DISCRIMINATIVE MULTI-DOMAIN PLDA FOR SPEAKER VERIFICATION

Alexey Sholokhov[1,2] *, Tomi Kinnunen*[1] *, Sandro Cumani*[3]

[1]University of Eastern Finland
[2]ITMO University
[3]Politecnico di Torino

sholok@cs.uef.fi, tkinnu@cs.uef.fi, sandro.cumani@polito.it

ABSTRACT

Domain mismatch occurs when data from application-specific target domain is related to, but cannot be viewed as iid samples from the source domain used for training speaker models. Another problem occurs when several training datasets are available but their domains differ. In this case training on simply merged subsets can lead to suboptimal performance. Existing approaches to cope with these problems employ generative modeling and consist of several separate stages such as training and adaptation. In this work we explore a discriminative approach which naturally incorporates both scenarios in a principled way. To this end, we develop a method that can learn across multiple domains by extending discriminative probabilistic linear discriminant analysis (PLDA) according to multi-task learning paradigm. Our results on the recent JHU Domain Adaptation Challenge (DAC) dataset demonstrate that the proposed multi-task PLDA decreases equal error rate (EER) of the PLDA without domain compensation by more than 35% relative and performs comparable to another competitive domain compensation technique.

Index Terms: multi-task learning, discriminative training, probabilistic linear discriminant analysis, speaker recognition

1. INTRODUCTION

For more than a decade, speaker recognition researchers have enjoyed the availability of large quantities of speech data provided by the National Institute of Standards and Technologies through their speaker recognition evaluations (NIST SREs). The NIST SRE data contains carefully annotated speaker labels along with the audio files and other metadata, enabling accurate training of system hyperparameters such as the universal background model (UBM) [1], i-vector extractor [2] and the probabilistic linear discriminant analysis (PLDA) [3].

Unfortunately, problems arise in many real world scenarios when the application domain data differs from the NIST SRE data. In particular, it was found that PLDA model trained on part of the Switchboard (SWB) corpus [4] experienced considerable performance degradation when tested on data from NIST SRE 2010 [5]. This is an example of *domain mismatch*: distributions of testing and training data are different [6]. It arises in cases when a lot of *source* data is available but data from the *target* domain is scarce or expensive to collect. This results in poor performance of models learned on training data because they do not generalize well to the testing domain. To reduce the effect of domain mismatch, *domain adaptation* [6] can be applied. It combines data from the source domain with the limited amount of in-domain data to improve the predictive performance in the target domain.

The domain mismatch problem was recently brought to the attention of speaker recognition community [7]. Domain mismatch has several causes, including mismatch in data collection caused by, for instance, different speech acquisition methods and languages. [8] illustrates that it leads to multi-modality in distribution of i-vectors. It was found that domain mismatch severely affects training of the PLDA parameters, while less impact on performance is due to data mismatch in training the UBM or the i-vector extractor. This observation brought the focus of research to the back-end part of the speaker verification system.

Accordingly, to address the problem of domain mismatch, three different categories of approaches have been explored. The first approach, based on generative modeling [5],[9], assumes that small number of *labeled* utterances from the target domain is available, and is therefore termed *supervised domain adaptation*. The experiments in [5] indicate that, even for just 10 speakers sampled from in-domain data, the methods filled up 45% relative performance gap between matched and mismatched training for the experimental setup of the Domain Adaptation Challenge [7].

The second approach assumes that *unlabeled* data from the target domain is available, that is, with missing speaker labels. This *unsupervised domain adaptation* is more appealing since human-based labeling is time consuming and expensive. It also enables the use of potentially much larger in-domain datasets. A straightforward way to utilize large unlabeled datasets is to apply clustering to obtain approximate pseudo speaker labels. The authors of [10], [11] explored various clustering algorithms and found that even imperfect clustering can provide recognition accuracy close to that obtained with oracle speaker labels. Assuming the clustering step is separated from the domain adaptation step, we can apply the same methods as in supervised domain adaptation: first we obtain the pseudo speaker labels, then apply a suitable supervised adaptation method. Otherwise, the possible solution is to employ Bayesian approaches or treating PLDA parameters as random variables and transfering information from source data encoded in posterior distributions to infer the missing speaker labels on target data [12].

The above studies assumed that large amount of labeled out-of-domain data is available. At the same time, lack or absence of in-domain data does not allow retraining all models on this data. Sometimes, however, despite the large amounts, training data can consist of several subsets, each from a different domain. Accordingly, the third approach assumes that there is *no data* from the target domain but relies solely on partitioning the heterogeneous out-

This work was financially supported by the Academy of Finland (proj. no. 253120 and 283256) and the Government of the Russian Federation (Grant 074-U01). A share of the computational resources for this work was provided by HPC@POLITO (www.hpc.polito.it)

of-domain data into a number of subsets with slightly different distributional properties. Heterogeneity of the training data allows estimating between-dataset variability and compensating for it so as to improve robustness against unseen conditions, i.e. the new domain. Techniques belonging to this third category include, for instance, *source normalization* (SN) [13] and *inter-dataset variability compensation* (IDVC) [14].

To cope with the aforementioned problems, we consider the goal of learning a classifier for speaker verification when multiple training datasets (from non-target domains, and possibly also from the target domain) are available, assuming a distribution mismatch across these sets. To this end, we propose a new training method for the PLDA model that takes advantage from so-called *multi-task learning* (MTL) [15]. In general, MTL aims at improving generalization performance by joint learning over several related learning problems or *tasks*. Specifically, inspired by [16], we extend discriminative training strategy for PLDA [17], [18] to cope with multiple domains. In our case, the tasks differ only by data sampled from different domains. Then, we aim at training a model that is more robust to domain mismatch archieved by using multiple source datasets and parameter sharing.

2. DISCRIMINATIVE MULTI-TASK PLDA

2.1. Multi-Task Learning

The MTL framework [15] considers the problem of *joint learning* over several learning tasks to improve generalization. To this end, all the learning tasks usually share the same feature space and their marginal distributions are assumed to be different, but not too much, which results in a set of related learning problems. Such parallel learning enables sharing of information across the tasks which can be beneficial for all the tasks and can lead to enhanced predictive performance. If the task is the same across all the domains, as in our case, MTL is similar to domain adaptation [6]. However, MTL aims at improving performance across *all* the tasks, while domain adaptation, introducing asymmetry, aims at improving performance only for a single target domain.

The MTL models differ in their assumptions made about relatedness of tasks. The simplest idea is *parameter sharing* across the tasks. It can be particularly useful when each task has limited data, as the tied parameters can be estimated more accurately. In practice, one can consider a set of task-specific classifiers trained with certain constraints on their parameters. These constraints can be set, for instance, by placing a *hierarchical Bayesian prior* on the parameters [19] or by regularizing the empirical risk [16].

2.2. I-Vector Based Speaker Verification

In the past few years, the i-vector approach [2] has become the *de facto* standard in speaker verification. It provides a convenient way to represent variable-length feature vector sequences as a low-dimensional vector with the help of Gaussian mixture model (GMM) and factor analysis techniques. In particular, the model states that,

$$\boldsymbol{\mu} = \boldsymbol{\mu}_0 + \boldsymbol{T}\boldsymbol{x},$$

where $\boldsymbol{\mu}$ is *mean supervector* of GMM that corresponds to one utterance, $\boldsymbol{\mu}_0$ is the global mean and $\boldsymbol{T}$ is a low-rank rectangular matrix whose columns span the so-called *total variability space* and $\boldsymbol{x}$ is a low-dimensional random vector which has a standard Gaussian prior distribution. The maximum *a posteriori* (MAP) estimate of $\boldsymbol{x}$ is known as an *i-vector*.

The parameters of the i-vector extractor are trained in an unsupervised fashion on a large corpus attempting to capture all possible variations in the training data. The problem of speaker verification then reduces to comparing whether a given pair of i-vectors (one for enrolment and the other one for test) originates from same or different speakers. To this end, the PLDA model, has turned out as one of the most robust techniques [20]. The *simplified* PLDA (SPLDA) [21] models a collection of n-dimensional i-vectors $\{\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, ..., \boldsymbol{x}_{iJ}\}$ for speaker i as follows:

$$\boldsymbol{x}_{ij} = \boldsymbol{m} + \boldsymbol{V}\boldsymbol{y}_i + \boldsymbol{e}_{ij},$$

where $\boldsymbol{m}$ is the speaker-independent dataset mean, $\boldsymbol{y}_i$ is a standard normal-distributed latent identity variable that represents a particular speaker, $\boldsymbol{e}_{ij}$ is a residual term distributed as $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ and $\boldsymbol{V}$ is a rectangular speaker subspace matrix. If $\mathrm{rank}(\boldsymbol{V}) = n$, SPLDA is equivalent to the 2-covariance model (2-cov) [22].

Given a trial (pair of i-vectors), the speaker verification score s is computed as the log-likelihood ratio (LLR) between two hypotheses: same speaker (H_s) and different speaker (H_d):

$$s(\boldsymbol{x}_1, \boldsymbol{x}_2) = \log \frac{p(\boldsymbol{x}_1, \boldsymbol{x}_2|H_s)}{p(\boldsymbol{x}_1, \boldsymbol{x}_2|H_d)}. \tag{1}$$

For the PLDA model, this expression has a closed form solution. In particular, it can be shown [17, 18] that it is a quadratic form that can be written as a dot product of the vector of weights and a function of i-vector pairs, thereby resulting in a linear classifier of the form:

$$\begin{aligned} s(\boldsymbol{x}_1, \boldsymbol{x}_2) &= \boldsymbol{x}_1^\top P \boldsymbol{x}_2 + \boldsymbol{x}_2^\top P \boldsymbol{x}_1 + \\ &\boldsymbol{x}_1^\top Q \boldsymbol{x}_1 + \boldsymbol{x}_2^\top Q \boldsymbol{x}_2 + \boldsymbol{c}^\top (\boldsymbol{x}_1 + \boldsymbol{x}_2) + d \\ &= \begin{bmatrix} \mathrm{vec}(\boldsymbol{P}) \\ \mathrm{vec}(\boldsymbol{Q}) \\ \boldsymbol{c} \\ d \end{bmatrix}^\top \begin{bmatrix} \mathrm{vec}(\boldsymbol{x}_1\boldsymbol{x}_2^\top + \boldsymbol{x}_2\boldsymbol{x}_1^\top) \\ \mathrm{vec}(\boldsymbol{x}_1\boldsymbol{x}_1^\top + \boldsymbol{x}_2\boldsymbol{x}_2^\top) \\ \boldsymbol{x}_1 + \boldsymbol{x}_2 \\ 1 \end{bmatrix} \\ &= \boldsymbol{w}^\top f(\boldsymbol{x}_1, \boldsymbol{x}_2), \end{aligned} \tag{2}$$

where $\mathrm{vec}(\cdot)$ is the column stacking operator while $\boldsymbol{P}, \boldsymbol{Q}, \boldsymbol{c}$ and d can be expressed in terms of the PLDA parameters $\boldsymbol{m}, \boldsymbol{V}$ and $\boldsymbol{\Sigma}$. That is, $s(\boldsymbol{x}_1, \boldsymbol{x}_2)$ can be written as linear classification rule in the space of vectors $f(\boldsymbol{x}_1, \boldsymbol{x}_2)$ representing trials. This form allows direct learning of the decision function without need to explicitly model data distribution. The weight vector $\boldsymbol{w}$ can be trained similar to standard logistic regression or support vector machine (SVM), leading to *discriminative* PLDA [17]. There are many possible choices of loss functions [23], including both convex and non-convex ones. In our study we focus only on the *hinge loss* (aka standard SVM objective) resulting in the model referred to as *pairwise* SVM (PSVM) [18]. We adopt PSVM to the case of multiple training datasets but other variants of discriminative PLDA could also be extended in the same way.

2.3. Multi-task Support Vector Machines

We assume a training set of T subsets that share the same feature space but represent different domains i.e. their distributions differ. Accordingly, the data available for training is,

$$\mathcal{D} = \{(\boldsymbol{X}_1, \boldsymbol{Y}_1), (\boldsymbol{X}_2, \boldsymbol{Y}_2), ..., (\boldsymbol{X}_T, \boldsymbol{Y}_T)\}$$

where $\boldsymbol{X}_t$ is a matrix of the feature vectors and $\boldsymbol{Y}_t$ the corresponding class labels with $\boldsymbol{x}_{it} \in \mathbb{R}^n$, $y_{it} \in \{-1, 1\}$ for i running over the t-th subset. We follow the multi-task learning extension of SVM introduced in [16]. In this approach, all the tasks share the same set of labels to be predicted by T classifiers, one for each task. That is,

the prediction function is allowed to change from task to task. Relatedness of tasks is expressed by assuming that the parameters of the t-th classifier is decomposed as

$$\boldsymbol{w}_t = \boldsymbol{w}_0 + \boldsymbol{v}_t, \tag{3}$$

Intuitively, for *similar* tasks, all the task-specific models $\boldsymbol{w}_t$ should be close to each other. In terms of (3), $\boldsymbol{w}_t$ should not deviate much from the shared "average" model $\boldsymbol{w}_0$, i.e. $\boldsymbol{v}_t$ are "small". Otherwise, for *weakly related* tasks, we expect $\boldsymbol{w}_0$ to be "small". In other words, $\boldsymbol{w}_0$ carries common information across the tasks while $\boldsymbol{w}_t$ adapts to a particular task. The authors of [16] formulated a convex optimization problem including task-coupling constraint specified through regularization. The goal is to estimate all $\boldsymbol{v}_t$ and $\boldsymbol{w}_0$ jointly by minimizing the following cost:

$$\min_{\boldsymbol{w}_0,\{\boldsymbol{v}_t\}} \Big\{ \lambda_0 \|\boldsymbol{w}_0\|^2 + \sum_t \lambda_t \|\boldsymbol{v}_t\|^2 + \sum_t \sum_i \max(0, 1 - y_{it} \boldsymbol{x}_{it}^\top (\boldsymbol{w}_0 + \boldsymbol{v}_t)) \Big\}, \tag{4}$$

where $y_{it} \in \boldsymbol{Y}_t$ is the label of i-th training vector $x_{it} \in \boldsymbol{X}_t$ from subset t. The latter term in the sum is the empirical risk evaluated on all subsets which is just the sum of the hinge loss functions. In fact, we have sum of T objective functions of the conventional SVM learning problem [24] with parameters decomposed in a specific way (3). The regularization parameters, λ_t, determine the tradeoff between closeness of each task-specific model to the mean parameter vector and optimality of these models.

For an alternative formulation, let us define a feature map and new parameter vector as follows :

$$\Phi(\boldsymbol{x}) = \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{0} \\ \vdots \\ \boldsymbol{x} \\ \vdots \end{bmatrix} \begin{matrix} 0 \\ 1 \\ \\ t \\ \end{matrix}, \quad \boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{w}_0 \\ \boldsymbol{v}_1 \\ \vdots \\ \boldsymbol{v}_t \\ \vdots \end{bmatrix}, \quad \boldsymbol{x} \in \boldsymbol{X}_t,$$

where $\boldsymbol{0} \in \mathbb{R}^n$ is the zero vector. Here Φ maps the original input vector $\boldsymbol{x}$ to the $n(T+1)$-dimensional space by augmenting it with zeros and its own copy placed at the $(t+1)$-th position. Then the problem 4 can be equivalently re-written in standard (single-task SVM) form:

$$\min_{\boldsymbol{\theta}} \Big\{ \boldsymbol{\theta}^\top \boldsymbol{H} \boldsymbol{\theta} + \sum_j \max(0, 1 - y_j \boldsymbol{\theta}^\top \Phi(\boldsymbol{x}_j)) \Big\},$$

where $\boldsymbol{H} = \mathrm{diag}([\lambda_0 \lambda_1 \dots \lambda_T]^\top \otimes \boldsymbol{1})$ is a diagonal matrix and $\boldsymbol{1} \in \mathbb{R}^n$ is the vector of ones. Here $\otimes$ denotes Kronecker product. Looking at form of the feature map Φ, one can see a parallel with the domain adaptation method proposed in [25].

When $\lambda_0 \to \infty$, the problem (4) reduces to T separate problems as there is no coupling between the tasks. In the other extreme case, $\lambda_t \to \infty$, optimization results in a single classifier because all the task-specific parameters are forced to be the same. Thus, it is important to decide what is the appropriate balance between learning shared and individual parameters.

2.4. Proposed Method: Multi-Task PSVM

Noting that the similarity score (1) can be written in the form of (2), it is straightforward to define multi-task formulation of the PSVM just by replacing feature vectors $\boldsymbol{x}$ in (4) by $f(\boldsymbol{x}_1, \boldsymbol{x}_2)$ from (2). For each subset we have the same pairwise classification problem: for each pair of i-vectors we must decide whether their speaker identities match or differ. For $T = 1$, this reduces to the single-task pairwise SVM optimization problem [18].

It should be noted that there is a difference between the standard multi-task SVM [16] and the multi-task PSVM (mtPSVM) proposed here. Since the input to PSVM is a pair of vectors, it may happen that they originate from different domains. In our formulation, we assume that there is no cross-domain trials and no overlap in speaker labels across subsets (tasks). At the same time it would be reasonable to expect that having sessions of the same speaker in different domains can considerably improve recognition performance because it allows more accurate estimation of how within-speaker distribution changes across domains. However, due to the pairwise formulation, the proposed model would require special extension to handle this case.

Once training is completed, $\boldsymbol{w}_0$ can be used to construct a classifier for unseen conditions, as the shared parameters capture domain-independent data structure, thereby making classification more robust under domain mismatch. We can obtain even better performance if some amount of labeled data from the target domain is available. If one of the training datasets, say the i-th one, is drawn from the target domain, then using $\boldsymbol{w}_i = \boldsymbol{w}_0 + \boldsymbol{v}_i$ as the final classifier can help lowering error rates through the use of task-specific parameters. We adopt this strategy for *supervised* domain adaptation.

Choosing among large-scale SVM solvers, we selected general framework of *bundle methods for risk minimization* (BMRM) [26] to solve the optimization problem (4). In particular, we use *optimized cutting plane algorithm* (OCAS) described in [27], which is a special case of BMRM customized for SVM problems. BMRM is an iterative procedure which requires only sub-gradients and loss function to be computed. Despite the fact that the training set consist of all possible pairs of i-vectors, efficient computation of scores and sub-gradient of the objective function is possible [18]. Therefore, no explicit expansion of all trials is necessary. For multi-task PSVM, subgradients of the empirical risk in (4) can be computed as follows:

$$\nabla \boldsymbol{v}_t = \begin{bmatrix} 2\mathrm{vec}\left(\boldsymbol{X}_t \boldsymbol{G}_t \boldsymbol{X}_t^\top\right) \\ 2\mathrm{vec}\left([\boldsymbol{X}_t \circ [\boldsymbol{1}_{nm}^\top \boldsymbol{G}_t]] \boldsymbol{X}_t^\top\right) \\ 2[\boldsymbol{X}_t \circ [\boldsymbol{1}_{nm}^\top \boldsymbol{G}_t] \boldsymbol{1}_m] \\ \boldsymbol{1}_m^\top \boldsymbol{G}_t \boldsymbol{1}_m \end{bmatrix}$$

$$\nabla \boldsymbol{w}_0 = \sum_t \nabla \boldsymbol{v}_t,$$

where $\boldsymbol{G}_t$ is the square matrix of derivatives of the hinge loss function with respect to the scores of all possible pairs in subset t, $\boldsymbol{X}_t = [\boldsymbol{x}_{1t} \boldsymbol{x}_{2t} \cdots \boldsymbol{x}_{mt}]$ is the matrix of m stacked i-vectors of dimension n belonging to this subset. Finally, $\boldsymbol{1}_{nm}$ and $\boldsymbol{1}_m$ denote, respectively, matrix and vector of ones of size $(n \times m)$ and $(m \times 1)$ and $\circ$ stands for element-wise matrix product.

3. EXPERIMENTS

We follow JHU 2013 *Domain Adaptation Challenge* (DAC) experimental setup detailed in [7]. As out-of-domain it defines the Switchboard (SWB) set consisting of 36,470 utterances. The in-domain set includes data from SRE 04, 05, 06, and 08 (telephone calls) – in total, 3,114 speakers (male and female) and 33,039 utterances.

We evaluate the proposed method under two multi-domain learning scenarios: (1) when only out-of-domain data is available *i.e.* domain robust training and (2) when there is some amount of labeled in-domain data *i.e.* supervised domain adaptation. We evaluate performance using male, female and pooled trials from SRE10

telephone data. In total, this evaluation protocol consists of 7,169 target and 408,950 non-target trials.

All the training and test utterances are represented as 600-dimensional i-vectors, obtained using gender-independent UBM and i-vector extractor trained on SWB [5]. Following the commonly used preprocessing steps, we center the i-vectors using mean computed on development data, then apply whitening transform or within-class covariance normalization (WCCN) [28] and length-normalization (LN) [21]. We use equal error rate (EER) and minimum detection cost function (minDCF) [29] with probability of target trial set to 10^{-3} to measure speaker verification system performance: $\mathrm{minDCF} = \min(P_{\mathrm{M}} + 999P_{\mathrm{FA}})$, where P_{M} and P_{FA}, respectively, are the miss and the false alarm probabilities at some operating point and the minimum is computed over the all operating points.

We used the open-source MATLAB[1] implementation of the SPLDA and 2-covariance models [30].

3.1. Domain Robust Training

In this scenario, the goal is to capture dataset-independent structure in data by generalizing across domains. In particular, we train PSVM so that it would be robust to unseen domains. Following [14], we divide SWB data into 6 gender-independent subsets according to the provided LDC labels. As the final classifier, we use the shared parameters $\boldsymbol{w}_0$ as detailed in Section 2.4.

Table 1 reports the results for three techniques: the proposed mtPSVM method, the IDVC method from [14] and the SPLDA. IDVC was applied to all the hyperparameters of a 2-cov model. Subspace dimension for mean was equal to 5. Results for SPLDA demonstrate the baseline performance without domain mismatch compensation (in this case, whitening instead of WCCN showed better results). The best result for IDVC, shown in Table 1, was obtained using subspace dimensions of 60 for the within-class and 0 for the between-class covariance matrices.

Table 1. Results for domain robustness (EER, % and minDCF). For mtPSVM the best results among different configurations are shown.

Model	All		Male		Female	
	EER	DCF	EER	DCF	EER	DCF
SPLDA	6.88	0.679	5.98	0.624	7.95	0.713
IDVC	**3.08**	**0.493**	**2.48**	**0.355**	**3.10**	**0.505**
mtPSVM	4.18	0.644	2.79	0.545	3.12	0.551

Discriminative approach increases domain robustness of the PLDA classifier substantially but performs worse than IDVC. This probably follows from the fact that IDVC aims at directly compensating for cross-domain variability, but the proposed approach encodes this goal through much weaker assumptions – form of parameter decomposition (3), which does not necessarily lead to the most robust model $\boldsymbol{w}_0$.

3.2. Supervised Domain Adaptation

In the supervised domain adaptation setting, we transfer the knowledge obtained on multiple source domains to a new target domain. This is done by including the available in-domain data as one of the tasks and using the corresponding task-specific parameters $\boldsymbol{w}_0 + \boldsymbol{v}_{\mathrm{in-domain}}$ at test time. In this experiment, we compare the performance of "domain-independent" model $\boldsymbol{w}_0$ and the model fitted to a target domain $\boldsymbol{w}_{\mathrm{SRE}}$. Additionally, we compare the proposed approach to IDVC [14] and parameter interpolation of 2-cov models. For IDVC, in-domain data were included as one of the subsets to find the subspace with the largest domain variability. After removing this subspace, full SWB was used to train the 2-cov model for final evaluation. Another method (2-cov interpolation) consists of training two 2-cov models on in-domain and out-of-domain data, and interpolating their parameters with a weight depending on amount of in-domain data, similar to [5].

It is important noting, that both whitening and WCCN transforms were trained on SWB data. This is in contrast with [5], where SRE data has been used to estimate these transforms. We also found that applying the combination of LN and WCCN twice is helpful for mtPSVM.

Table 2 shows the speaker verification system performance for different amounts of in-domain speakers. Interestingly, for small amount of in-domain data, $\boldsymbol{w}_0$ still performs better than $\boldsymbol{w}_{\mathrm{SRE}}$. We hypothesize that the large number of model parameters leads to overfitting the target task for small amount of adaptation data. The results also indicate that small sizes of in-domain subset for IDVC can decrease its performance because of inaccurate parameter estimation. It should be added that for 1000 in-domain speakers mtPSVM performs better than the 2-cov model trained on the full SRE dataset, which achieves an EER of 2.58%

To set the regularization parameters, one can use heuristic rule from [18] to set the λ_t for each subset separately, then λ_0 can be set to be a few times (2-5) less than their average. We found that this strategy leads to satisfactory accuracy, close to the one that uses the values of λ tuned over evaluation set. We leave the problem of finding better values for these regularizers as future work.

Table 2. Results (EER,%) for supervised domain adaptation with different amounts of in-domain (SRE) data (pooled genders).

Model	Number of in-domain speakers				
	0	10	100	500	1000
$\boldsymbol{w}_0$	4.18	4.26	4.12	3.34	2.77
$\boldsymbol{w}_0 + \boldsymbol{v}_{\mathrm{SRE}}$	—	4.33	4.14	3.08	**2.41**
2-cov interp	7.02	4.70	3.64	3.14	2.91
IDVC	**3.08**	**3.67**	**3.28**	**2.69**	2.59

4. CONCLUSION

We adopted discriminative multi-task learning formulation to the PLDA model for the problem of speaker verification in the presence of domain mismatch. Our results indicate that the new method leads to substantial improvement over the case of no domain robust training/adaptation, similar to the findings reported for conventional PLDA in [5, 14]. Moreover, as the number of in-domain speakers increase, mtPSVM accuracy improves as expected. Unfortunately, the proposed method requires a large number of in-domain utterances to be effective. The primary reason, as we suspect, is the high dimensionality of PSVM model which calls either for a larger out-of-domain dataset or different regularization strategies such as low-rank or orthogonal solutions. Another potential extension is to reformulate the current model for simultaneous learning of domain-specific classifiers and a domain-invariant subspace.

We provide an open-source package, containing the codes for training mtPSVM model[2].

[1] Source code can be found at `https://sites.google.com/site/fastplda/`

[2] `http://cs.uef.fi/~sholok/mtPSVM.tar.gz`

5. REFERENCES

[1] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[2] Najim Dehak, Patrick Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[3] Simon J. D. Prince and James H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *IEEE 11th International Conference on Computer Vision, ICCV 2007, Rio de Janeiro, Brazil, October 14-20, 2007*, 2007, pp. 1–8.

[4] "The lingusitic data consortium (LDC) catalog," `http://catalog.ldc.upenn.edu`, Accessed: 2015-03-12.

[5] Daniel Garcia-Romero and Alan McCree, "Supervised domain adaptation for i-vector based speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014*, 2014, pp. 4047–4051.

[6] Sinno Jialin Pan and Qiang Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, 2010.

[7] "JHU 2013 speaker recognition workshop," `http://www.clsp.jhu.edu/workshops/archive/ws13-summer-workshop/groups/spk-13/`, Accessed: 2015-03-12.

[8] Ondrej Glembek, Jeff Z. Ma, Pavel Matejka, Bing Zhang, Oldrich Plchot, Lukas Burget, and Spyros Matsoukas, "Domain adaptation via within-class covariance correction in i-vector based speaker recognition systems," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014*, 2014, pp. 4032–4036.

[9] Jesús A. Villalba and Eduardo Lleida, "Bayesian adaptation of PLDA based speaker recognition to domains with scarce development data," in *Odyssey 2012: The Speaker and Language Recognition Workshop, Singapore, June 25-28, 2012*, 2012, pp. 47–54.

[10] Daniel Garcia-Romero, Alan McCree, Stephen Shum, Niko Brummer, and Carlos Vaquero, "Unsupervised domain adaptation for i-vector speaker recognition," in *Odyssey, The Speaker and Language Recognition Workshop*, 2014.

[11] Stephen Shum, Douglas Reynolds, Daniel Garcia-Romero, and Alan McCree, "Unsupervised clustering approaches for domain adaptation in speaker recognition systems," in *Odyssey, The Speaker and Language Recognition Workshop*, 2014.

[12] Jesús A. Villalba and Eduardo Lleida, "Unsupervised adaptation of PLDA by using variational Bayes methods," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014*, 2014, pp. 744–748.

[13] Mitchell McLaren and David van Leeuwen, "Source-normalized LDA for robust speaker recognition using i-vectors from multiple speech sources," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 20, no. 3, pp. 755–766, 2012.

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

[15] Rich Caruana, "Multitask learning," in *Machine Learning*, 1997, pp. 41–75.

[16] Theodoros Evgeniou and Massimiliano Pontil, "Regularized multi–task learning," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004*, 2004, pp. 109–117.

[17] Lukas Burget, Oldrich Plchot, Sandro Cumani, Ondrej Glembek, Pavel Matejka, and Niko Brümmer, "Discriminatively trained probabilistic linear discriminant analysis for speaker verification," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2011, May 22-27, 2011, Prague Congress Center, Prague, Czech Republic*, 2011, pp. 4832–4835.

[18] Sandro Cumani, Niko Brummer, Lukás Burget, Pietro Laface, Oldrich Plchot, and Vasileios Vasilakakis, "Pairwise discriminative speaker verification in the i-vector space," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 21, no. 6, pp. 1217–1227, 2013.

[19] Jenny Rose Finkel and Christopher D. Manning, "Hierarchical bayesian domain adaptation," in *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, May 31 - June 5, 2009, Boulder, Colorado, USA*, 2009, pp. 602–610.

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

[21] Daniel Garcia-Romero and Carol Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011*, 2011, pp. 249–252.

[22] Niko Brümmer and Edward de Villiers, "The speaker partitioning problem," in *Odyssey 2010: The Speaker and Language Recognition Workshop, Brno, Czech Republic, June 28 - July 1, 2010*, 2010, p. 34.

[23] Johan Rohdin, Sangeeta Biswas, and Koichi Shinoda, "Discriminative PLDA training with application-specific loss functions for speaker verification," in *Odyssey, The Speaker and Language Recognition Workshop*, 2014.

[24] Chris Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[25] Hal Daumé III, "Frustratingly easy domain adaptation," in *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*, 2007.

[26] Choon Hui Teo, S. V. N. Vishwanathan, Alex J. Smola, and Quoc V. Le, "Bundle methods for regularized risk minimization," *Journal of Machine Learning Research*, vol. 11, pp. 311–365, 2010.

[27] Vojtech Franc and Sören Sonnenburg, "Optimized cutting plane algorithm for large-scale risk minimization," *Journal of Machine Learning Research*, vol. 10, pp. 2157–2192, 2009.

[28] Andrew O. Hatch, Sachin S. Kajarekar, and Andreas Stolcke, "Within-class covariance normalization for svm-based speaker recognition," in *INTERSPEECH 2006 - ICSLP, Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, September 17-21, 2006*, 2006.

[29] "The NIST year 2010 speaker recognition evaluation plan," `http://www.itl.nist.gov/iad/mig/tests/sre/2010/NIST_SRE10_evalplan.r6.pdf`, Accessed: 2015-03-12.

[30] Aleksandr Sizov, K.-A Lee, and Tomi Kinnunen, "Unifying probabilistic linear discriminant analysis variants in biometric authentication," in *Structural, Syntactic, and Statistical Pattern Recognition - Joint IAPR International Workshop, S+SSPR 2014, Joensuu, Finland, August 20-22, 2014. Proceedings*, 2014, pp. 464–475.