the performance gap is about 3x. This validates the setup of the *domain adaptation challenge* and provide a significant gap to explore the effect of adaptation approaches.

[2] The authors thank MIT-LL for designing 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/

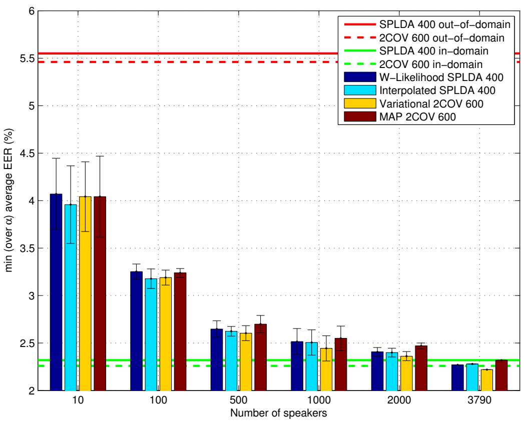

Fig. 2. Comparison of 4 adaptation approaches in terms of the amount of in-domain speakers. Two approaches are applied to the 2-cov model and the other two to SPLDA with 400 dimensional speaker space. Results averaged over 5 random draws.

4.3.2. *UBM, $\mathbf{T}$, and length-normalization*

In this section we explore the impact of adapting the parameters that do not require speaker labels. Comparing rows 3 and 4 of Table 1 we can evaluate the impact of training the UBM and $\mathbf{T}$ on the in-domain SRE set. The main observation is that the impact is small (this is consistent with [10]). We speculate that using SWB produces slightly better results due to the disjoint datasets used for i-vector extraction and PLDA training, but at the moment this is still uncertain. Comparing rows 1 and 2, we can see the effects of length normalization. In the first row, the centering and whitening of the out-of-domain SWB and the evaluation data are based on SWB statistics. For the second row, the centering of SWB is based on SWB statistics, but the evaluation data is centered using SRE statistics. Also, the whitening of both sets is based on SRE statistics. This strategy of dataset-dependent centering and common whitening (based on in-domain statistics) produces the best results.

4.3.3. *Adaptation of $\mathbf{\Gamma}$ and $\mathbf{\Lambda}$*

To decouple the effects of adapting $\mathbf{\Gamma}$ and $\mathbf{\Lambda}$ from the choices of UBM, $\mathbf{T}$, and length-normalization, we use the configurations in rows 2 and 3 of Table 1 for the remaining experiments. We refer to 2 as the out-of-domain system (i.e. using all SWB to train $\mathbf{\Gamma}$ and $\mathbf{\Lambda}$), and to 3 as the in-domain system. The red horizontal lines of Figure 2 are the performance of the out-of-domain systems (SPLDA is solid and 2-cov is dashed), and represent the starting point. The green lines are the performance of the in-domain systems, and represent the target performance. We can see that using a full-dimensional speaker space (600) provides slightly better results. This is quite unusual, since in other evaluation setups [6], using a low-rank $\mathbf{\Gamma}$ ($\sim$200) was better. However, during our SRE12 participation, we observed that for gender-independent PLDA, 400 dimensions was better than 200. Nonetheless, to exercise the two adaptation techniques that apply to low-rank $\mathbf{\Gamma}$, we also present results for SPLDA.

Figure 2 also shows the performance of the 4 adaptation techniques for different amounts of in-domain speakers, and optimal adaptation α. Although it is possible to use different amounts of adaptation for each covariance matrix, we tied them together to facilitate the analysis. The results are presented in terms of EER and are averaged over 5 random draws from the entire in-domain SRE.

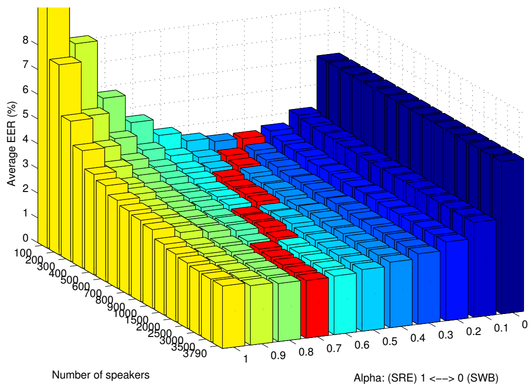

Fig. 3. Grid search to show the optimal value of α as a function of the number of speakers (on average 10 cuts per speaker). Results averaged over 5 random draws.

We can see that the four techniques perform similarly and leverage the increasing amount of in-domain data. However, we observe a clear diminishing return as the amount of data grows. For example, using 10 speakers we recover about 45% of the gap, and we need around 1000 speakers to recover 90%. For the techniques specific to the 2-cov model (variational and MAP), we observe that the simpler MAP approximation initially gets close to the variational approach, but underperforms as the amount of in-domain adaptation data increases. For the SPLDA system, the interpolated SPLDA approach is as good as weighted likelihood without requiring access to the out-of-domain i-vectors. Note that, except for MAP, when all the in-domain data was provided, the performance was slightly better than that of the target performance. This implies that, even when using the complete set of in-domain data, $\alpha = 1$ was not optimal and the out-of-domain SWB data was helpful. In the next section we explore system performance with respect to α.

4.3.4. *Amount of adaptation*

Figure 3 shows the average EER (over 5 random draws) of the 2-cov model adapted with the parameter interpolation method. The performance is presented as a function of the amount of in-domain data, as well as the adaptation amount α. As expected, the more in-domain data, the larger α. Also, the error surface is quite smooth as a function of α. This indicates that, for this adaptation challenge, the selection of the optimal α is not hard, and therefore, a small held-out set should be sufficient to get a good estimate. The same behavior was observed for the other adaptation techniques.

5. CONCLUSION

In this paper, we presented a comparative study of four supervised domain adaptation techniques. Three of them were proposed in this work, and a fourth one was previously published. The four approaches were experimentally validated on the recently designed *domain adaptation challenge*. Overall, the four techniques were successful at leveraging varying amounts of labeled in-domain data. The performance was quite similar across techniques. We observed that among all the parameters of a state-of-the-art i-vector recognizer, the adaptation of the across-class and within-class covariances (which required labeled data) produce the largest gains. Nonetheless, length-normalization was also important; whereas using an in-domain UBM and $\mathbf{T}$ matrix was not crucial.

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] J. Villalba and E. Lleida, “Bayesian adaptation of PLDA based speaker recognition to domains with scarce development data,” in *Odyssey: The Speaker and Language Recognition Workshop*, Singapore, 2012.

[10] C. Vaquero, “Dataset shift in PLDA based speaker verification,” in *Odyssey: The Speaker and Language Recognition Workshop*, Singapore, 2012.

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

[12] N. Brümmer, “EM for probabilistic LDA,” (Available at `https://sites.google.com/site/nikobrummer/`), February 2010.