Improving the Effectiveness of Speaker Verification Domain Adaptation With Inadequate In-Domain Data

Bengt J. Borgström[1], Elliot Singer[1], Douglas Reynolds[1], and Omid Sadjadi[2,†]

[1]MIT Lincoln Laboratory, Lexington, MA
[2]National Institute of Standards and Technology, Gaithersburg, MD

jonas.borgstrom@ll.mit.edu.edu, dar@ll.mit.edu, es@ll.mit.edu, omid.sadjadi@nist.gov

Abstract

This paper addresses speaker verification domain adaptation with inadequate in-domain data. Specifically, we explore the cases where in-domain data sets do not include speaker labels, contain speakers with few samples, or contain speakers with low channel diversity. Existing domain adaptation methods are reviewed, and their shortcomings are discussed. We derive an unsupervised version of fully Bayesian adaptation which reduces the reliance on rich in-domain data. When applied to domain adaptation with inadequate in-domain data, the proposed approach yields competitive results when the samples per speaker are reduced, and outperforms existing supervised methods when the channel diversity is low, even without requiring speaker labels. These results are validated on the SRE16, which uses a highly inadequate in-domain data set.

Index Terms: speaker verification, unsupervised domain adaptation, Bayesian adaptation.

1. Introduction

In recent years, i-vectors have become the dominant representation of speech signals for speaker verification, since they allow the mapping of utterances of arbitrary duration to a single low-dimensional vector [1]. Due to their low dimensionality, sophisticated techniques can be used to model i-vectors and generate verification scores. One such scoring method is probabilistic linear discriminant analysis (PLDA) [2], which provides a statistical tool for emphasizing speaker information while compensating for undesired sources of variability.

Statistical approaches to i-vector speaker verification often require explicit modeling of across-class and within-class variabilities. Commonly, these sources of variability are modeled as Gaussian distributions, and their means and covariance matrices are trained on large sets of speech data. Typically, such data sets may include tens of thousands of utterances, with thousands of individual labeled speakers.

It has been widely shown that the performance of speaker verification systems degrades when facing unseen types of data [3, 4, 5, 6, 7]. This is caused in part by a mismatch between the out-of-domain data used to train system hyperparameters, and the in-domain data encountered during enrollment and testing. However, such performance degradation can be mitigated via domain adaptation of system parameters using in-domain development data. Typically, in-domain sets are inadequate in some respect, and we focus on three such properties. First, speaker labels may not be included, making supervised approaches unusable. Secondly, the data may include speakers with few samples. Finally, the data may include speakers with low channel diversity, where samples may exhibit similar channel types. In the case of inadequate in-domain data, thoughtful strategies must be employed to successfully leverage this data.

In [3], the authors treat across-class and within-class covariance matrices as random variables, and propose using maximum a posteriori (MAP) estimates of these parameters conditioned on the in-domain data. In this work, point estimate approximations are used for speaker means for the sake of computational efficiency. However, such methods may suffer if the in-domain set includes individual speakers with few samples or low channel diversity. In [5], the authors present a fully Bayesian framework to domain adaptation. This approach explicitly models the uncertainty due to speakers with few samples, and is therefore less sensitive to such data. However, it still does not address the low channel diversity problem.

In this paper, we derive an unsupervised version of fully Bayesian domain adaptation. We assume in-domain data samples to be independent, which implies that each sample was produced by a unique speaker, leading to a reduced reliance on a rich in-domain data set. It follows that the proposed method does not require speaker labels for the in-domain data. When applied to domain adaptation with inadequate in-domain data, the proposed technique provides competitive results for data with few samples per speaker, and outperforms existing supervised methods for data with low channel diversity.

This paper is organized as follows. In Sec. 2, we present a statistical framework for i-vector domain adaptation. Sec. 3 includes a discussion of existing techniques for supervised domain adaptation. In Sec. 4 we derive unsupervised Bayesian adaptation. Experimental results are presented in Sec. 5, and conclusions are provided in Sec. 6.

2. Statistical Framework

In this paper, we assume the additive noise model for i-vectors:

$$\mathbf{x}_{mn} = \mathbf{y}_m + \mathbf{c}_{mn}, \tag{1}$$

where $\mathbf{x}_{mn}$ denotes the n^{th} sample from the m^{th} speaker, $\mathbf{y}_m$ is the latent speaker component, and $\mathbf{c}_{mn}$ is the channel component. Speaker components are assumed Gaussian i.i.d.:

$$p\left(\mathbf{y}_m\right) = \mathcal{N}\left(\mathbf{y}_m; \boldsymbol{\mu}, \boldsymbol{\Sigma}_a\right), \tag{2}$$

and are collectively denoted by $\mathcal{Y}$. Channel components are

[†]Contractor (Systems Plus, Inc)

The editor coordinating the review of this manuscript and approving it for publication from NIST was Greg Sanders.

This work is sponsored by the Department of Defense under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

assumed Gaussian i.i.d.:

$$p(\mathbf{c}_{mn}) = \mathcal{N}(\mathbf{c}_{mn}; \mathbf{0}, \boldsymbol{\Sigma}_w). \quad (3)$$

Domain adaptation involves using a limited set of in-domain data to adapt hyperparameters trained in a more resource-rich domain. Specifically, in the context of i-vector-based speaker verification, domain adaptation refers to estimation of the set $\{\boldsymbol{\mu}, \boldsymbol{\Sigma}_a, \boldsymbol{\Sigma}_w\}$ based on a set of in-domain samples, $\mathcal{X}$, and out-of-domain hyperparameter estimates $\{\boldsymbol{\mu}^{\boldsymbol{out}}, \boldsymbol{\Sigma}_a^{out}, \boldsymbol{\Sigma}_w^{out}\}$. Let $\mathcal{X}$ contain N_T samples from M speakers. We seek a probabilistic solution to domain adaptation, and so we encode knowledge of the out-of-domain data in prior distributions, which can be designed to reflect the out-of-domain parameter estimates. We model $\boldsymbol{\Sigma}_a$ and $\boldsymbol{\Sigma}_w$ with inverse-Wishart distributions, as:

$$p(\boldsymbol{\Sigma}_a) = \mathcal{IW}\left(\boldsymbol{\Sigma}_a; \nu_a \boldsymbol{\Sigma}_a^{out}, \nu_a\right), \quad (4)$$

and:

$$p(\boldsymbol{\Sigma}_w) = \mathcal{IW}\left(\boldsymbol{\Sigma}_w; \nu_w \boldsymbol{\Sigma}_w^{out}, \nu_w\right). \quad (5)$$

In this way, $\boldsymbol{\Sigma}_a^{out}$ and $\boldsymbol{\Sigma}_w^{out}$ approximate the modes of the respective prior distributions, and $\nu_a$and ν_w reflect the certainty of the priors, as discussed in [3].

3. Supervised Domain Adaptation

3.1. Parameter Estimation

A popular approach for the estimation of across-class and within-class covariance matrices is to use point estimates for speaker components:

$$\tilde{\mathbf{y}}_m = \frac{1}{N_m} \sum_{n=1}^{N_m} \mathbf{x}_{mn}. \quad (6)$$

where N_m denotes the number of samples provided by the m^{th} speaker. Maximum likelihood parameter estimation leads to:

$$\boldsymbol{\Sigma}_a = \frac{1}{N_T} \sum_{m=1}^{M} N_m (\tilde{\mathbf{y}}_m - \tilde{\boldsymbol{\mu}})(\tilde{\mathbf{y}}_m - \tilde{\boldsymbol{\mu}})^T, \quad (7)$$

$$\boldsymbol{\Sigma}_w = \frac{1}{N_T} \sum_{m=1}^{M} \sum_{n=1}^{N_m} (\mathbf{x}_{mn} - \tilde{\mathbf{y}}_m)(\mathbf{x}_{mn} - \tilde{\mathbf{y}}_m)^T, \quad (8)$$

where $\tilde{\boldsymbol{\mu}}$ is the sample mean of $\mathcal{X}$. Note that these equations can be manipulated slightly to normalize the effect of individual speakers. It can be observed from (6)-(8) that modeling speaker components plays a vital role in parameter estimation, and inaccurate speaker component estimates may lead to poor estimates for $\boldsymbol{\Sigma}_a$ and $\boldsymbol{\Sigma}_w$.

The approximation in (6) is valid when each speaker provides a large number of samples and includes rich channel diversity, but may otherwise result in inaccurate estimates. If (1) is substituted into (6), the speaker estimate can be expressed as:

$$\tilde{\mathbf{y}}_m = \mathbf{y}_m + \frac{1}{N_m} \sum_{n=1}^{N_m} \mathbf{c}_{mn}, \quad (9)$$

If channel components are truly distributed according to (3), it is clear from (9) that the speaker component estimate is unbiased with a variance of $\boldsymbol{\Sigma}_w/N_m$. Thus, the estimate is highly variable for speakers with few samples, but will approach the underlying speaker component, $\mathbf{y}_m$, as the number of samples increases. Additionally, if a speaker provides samples from similar channels, so that the channel components are not zero-mean, the estimate from (9) will not approach the underlying speaker component regardless of the value of N_m.

3.2. Adaptation Using Speaker Point Estimates

In [3], the authors find approximated MAP estimates of the in-domain parameters assuming (6):

$$\begin{aligned}\boldsymbol{\Sigma}_a =& \frac{\alpha}{N_T} \sum_{m=1}^{M} N_m (\tilde{\mathbf{y}}_m - \tilde{\boldsymbol{\mu}})(\tilde{\mathbf{y}}_m - \tilde{\boldsymbol{\mu}})^T \\ &+ (1-\alpha)\boldsymbol{\Sigma}_a^{out} \\ &+ \alpha(1-\alpha)\left(\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}^{\boldsymbol{out}}\right)\left(\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}^{\boldsymbol{out}}\right)^T, \end{aligned} \quad (10)$$

$$\begin{aligned}\boldsymbol{\Sigma}_w =& \frac{\alpha}{N_T} \sum_{m=1}^{M} \sum_{n=1}^{N_m} (\mathbf{x}_{mn} - \tilde{\mathbf{y}}_m)(\mathbf{x}_{mn} - \tilde{\mathbf{y}}_m)^T \\ &+ (1-\alpha)\boldsymbol{\Sigma}_w^{out}. \end{aligned} \quad (11)$$

where α is a function of the number of samples in the in-domain set, and approaches 1 as the data set grows. However, α can also be tuned manually to control the emphasis placed on the in-domain data during adaptation. The technique was applied to the 2013 *domain adaptation challenge* (DAC13) [8], where the in-domain set was limited with respect to the number of speakers, and showed graceful degradation as the speaker diversity was made increasingly scarce. However, such domain adaptation methods can suffer for in-domain sets containing limited numbers of samples per speaker, or containing speakers with low channel diversity.

3.3. Bayesian Adaptation

In [5], the authors propose Bayesian adaptation of hyperparameters, where speaker components are assumed to be latent random variables, as in (2). The joint posterior distribution of the set $\{\boldsymbol{\mu}, \boldsymbol{\Sigma}_a, \boldsymbol{\Sigma}_w, \mathcal{Y}\}$ is approximated by a factorized form using variational Bayes (VB) [9]:

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}_a, \boldsymbol{\Sigma}_w, \mathcal{Y} | \mathcal{X}) \approx q(\mathcal{Y})\, q(\boldsymbol{\mu}, \boldsymbol{\Sigma}_a)\, q(\boldsymbol{\Sigma}_w). \quad (12)$$

MAP parameter estimates are then found for each factoring distribution, and used during scoring. According to variational Bayes, the optimal factoring distributions are found iteratively. The update equation for the distribution of speaker components is given by [9]:

$$\log q(\mathcal{Y}) = E_{\boldsymbol{\mu}, \boldsymbol{\Sigma}_a, \boldsymbol{\Sigma}_w}\{\log p(\mathcal{X}, \mathcal{Y}, \boldsymbol{\mu}, \boldsymbol{\Sigma}_a, \boldsymbol{\Sigma}_w)\} + \text{const}, \quad (13)$$

with analogous expressions for the other hidden variables. Central to the variational Bayes approach is the total data log-likelihood, which for our statistical framework is given by:

$$\begin{aligned}\log p(\mathcal{X}, \mathcal{Y}, \boldsymbol{\mu}, \boldsymbol{\Sigma}_a, \boldsymbol{\Sigma}_w) =& \log p(\mathcal{X} | \mathcal{Y}, \boldsymbol{\Sigma}_w) \\ &+ \log p(\mathcal{Y} | \boldsymbol{\mu}, \boldsymbol{\Sigma}_a) \\ &+ \log p(\boldsymbol{\mu}, \boldsymbol{\Sigma}_a) + \log p(\boldsymbol{\Sigma}_w), \end{aligned} \quad (14)$$

where:

$$\log p(\mathcal{X} | \mathcal{Y}, \boldsymbol{\Sigma}_w) = \sum_{m=1}^{M} \sum_{n=1}^{N_m} \mathcal{N}(\mathbf{x}_{mn}; \mathbf{y}_m, \boldsymbol{\Sigma}_w). \quad (15)$$

Applying the variational Bayes method, and using the statistical framework described in Sec. 2, the optimal factors are deter-

mined via the iterative equations (see [5] for details):

$$\boldsymbol{\mu} = \alpha\bar{\mathbf{y}} + (1-\alpha)\,\boldsymbol{\mu}^{\boldsymbol{out}}, \tag{16}$$

$$\boldsymbol{\Sigma}_a = \alpha\left(\frac{1}{M}\sum_{m=1}^{M}\langle \mathbf{y}_m\mathbf{y}_m^T\rangle - \bar{\mathbf{y}}\bar{\mathbf{y}}^T\right) + (1-\alpha)\,\boldsymbol{\Sigma}_a^{out} + \alpha\,(1-\alpha)\left(\bar{\mathbf{y}} - \boldsymbol{\mu}^{\boldsymbol{out}}\right)\left(\bar{\mathbf{y}} - \boldsymbol{\mu}^{\boldsymbol{out}}\right)^T, \tag{17}$$

$$\boldsymbol{\Sigma}_w = \frac{\alpha}{N_T}\sum_{m=1}^{M}\sum_{n=1}^{N_m}\Big(\mathbf{x}_{mn}\mathbf{x}_{mn}^T - \langle \mathbf{y}_m\rangle\mathbf{x}_{mn}^T - \mathbf{x}_{mn}\langle \mathbf{y}_m\rangle^T + \langle \mathbf{y}_m\mathbf{y}_m^T\rangle\Big) + (1-\alpha)\,\boldsymbol{\Sigma}_w^{out}, \tag{18}$$

where:

$$\bar{\mathbf{y}} = \frac{1}{M}\sum_{m=1}^{M}\langle \mathbf{y}_m\rangle, \tag{19}$$

$$\langle \mathbf{y}_m\rangle = \boldsymbol{\Sigma}_a\left(\boldsymbol{\Sigma}_a + \frac{1}{N_m}\boldsymbol{\Sigma}_w\right)^{-1}\frac{1}{N_m}\sum_{n=1}^{N_m}\mathbf{x}_{mn} + \boldsymbol{\Sigma}_w\left(N_m\boldsymbol{\Sigma}_a + \boldsymbol{\Sigma}_w\right)^{-1}\boldsymbol{\mu}, \tag{20}$$

$$\langle \mathbf{y}_m\mathbf{y}_m^T\rangle = \boldsymbol{\Sigma}_w\left(N_m\boldsymbol{\Sigma}_a + \boldsymbol{\Sigma}_w\right)^{-1}\boldsymbol{\Sigma}_a + \langle \mathbf{y}_m\rangle\langle \mathbf{y}_m\rangle^T. \tag{21}$$

As previously mentioned, Bayesian adaptation models speaker means as posterior distributions, taking into account the uncertainty resulting from parameter estimation with finite data, and is therefore less sensitive to in-domain data sets with few samples per speaker.

However, Bayesian adaptation may still suffer from in-domain data with low channel diversity. If (1) is substituted into (20), the mean of the posterior distribution of $\mathbf{y}_m$ is:

$$\langle \mathbf{y}_m\rangle = \boldsymbol{\Sigma}_a\left(\boldsymbol{\Sigma}_a + \frac{1}{N_m}\boldsymbol{\Sigma}_w\right)^{-1}\mathbf{y}_m + \boldsymbol{\Sigma}_w\left(N_m\boldsymbol{\Sigma}_a + \boldsymbol{\Sigma}_w\right)^{-1}\boldsymbol{\mu} + \boldsymbol{\Sigma}_a\left(\boldsymbol{\Sigma}_a + \frac{1}{N_m}\boldsymbol{\Sigma}_w\right)^{-1}\frac{1}{N_m}\sum_{n=1}^{N_m}\mathbf{c}_{mn}. \tag{22}$$

The expected value of $\mathbf{y}_m$ in (22) includes an additive noise term due only to channel effects. If channel components are zero-mean, which can be expected in the case of high channel-diversity, the noise term will also be zero-mean. Conversely, for low channel diversity, the noise term will not be zero-mean, introducing distortion to $\langle \mathbf{y}_m\rangle$. Furthermore, the posterior covariance of $\mathbf{y}_m$, given by $\boldsymbol{\Sigma}_w\left(N_m\boldsymbol{\Sigma}_a + \boldsymbol{\Sigma}_w\right)^{-1}\boldsymbol{\Sigma}_a$, shrinks as N_m increases, causing the model to become increasingly confident in this inaccurate estimate.

4. Unsupervised Bayesian Adaptation

In the case of low channel diversity, speaker labels can be adjusted to keep Bayesian adaptation from becoming overly confident. For example, sets of samples from individual speakers can be limited so that N_m does not exceed a certain value. In the extreme case, all in-domain data can be assumed to be independent, implying that each sample is provided by a unique speaker. Note that this assumption eliminates the requirement for speaker labels for the in-domain data set.

If data samples in $\mathcal{X}$ are assumed independent, implying that each speaker contributed a single sample, the conditional likelihood from (15) reduces to:

$$\log p\left(\mathcal{X}\,|\mathcal{Y}, \boldsymbol{\Sigma}_w\right) = \sum_{n=1}^{N_T}\mathcal{N}\left(\mathbf{x}_n; \mathbf{y}_n, \boldsymbol{\Sigma}_w\right). \tag{23}$$

where subscripts can be changed to omit speaker label, and N_T denotes the number of samples in $\mathcal{X}$. If (23) is substituted into (14), the VB solution from (16)-(21) becomes:

$$\boldsymbol{\mu} = \alpha\bar{\mathbf{y}} + (1-\alpha)\,\boldsymbol{\mu}^{\boldsymbol{out}}, \tag{24}$$

$$\boldsymbol{\Sigma}_a = \alpha\left(\frac{1}{N_T}\sum_{n=1}^{N_T}\langle \mathbf{y}_n\mathbf{y}_n^T\rangle - \bar{\mathbf{y}}\bar{\mathbf{y}}^T\right) + (1-\alpha)\,\boldsymbol{\Sigma}_a^{out} + \alpha\,(1-\alpha)\left(\bar{\mathbf{y}} - \boldsymbol{\mu}^{\boldsymbol{out}}\right)\left(\bar{\mathbf{y}} - \boldsymbol{\mu}^{\boldsymbol{out}}\right)^T, \tag{25}$$

$$\boldsymbol{\Sigma}_w = \frac{\alpha}{N_T}\sum_{n=1}^{N_T}\Big(\mathbf{x}_n\mathbf{x}_n^T - \langle \mathbf{y}_n\rangle\mathbf{x}_n^T - \mathbf{x}_n\langle \mathbf{y}_n\rangle^T + \langle \mathbf{y}_n\mathbf{y}_n^T\rangle\Big) + (1-\alpha)\,\boldsymbol{\Sigma}_w^{out}, \tag{26}$$

where:

$$\bar{\mathbf{y}} = \frac{1}{N_T}\sum_{n=1}^{N_T}\langle \mathbf{y}_n\rangle, \tag{27}$$

$$\langle \mathbf{y}_n\rangle = \boldsymbol{\Sigma}_a\left(\boldsymbol{\Sigma}_a + \boldsymbol{\Sigma}_w\right)^{-1}\mathbf{x}_n + \boldsymbol{\Sigma}_w\left(\boldsymbol{\Sigma}_a + \boldsymbol{\Sigma}_w\right)^{-1}\boldsymbol{\mu}, \tag{28}$$

$$\langle \mathbf{y}_n\mathbf{y}_n^T\rangle = \left(\boldsymbol{\Sigma}_a^{-1} + \boldsymbol{\Sigma}_w^{-1}\right)^{-1} + \langle \mathbf{y}_n\rangle\langle \mathbf{y}_n\rangle^T. \tag{29}$$

It may provide insight to substitute (27)-(29) into (24)-(26) to obtain update equations for $\boldsymbol{\Sigma}_a$ and $\boldsymbol{\Sigma}_w$. If $\boldsymbol{\Sigma}_t$ denotes the global covariance of $\mathcal{X}$ (and the sample means of $\mathcal{X}$ and the out-of-domain sets are assumed to be zero for illustrative purposes), the update equations become:

$$\boldsymbol{\Sigma}_a \Leftarrow \alpha\left(\mathbf{H}_a\boldsymbol{\Sigma}_t\mathbf{H}_a^T + \left(\boldsymbol{\Sigma}_a^{-1} + \boldsymbol{\Sigma}_w^{-1}\right)^{-1}\right) + (1-\alpha)\,\boldsymbol{\Sigma}_a^{out}, \tag{30}$$

$$\boldsymbol{\Sigma}_w \Leftarrow \alpha\left(\mathbf{H}_w\boldsymbol{\Sigma}_t\mathbf{H}_w^T + \left(\boldsymbol{\Sigma}_a^{-1} + \boldsymbol{\Sigma}_w^{-1}\right)^{-1}\right) + (1-\alpha)\,\boldsymbol{\Sigma}_w^{out}, \tag{31}$$

where $\mathbf{H}_a{=}\boldsymbol{\Sigma}_a\left(\boldsymbol{\Sigma}_a + \boldsymbol{\Sigma}_w\right)^{-1}$ and $\mathbf{H}_w{=}\boldsymbol{\Sigma}_w\left(\boldsymbol{\Sigma}_a + \boldsymbol{\Sigma}_w\right)^{-1}$. The proposed adaptation technique can then be interpreted as iteratively designing Wiener filters to extract the across-class and within-class variabilities from the total variability of the data set. Furthermore, the out-of-domain hyperparameter estimates serve both to provide initial estimates of these filters, and to constrain the across-class and within-class variabilities during optimization. It should be noted that the adaptation coefficient can not be set too close to 1, since the out-of-domain estimates are required to constrain the optimization.

It is of interest to note that the update equations for $\boldsymbol{\Sigma}_a$ and $\boldsymbol{\Sigma}_w$ show strong similarity to iterative Wiener filters (IWFs). Specifically, in the trivial case of α=1, the update equations are each identical to the IWF with an additive correction factor proposed in [10], where the term $\left(\boldsymbol{\Sigma}_a^{-1} + \boldsymbol{\Sigma}_w^{-1}\right)^{-1}$ ensures that the property $\boldsymbol{\Sigma}_t{=}\boldsymbol{\Sigma}_a{+}\boldsymbol{\Sigma}_w$ holds at the stationary point.

5. Experimental Results

5.1. System Description

This section presents experimental results for domain adaptation with inadequate in-domain data. The baseline system uses 600-dimensional i-vectors, with global centering and whitening applied prior to length normalization [11]. The system uses 40-dimensional cepstral features including deltas, with mean and variance normalization. All speaker verification results are presented for PLDA scoring, in terms of equal error rate (EER) and

minimum decision cost function (mindcf), and pooled across gender. For all adaptation methods, an adaptation coefficient of α=0.5 was used. As baseline methods, we use MAP adaptation with point estimates [3] and supervised Bayesian adaptation [5].

5.2. Domain Adaptation with Few Samples Per Speaker

Table 1 provides speaker verification results for a variety of domain adaptation strategies, when applied to the DAC13 [8]. The out-of-domain system was trained on the Switchboard-I and Switchboard-II corpora, and the in-domain data included telephone calls from SRE04-SRE08. The in-domain data was reduced to contain only 2 randomly drawn samples per speaker. Five random draws of the in-domain set were tested, and results represent the average of these. In Table 1, *none* refers to the unadapted out-of-domain system, and *whitening* refers to using in-domain data solely to adapt i-vector whitening and centering prior to length normalization. In Table 1, MAP adaptation with point estimates [3] suffers degradation since speaker components estimated with (6) are highly variable for small values of N_m, as discussed in Sec. 3.2. However, Bayesian adaptation from [5] and the proposed unsupervised Bayesian adaptation perform well since they take into account the uncertainty present when estimating speaker components from limited data.

Table 1: *Speaker verification results on the DAC13 task, using an in-domain set with 2 samples per speaker*

adaptation method	EER (%)	mindcf
none	6.41	0.471
whitening	5.25	0.412
MAP with point estimates [3]	3.18	0.296
Supervised Bayesian [5]	2.74	0.271
Unsupervised Bayesian	2.92	0.270

5.3. Domain Adaptation with Low Channel Diversity

Realistic in-domain data sets may also be inadequate due to limited channel diversity. Table 2 provides results for the DAC13 task when the in-domain set only includes samples from the dominant phone number for each speaker, so that the in-domain set consisted of $\sim$24k samples from $\sim$3800 speakers. In this case, the baseline methods suffer degradation since both will produce distorted speaker component estimates when channel components have non-zero mean, as discussed in Sec. 3.3. The proposed method, however, does not rely on channel diversity per speaker, and suffers no such degradation.

Table 2: *Speaker verification results on the DAC13 task, using an in-domain set with a single phone number from each speaker*

adaptation method	EER (%)	mindcf
none	6.41	0.471
whitening	5.17	0.413
MAP with point estimates [3]	5.51	0.424
Supervised Bayesian [5]	5.47	0.422
Unsupervised Bayesian	3.07	0.278

5.4. Domain Adaptation with Resource-rich Data

Table 3 provides verification results for the DAC13 task for the full in-domain set, consisting of $\sim$36k samples from $\sim$3800 speakers, with an average of 9.6 samples per speaker and 2.8 phone numbers per speaker. This can be considered a resource-rich adaptation set, and the systems from [3] and [5] can both be expected to perform well. However, we wish to verify that the proposed method remains competitive, even though it is not able to leverage the rich information provided by speaker labels. It can be observed in Table 3 that the baseline methods provide excellent results, and that the proposed technique suffers only a slight degradation.

Table 3: *Speaker verification results on the DAC13 task, using the full in-domain set*

adaptation method	EER (%)	mindcf
none	6.41	0.471
whitening	5.20	0.413
MAP with point estimates [3]	2.54	0.254
Supervised Bayesian [5]	2.40	0.245
Unsupervised Bayesian	2.96	0.269

5.5. Domain Adaptation on the SRE16

To verify that the observations made from Tables 1-3 generalize to other data sets, we performed experiments on the 2016 Speaker Recognition Evaluation (SRE16) [12] fixed task. In this case, the out-of-domain system was was trained using data from SRE04-SRE12. The SRE16 included trials in two non-English languages and from unseen channels. An inadequate in-domain set was provided, which consisted of 2272 samples from 1164 speakers, and spanned both unseen languages. An adaptation coefficient of α=0.1 was used. In the set, speakers provided an average of 1.9 samples from 1.1 different phone numbers, making this data deficient with respect to both samples per speaker and channel diversity. It can be observed in Table 4 that the baseline technique from [3] provides benefit in terms of mindcf, and supervised Bayesian adaptation [5] provides improvements in terms of both EER and mindcf. The unsupervised Bayesian approach provides results which are competitive with [5], even though speaker labels are not required for the in-domain data.

Table 4: *Speaker Verification Results on the SRE16 Fixed Task*

adaptation method	EER (%)	mindcf
none	19.35	0.999
whitening	16.70	0.972
MAP with point estimates [3]	18.26	0.775
Supervised Bayesian [5]	15.49	0.776
Unsupervised Bayesian	15.32	0.723

6. Conclusions

We have proposed a technique for unsupervised Bayesian domain adaptation for i-vector speaker verification. The method shows improved effectiveness for inadequate in-domain data. Specifically, the method performs well even when in-domain sets include speakers with few samples or low channel diversity. The proposed technique provides competitive results on a range of domain adaptation experiments with inadequate data, even when compared to supervised systems which require speaker labels for the in-domain set.

7. References

[1] N. Dehak, P. Kenny, R. Dehak, P. Ouellet, and P. Dumouchel, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 788–798, 2011.

[2] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *ICCV*, 2007.

[3] D. Garcia-Romero and A. McCree, "Supervised domain adaptation for i-vector speaker recognition," in *ICASSP*, 2014.

[4] C. Vaquero, "Dataset shift in plda based speaker verification," in *Odyssey*, 2012.

[5] J. Villalba and E. Lleida, "Bayesian adaptation of plda based speaker recognition to domains with scarce development data," in *Odyssey*, 2012.

[6] D. Garcia-Romero, A. McCree, S. Shum, N. Brummer, and C. Vaquero, "Unsupervised domain adaptation for i-vector speaker recognition," in *Odyssey*, 2014.

[7] Q. Wang, H. Yamamoto, and T. Koshinaka, "Domain adaptation using maximum likelihood linear transformation for plda-based speaker verification," in *ICASSP*, 2016.

[8] "https://engineering.jhu.edu/clsp/wp-content/uploads/sites/75/2015/10/ws13-speaker-dac.pdf."

[9] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2013.

[10] A. D. Hillery and R. T. Chin, "Iterative wiener filters for image restoration," *IEEE Transactions on Signal Processing*, vol. 39, no. 8, pp. 1892–1899, 1991.

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

[12] "https://www.nist.gov/itl/iad/mig/speaker-recognition-evaluation-2016."