SOURCE-SPECIFIC INFORMATIVE PRIOR FOR I-VECTOR EXTRACTION

Sven Ewan Shepstone[1,2,3], *Kong Aik Lee*[2], *Haizhou Li*[2], *Zheng-Hua Tan*[3], *Søren Holdt Jensen*[3]

[1]Bang and Olufsen A/S,
Peter Bangs Vej 15, 7600 Struer, Denmark
[2]Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore
[3]Department of Electronic Systems, Aalborg University,
Fredrik Bajers Vej 7B, 9220 Aalborg, Denmark
ssh@bang-olufsen.dk, {kalee,hli}@i2r.a-star.edu.sg, {zt,shj}@es.aau.dk

ABSTRACT

An i-vector is a low-dimensional fixed-length representation of a variable-length speech utterance, and is defined as the posterior mean of a latent variable conditioned on the observed feature sequence of an utterance. The assumption is that the prior for the latent variable is non-informative, since for homogeneous datasets there is no gain in generality in using an informative prior. This work shows that extracting i-vectors for a heterogeneous dataset, containing speech samples recorded from multiple sources, using informative priors instead is applicable, and leads to favorable results. Tests carried out on the NIST 2008 and 2010 Speaker Recognition Evaluation (SRE) dataset show that our proposed method beats three baselines: For the short2-short3 core-task in SRE'08, for the female and male cases, five and six respectively, out of eight common conditions were beaten, and for the core-core task in SRE'10, for both genders, five out of nine common conditions were beaten.

Index Terms— i-vector, informative prior, total variability, source variation

1. INTRODUCTION

In the i-vector approach, variable-length speech utterances are mapped into fixed-length low dimensional vectors that reside in the so-called total variability space [1]. The i-vectors capture the *total* variability, which is usually understood to include both speaker and channel variability. The ease of dealing with i-vectors has resulted in a myriad of techniques being proposed to maximize speaker discrimination and reduce channel effects, which include amongst others *within-class covariance normalization* (WCCN) [2], *linear discriminant analysis* (LDA) [3], and *probabilistic LDA* (PLDA) [4].

When i-vectors are extracted from a heterogeneous dataset, as encountered in the recent NIST SREs [5, 6], not only will they capture both speaker and channel variability, but also source variation. If this source variation is not dealt with, it will adversely affect speaker recognition performance [3, 7]. The notion of source variation was introduced in the recent SREs and it is related to the speech acquisition method (e.g., telephone versus microphone channel types) and recording scenario (e.g., telephone conversation versus interview styles). The various combinations of styles and channel types (e.g., interview speech recorded over microphone channel) form relatively homogeneous subsets of the dataset. In this work, the dataset consists of *telephone*, *microphone* (telephone conversation recorded over microphone channel), and *interview* subsets, or sources.

Several proposals consider the issue of source variation within the context of total variability modeling. In [8], the authors address the issue of estimating the inter-speaker scatter matrix given a heterogeneous dataset where most speakers appear only once in any one of the sources. The source variation will be strongly represented and seen as part of the inter-speaker variability and will therefore be optimized in the resulting LDA transform. Another proposal involves training of a supplementary matrix for the *microphone* subset on top of an already trained total variability matrices on *telephone* data [3]. I-vectors are then extracted from a total variability matrix formed by concatenating the two matrices. PLDA has also been used to further project microphone and telephone factors to a common space [9]. Compensation using heavy-tailed PLDA has also been successful [10]. Finally, a total variability matrix can be trained from a pooled set of the training data. All these schemes require either training of a supplementary matrix or retraining of the total variability matrix.

This work proposes to deal with the source variability by using an informative prior at the i-vector extraction stage. The objective is to use the same total variability matrix to describe the speaker and channel variability across sources of data from a heterogeneous dataset, with the source variation modeled at the priors. Re-training of the total variability matrix is not required, neither in whole or in part. Instead we assume a matrix already trained using abundantly available data. We show how a *source-specific* prior can be used in the i-vector extraction phase to compensate for unwanted source variability. The extracted i-vectors, which now only capture speaker and channel variability, can be processed at the LDA or PLDA stages without needing to carry out any source variation suppression.

This paper is structured as follows: Section 2 reviews the i-vector paradigm and the use of the non-informative prior. Section 3 gives the motivation for using an informative prior when a heterogeneous dataset is concerned. Section 4 presents theory for estimating the source-specific priors and using them effectively in extracting i-vectors. The following two section present the experiments that were carried out and our results, and the final section concludes the paper.

2. THE I-VECTOR PARADIGM

The total variability model assumes that a speaker- and channel-dependent GMM supervector $\mathbf{m}$ of an utterance [11] is modeled as

$$\mathbf{m} = \mathbf{m}_0 + \mathbf{T}\mathbf{w} \tag{1}$$

This work is supported by the Danish Ministry of Science, Innovation and Higher Education under Grant no 12-122802.

where $\mathbf{m_0}$ is the speaker-independent supervector obtained by concatenating the mean vectors from the UBM. The hidden variable $\mathbf{w}$ weights the columns of the matrix $\mathbf{T}$ to explain the observed deviation from $\mathbf{m_0}$. The matrix $\mathbf{T}$ is defined to have low rank so as to model the subspace where both the speaker and channel variability (hence the name total variability matrix) correlate the most. The training of the total variability matrix follows the same process as that of training an eigenvoice matrix [12, 13]. The major difference is that utterances from the same speakers are treated individually as unrelated sessions [1].

Let $\{\mathbf{o}_1, \mathbf{o}_2, ...\mathbf{o}_T\}$ represent the feature sequence of a given utterance O. The feature vectors are assumed to be drawn from a GMM with its mean supervector as in (1). For each mixture component c of the GMM, the following Baum-Welch statistics are defined:

$$N(c) = \sum_t \gamma_t(c) \tag{2}$$

where t extends over all frames of an utterance and $\gamma_t(c)$ is the occupancy of frame $\mathbf{o}_t$ to the c-th Gaussian. We further denote the centered first-order statistics as

$$\widetilde{\mathbf{F}}(c) = \sum_t \gamma_t(c)(\mathbf{o}_t - \mathbf{m_0}(c)) \tag{3}$$

Also, let $\mathbf{N}$ represent the diagonal matrix whose diagonal blocks are $N(c) \times \mathbf{I}$ and let $\widetilde{\mathbf{F}}$ represent the supervector obtained by concatenating the $\widetilde{\mathbf{F}}(c)$, where c extends over all mixtures in both cases. In order to extract an i-vector, given an already trained $\mathbf{T}$, we compute the posterior distribution over the latent variable $\mathbf{w}$ conditioned on the observations. Assuming a standard normal prior $\mathbf{w} \sim \mathcal{N}(0, I)$, the posterior distribution is also Gaussian [12], as follows

$$p(\mathbf{w}|O) = \mathcal{N}(\mathbf{L}^{-1} \cdot \mathbf{T}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\widetilde{\mathbf{F}},\ \mathbf{L}^{-1}) \tag{4}$$

with mean vector

$$\phi = \mathbf{L}^{-1} \cdot \mathbf{T}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\widetilde{\mathbf{F}} \tag{5}$$

and precision $\mathbf{L} = (\mathbf{I} + \mathbf{T}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{N}\mathbf{T})$. The i-vector is then given by the mean vector ϕ of the posterior distribution [1]. Similar to that of $\mathbf{N}$, the matrix $\mathbf{\Sigma}$ in (4) is constructed by having its diagonal blocks made up by the covariance matrices of the UBM.

The prior over the hidden variable $\mathbf{w}$ is usually taken to be a standard normal distribution. While it is indeed possible to define an informative prior, this prior can always be absorbed to the global mean vector $\mathbf{m_0}$ and the loading matrix $\mathbf{T}$ [13, 14]. This step causes the resulting prior to become non-informative, thereby requiring no alteration to (4). As such, there is no compelling reason to use an informative prior at least for the case when the dataset is homogeneous. In the following, we show how informative priors of the form $\mathbf{w} \sim \mathcal{N}(\mu_{\mathbf{p}}, \mathbf{\Sigma_p})$, where $\mu_{\mathbf{p}} \neq 0$ and $\mathbf{\Sigma_p} \neq \mathbf{I}$, could be modeled and used for i-vector extraction, and the benefit of doing so when a heterogeneous dataset is concerned. In the NIST series of speaker recognition evaluations (SREs), for instance, the dataset contains "telephone", "interview" or "microphone" speech sources [5, 6].

3. INTRODUCING INFORMATIVE PRIORS

An informative prior encodes domain knowledge (i.e., the source variation) by capturing underlying dependencies between the parameters [15]. In this section, we propose using minimum divergence criterion for estimating source-specific priors from a heterogeneous dataset. We then show how to incorporate the informative prior in the i-vector extraction formula.

3.1. Minimum divergence estimation

Consider the case where individual speech sources (e.g., telephone, microphone, or interview in NIST SRE) forms a relatively homogeneous subset and each speech source has I number of utterances. For each utterance we compute the posterior distribution according to (4) using the already trained $\mathbf{T}$ matrix. Given the set of posterior distributions, we seek for a Gaussian distribution $\mathcal{N}(\mu_{\mathbf{p}}, \mathbf{\Sigma_p})$ that best describes the I posterior distributions. This could be achieved by minimizing the Kullback-Leibler (KL) divergence of the desired distribution $\mathcal{N}(\mu_{\mathbf{p}}, \mathbf{\Sigma_p})$ from all the I posteriors $\mathcal{N}(\phi_i, \mathbf{L}_i^{-1})$. As shown in [16], the closed form solution consists of the mean vector

$$\mu_{\mathbf{p}} = \frac{1}{I}\sum_{i=1}^{I}\phi_i \tag{6}$$

and the covariance matrix

$$\mathbf{\Sigma_p} = \frac{1}{I}\sum_{i=1}^{I}(\phi_i - \mu_{\mathbf{p}})(\phi_i - \mu_{\mathbf{p}})^{\mathrm{T}} + \frac{1}{I}\sum_{i=1}^{I}\mathbf{L}_i^{-1} \tag{7}$$

Notice that the number of utterances I is generally different for each speech source. The central idea here is to use a single $\mathbf{T}$ matrix for all sources of data, where the variability due to the different sources is modeled at the prior. Together, the combination of $\mathbf{T}$ and the source-specific priors better models the variation across sources from the heterogeneous dataset.

Notice that the mean $\mu_{\mathbf{p}}$ of the informative prior is given by the average of all the i-vectors belonging to a target set (recall that an i-vector is given by the mean of the posterior distribution). The deviation of the i-vectors from $\mu_{\mathbf{p}}$ forms the empirical term in the covariance $\mathbf{\Sigma_p}$, while the second term accounts form posterior covariances of the i-vectors.

3.2. Posterior inference with informative prior

We formulate the expression for the posterior distribution for the general case when the informative prior as estimated above is used in place of a non-informative one.

Proposition 1: Consider an informative prior $p(\mathbf{w}) \sim \mathcal{N}(\mu_{\mathbf{p}}, \mathbf{\Sigma_p})$ with mean $\mu_{\mathbf{p}}$ and the covariance matrix $\mathbf{\Sigma_p}$. The posterior distribution $p(\mathbf{w}|O)$ is Gaussian with mean

$$\phi = \mathbf{L}^{-1}(\mathbf{T}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\widetilde{\mathbf{F}} + \mathbf{\Sigma_p}^{-1}\mu_{\mathbf{p}}) \tag{8}$$

and precision

$$\mathbf{L} = \mathbf{T}^{\mathrm{T}}\mathrm{N}\mathbf{\Sigma}^{-1}\mathbf{T} + \mathbf{\Sigma_p}^{-1} \tag{9}$$

Note that by setting $\mu_{\mathbf{p}} = \mathbf{0}$ and $\mathbf{\Sigma_p} = \mathbf{I}$, the posterior mean ϕ (i.e., the i-vector) and precision $\mathbf{L}$ reduce to the standard form of i-vector extraction with a non-informative prior as in (4).

Proof. Assume that we have the parameter set $(\mathbf{T}, \mathbf{\Sigma})$, the hidden variable $\mathbf{w}$ and the observation O. From Lemma 1 in [12] we know that the log likelihood of O given $\mathbf{w}$ and the parameters $(\mathbf{T}, \mathbf{\Sigma})$ can be expressed as the sum of two terms:

$$\log p_{\mathbf{T},\mathbf{\Sigma}}(O|\mathbf{w}) = G_{\mathbf{T}} + H_{\mathbf{T},\mathbf{\Sigma}} \tag{10}$$

where $G_{\mathbf{T}}$ is defined by (3) in [12], and $H_{\mathbf{T},\mathbf{\Sigma}}$ is defined as

$$H_{\mathbf{T},\mathbf{\Sigma}} = \mathbf{w}^{\mathrm{T}}\mathbf{T}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\widetilde{\mathbf{F}} - \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{T}^{\mathrm{T}}\mathbf{N}\mathbf{\Sigma}^{-1}\mathbf{T}\mathbf{w} \tag{11}$$

Since $G_{\mathbf{T}}$ does not depend on $\mathbf{w}$, this term is not considered further. Given the mean $\mu_{\mathbf{p}}$ and covariance ${\mathbf{\Sigma}_{\mathbf{p}}}^{-1}$, we express the prior as:

$$p(\mathbf{w}) \propto exp(-\frac{1}{2}(\mathbf{w} - \mu_{\mathbf{p}})^{\mathrm{T}}{\mathbf{\Sigma}_{\mathbf{p}}}^{-1}(\mathbf{w} - \mu_{\mathbf{p}})) \tag{12}$$

The posterior distribution of $\mathbf{w}$ given O could be obtained by taking the product of (11) and (12), as follows:

$$\begin{aligned} p(\mathbf{w}|O) \propto exp(&\mathbf{w}^{\mathrm{T}}\mathbf{T}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{Ft} - \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{T}^{\mathrm{T}}\mathbf{N}\mathbf{\Sigma}^{-1}\mathbf{T}\mathbf{w} - \\ &\frac{1}{2}(\mathbf{w} - \mu_{\mathbf{p}})^{\mathrm{T}}{\mathbf{\Sigma}_{\mathbf{p}}}^{-1}(\mathbf{w} - \mu_{\mathbf{p}})) \\ \propto exp(&-\frac{1}{2}(\mathbf{w} - \phi)^{\mathrm{T}}\mathbf{L}(\mathbf{w} - \phi)) \end{aligned} \tag{13}$$

with ϕ and $\mathbf{L}$ in the form as stated above. □

4. PRIOR-COMPENSATED I-VECTOR EXTRACTION

In the Bayesian sense, an informative prior increases the prior belief of the location and dispersion of each source in a heterogeneous dataset. We note that a different spread is observed for each source in the i-vector space, as was also reported in a previous study [7]. In the case of cross-source trials, the test i-vectors belonging to one source and target i-vector belonging to another can no longer be assumed to lie close to one another, even when representing the same speaker. The implication of applying (8) directly would intensify the difference across speech sources, resulting in poorer performance.

We propose to compensate for the differences across speech sources (e.g., telephone versus microphone) by applying the prior mean and covariance at separate stages in the i-vector extraction phase. More specifically, we project the prior mean to the acoustic space, while the covariance remains intact as part of the prior. The operation of separating the prior mean and covariance is based on the equality of marginalization which we shall now demonstrate.

Proposition 2: Let $\Pi(c)$ be the marginal distribution for Gaussian c obtained by modeling $\mathbf{m} = \mathbf{m_0} + \mathbf{T}\mathbf{w}$ with the prior $\mathbf{w} \sim \mathcal{N}(\mu_{\mathbf{p}}, \mathbf{\Sigma}_{\mathbf{p}})$. For this source, the same marginalization $\Pi(c)$ can be realized by modeling $\mathbf{m} = \mathbf{m_0} + \mathbf{T}\mathbf{w} + \mathbf{T}\mu_{\mathbf{p}}$ with the prior $\mathbf{w} \sim \mathcal{N}(0, \mathbf{\Sigma}_{\mathbf{p}})$. This gives the following equality:

$$\begin{aligned} \Pi(c) &= \int \mathcal{N}(O|\mathbf{m_0}(c) + \mathbf{T}_c\mathbf{w}, \Sigma_0)\mathcal{N}(\mathbf{w}|\mu_{\mathbf{p}}, \mathbf{\Sigma}_{\mathbf{p}})d\mathbf{w} \\ &= \int \mathcal{N}(O|\mathbf{m_0}(c) + \mathbf{T}_c\mu_{\mathbf{p}} + \mathbf{T}_c\mathbf{w}, \Sigma_0)\mathcal{N}(\mathbf{w}|0, \mathbf{\Sigma}_{\mathbf{p}})d\mathbf{w} \end{aligned} \tag{14}$$

The proof of the proposition is given in the appendix.

Comparing the first and second rows of (14), the prior mean $\mu_{\mathbf{p}}$ is brought forward to the conditional density, which describes the acoustic observation O. By doing so, the projection $\mathbf{T}_c\mu_{\mathbf{p}}$ of the prior mean imposes a shift on the global mean vector $\mathbf{m_0}(c)$. This also gives rise to prior distributions with a common mode at the origin (i.e., zero mean) but different dispersions $\mathbf{\Sigma}_{\mathbf{p}}$ for individual sources. Algorithmically, the projection $\mathbf{T}_c\mu_{\mathbf{p}}$ is applied on the observation by re-centering the first order statistics $\widetilde{\mathbf{F}}(c)$, as follows

$$\begin{aligned} \widetilde{\widetilde{\mathbf{F}}}(c) &= \sum_t \gamma_t(c)(\mathbf{o}_t - \mathbf{m_0}(c) - \mathbf{T}_c\mu_{\mathbf{p}}) \\ &= \widetilde{\mathbf{F}}(c) - N(c)\mathbf{T}_c\mu_{\mathbf{p}} \end{aligned} \tag{15}$$

In a sense, the re-centering brings heterogeneous sources to a common mode at the origin of the total variability space and allows the priors to differ only with regard to one anothers' covariance.

The proposed prior-compensated i-vector extraction can be summarized into the following steps:

1. Start out with an already trained $\mathbf{T}$ matrix. For each source, extract an informative prior $\mathcal{N}(\mu_{\mathbf{p}}, \mathbf{\Sigma}_{\mathbf{p}})$ using the minimum divergence estimation as described in Section 3.1.

2. Re-center the first order statistics $\widetilde{\mathbf{F}}$ around the relevant source-specific mean to give $\widetilde{\widetilde{\mathbf{F}}}$, as in (15).

3. Extract i-vectors, by matching the now zero-mean informative prior $\mathcal{N}(0, \mathbf{\Sigma}_{\mathbf{p}})$ for each source to the relevant re-centered first-order statistics:

$$\begin{aligned} \phi =& \mathbf{L}^{-1}(\mathbf{T}^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\widetilde{\mathbf{F}} - \mathbf{N}\mathbf{T}\mu_{\mathbf{p}})) \\ =& \mathbf{L}^{-1}(\mathbf{T}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\widetilde{\widetilde{\mathbf{F}}}) \end{aligned} \tag{16}$$

   where the precision $\mathbf{L}$ is as given in (9).

5. EXPERIMENTS

5.1. Datasets and system setup

Our experiments were carried out on the short2-short3 core-task of SRE'08 [5] and the core-core task of SRE'10 [6]. For all experiments, a gender dependent setup was used. The features used for training the 512-Gaussian UBMs were 57-dimensional MFCCs (including the first and second derivatives). The first order statistics used for training each total variability matrix were centered and whitened [17]. For all experimental setups, a total variability matrix was trained with non-informative priors being used in the E-step.

We compare four individual experimental setups in this work, of which three are reference systems and one is the proposed system. In the *telephone only* setup, a 600 dimensional $\mathbf{T}$ matrix was trained using only the telephone data. In the *pooled* system, a 600 dimensional $\mathbf{T}$ matrix was trained using pooled telephone and microphone data. In the *cascade* system, a 400 dimensional $\mathbf{T}$ matrix was trained using the telephone data, and a 200 dimensional $\mathbf{T}$ matrix was trained using microphone data [3]. The telephone data used to train these systems was taken from SRE'04, 05 and 06. The microphone data was taken from SRE'05, 06 and MIXER 5. The same dataset was used to derive the informative priors.

In the *2-prior* system, the already trained *pooled* $\mathbf{T}$ matrix was used as the starting point. Using minimum divergence estimation (Section 3.1), we trained one prior for the telephone subset and another prior for microphone and interview subsets. We chose to use only one prior for both microphone and interview since there was not enough interview data to reliably estimate the interview prior. I-vectors were extracted by performing re-centering of the first-order

	CC1: int-int		CC2: int-int		CC3: int-int		CC4: int-tel		CC5: tel-mic		CC6: tel-tel		CC7: tel-tel		CC8: tel-tel	
EER	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F	M
Telephone only	3.51	2.84	1.50	0.32	3.61	2.97	5.69	4.07	6.65	4.17	5.85	4.67	2.73	2.32	3.24	1.43
Pooled	3.22	2.54	1.28	0.33	3.29	2.64	4.65	3.89	5.62	3.05	5.86	4.15	2.84	1.60	3.32	1.04
Cascade	3.17	3.01	1.25	0.41	3.26	3.22	5.38	4.27	6.10	4.12	5.86	4.06	2.98	1.66	3.81	1.32
2-prior	**2.34**	**1.95**	1.32	**0.32**	**2.39**	**2.04**	**4.32**	3.91	**5.37**	3.21	**5.79**	**3.84**	2.87	**1.39**	3.27	**0.90**

Table 1. SRE'08 Performance comparison for the sub-task short2-short3. Left: FEMALE Trials, Right: MALE Trials

	CC1: int-int-same-mic		CC2: int-int-diff-mic		CC3: int-tel		CC4: int-mic		CC5: nve-nve-diff-tel		CC6: nve-hve-diff-tel		CC7: nve-hve-mic		CC8: nve-lve-diff-tel		CC9: nve-lve-mic	
EER	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F	M
Telephone only	3.06	2.02	5.65	3.45	4.21	3.55	3.96	2.72	3.59	3.47	8.09	4.64	8.49	4.91	2.01	1.14	2.46	1.54
Pooled	3.16	2.22	5.13	3.14	3.34	2.82	3.78	2.54	3.00	2.60	7.13	4.01	7.98	4.95	1.66	1.54	2.55	1.34
Cascade	3.12	2.29	5.60	3.29	4.01	2.62	4.04	2.87	3.41	3.13	7.10	4.33	8.19	5.25	1.83	1.68	3.08	1.61
2-prior	**2.43**	**1.67**	**4.44**	**2.25**	3.87	3.19	**3.33**	**2.22**	**3.00**	2.89	7.11	4.13	**7.49**	**4.16**	**1.59**	1.56	2.48	**1.15**

Table 2. SRE'10 Performance comparison for the sub-task core-core. Left: FEMALE Trials, Right: MALE Trials

statistics using the prior's mean, followed by computation of the posteriors using the prior's informative covariance. LDA was used to bring the dimension of the 600-dimensional i-vectors down to 400. After carrying out length normalization, PLDA was used to model the channel variability. For the PLDA model, a separate 200 dimensional telephone matrix and 50 dimensional microphone matrix were trained, in a decoupled manner, similar to the setup in [18].

5.2. Results

We present results for the four systems, for both male and female trials. For all results, we used Equal Error Rate (EER). For the SRE'08 results, shown in Table 2 for both male and female trials, a substantial improvement was seen in sub-tasks 1 and 3, corresponding to the *int-int* condition. We could not beat the baseline for sub-task 2, which we believe is due to the smaller number of trials. For the mixed trials, i.e. sub-tasks 4 and 5, source-specific informative priors showed improved robustness against both the telephone-only and cascade cases. For the pooled case however, the results were a lot closer and we did not beat this baseline in all cases. Interestingly, our approach improved on several of the *tel-tel* only conditions, especially in the male case. From these results, it appears that source-specific informative priors offer the greatest strength in enhancing performance trials where the sources of the trail and target match.

We now discuss the SRE'10 results shown in Table 2. For the single source interview and mic sub-tasks, as given by sub-tasks 1, 2, 7 and 9, we were able to beat all baselines in 3 out of 4 sub-tasks in the female case and all cases in the male case. For telephone only trials, given by sub-tasks 5, 6 and 8, in only one case could all baselines be beaten. We believe the reason for the slightly worse results for SRE'10 is the similarity of the data used to train the $\mathbf{T}$ matrices and subspace PLDA models to that of SRE'08. For the cross-channel conditions, we noted better performance for the int-mic cross channel than for int-tel, strengthening our belief that best performance is gained where source and target trials are better.

6. CONCLUSION

In this paper, we proposed a novel method of using a single $\mathbf{T}$ matrix to better describe the source variation from a heterogeneous dataset. The gist of our proposal is to compensate for source variation by applying the prior mean and covariance at separate stages in the i-vector extraction. We showed that by using an existing $\mathbf{T}$ matrix, introducing informative priors for each source into the i-vector extraction stage leads to performance gains in 5 out of 8 and 6 out of 8 common conditions for the short2-short3 core-task in SRE'08 for the female and male case, respectively, and 5 out of 9 common conditions for the core-core task in SRE'10, for both the female and male case. The results show that source-specific informative priors offer the greatest strength in enhancing performance trials where the sources of the trail and target are similar, or match.

7. PROOF OF PROPOSITION 2

Proof. We first derive the probability distribution of $p(\mathbf{m})$ where $\mathbf{m} = \mathbf{m_0} + \mathbf{Tw}$ and $\mathbf{w} \sim \mathcal{N}(\mu_\mathbf{p}, \mathbf{\Sigma_p})$. The mean is computed as:

$$\mathrm{E}[\mathbf{m}] = \mathbf{m_0} + \mathbf{T}\mu_\mathbf{p} \tag{17}$$

and covariance as:

$$\mathrm{E}[(\mathbf{m} - \mathrm{E}[\mathbf{m}])^2] = \mathbf{T}\mathrm{E}[\mathbf{w}\mathbf{w}^\mathrm{T}]\mathbf{T}^\mathrm{T} - \mathbf{T}\mu_\mathbf{p}{\mu_\mathbf{p}}^\mathrm{T}\mathbf{T}^\mathrm{T} \tag{18}$$

Realizing that the covariance of the prior distribution for $P(\mathbf{w})$ is simply $\mathbf{\Sigma_p} = \mathrm{E}[(\mathbf{w} - \mathrm{E}[\mathbf{w}]^2)] = \mathrm{E}[\mathbf{w}\mathbf{w}^\mathrm{T}] - \mu_\mathbf{p}{\mu_\mathbf{p}}^\mathrm{T}$, substituting back into into (18) and simplifying, gives:

$$\mathrm{E}[(\mathbf{m} - \mathrm{E}[\mathbf{m}])^2] = \mathbf{T}\mathbf{\Sigma}_p\mathbf{T}^\mathrm{T} \tag{19}$$

Note that for the case of the non-informative prior, the mean and covariance are reduced to $\mathbf{m_0}$ and $\mathbf{T}\mathbf{T}^\mathrm{T}$, respectively. In the same vein, we compute the mean for the marginalization modeled by $\mathbf{m} = \mathbf{m_0} + \mathbf{Tw} + \mathbf{T}\mu_\mathbf{p}$ and $\mathbf{w} \sim \mathcal{N}(0, \mathbf{\Sigma_p})$. We find the mean to be

$$\mathrm{E}[\mathbf{m}] \quad = \mathbf{m_0} + \mathbf{T}\mu_\mathbf{p} \tag{20}$$

which is identical to the formally derived mean. The covariance is computed as:

$$\mathrm{E}[(\mathbf{m} - \mathrm{E}[\mathbf{m}])^2] = \mathbf{T}\mathrm{E}[\mathbf{w}\mathbf{w}^\mathrm{T}]\mathbf{T}^\mathrm{T} \tag{21}$$

Now the covariance $\mathbf{\Sigma}_p = \mathrm{E}[(\mathbf{w} - \mathrm{E}[\mathbf{w}])^2] = \mathrm{E}[\mathbf{w}\mathbf{w}^\mathrm{T}]$, which when substituted back into (21), gives:

$$\mathrm{E}[(\mathbf{m} - \mathrm{E}[\mathbf{m}^2)] = \mathbf{T}\mathbf{\Sigma}_p\mathbf{T}^\mathrm{T} \tag{22}$$

which is identical to the formally derived covariance. These will contribute equally to the marginalization $\Pi(c)$ given in (14). This concludes the proof.

□

8. REFERENCES

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

[2] Andrew O Hatch, Sachin S Kajarekar, and Andreas Stolcke, "Within-class covariance normalization for SVM-based speaker recognition.," in *Interspeech*, 2006, pp. 1471–1474.

[3] Mohammed Senoussaoui, Patrick Kenny, Najim Dehak, and Pierre Dumouchel, "An i-vector extractor suitable for speaker recognition with both microphone and telephone speech.," in *Odyssey*, 2010, p. 6.

[4] Ye Jiang, Kong Aik Lee, and Longbiao Wang, "PLDA in the i-supervector space for text-independent speaker verification," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2014, no. 1, pp. 1–13, 2014.

[5] National Institute of Standards and Technology, "The NIST 2008 SRE Evaluation Plan," 2008.

[6] National Institute of Standards and Technology, "The NIST 2010 SRE Evaluation Plan," 2010.

[7] Mitchell McLaren and David Van Leeuwen, "Source-normalised-and-weighted LDA for robust speaker recognition using i-vectors," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5456–5459.

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

[9] Najim Dehak, Zahi N Karam, Douglas A Reynolds, Reda Dehak, William M Campbell, and James R Glass, "A channel-blind system for speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 4536–4539.

[10] Mohammed Senoussaoui, Patrick Kenny, Pierre Dumouchel, and Fabio Castaldo, "Well-calibrated heavy tailed bayesian speaker verification for microphone speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 4824–4827.

[11] Tomi Kinnunen and Haizhou Li, "An overview of text-independent speaker recognition: from features to supervectors," *Speech communication*, vol. 52, no. 1, pp. 12–40, 2010.

[12] Patrick Kenny, Gilles Boulianne, and Pierre Dumouchel, "Eigenvoice modeling with sparse training data," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 3, pp. 345–354, 2005.

[13] Patrick Kenny, Pierre Ouellet, Najim Dehak, Vishwa Gupta, and Pierre Dumouchel, "A study of interspeaker variability in speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 5, pp. 980–988, 2008.

[14] Kevin P Murphy, *Machine learning: a probabilistic perspective*, MIT press, 2012.

[15] Rajat Raina, Andrew Y Ng, and Daphne Koller, "Constructing informative priors using transfer learning," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 713–720.

[16] Liping Chen, Kong Aik Lee, Bin Ma, Wu Guo, Haizhou Li, and Li Rong Dai, "Minimum divergence estimation of speaker prior in multi-session PLDA scoring," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 4007–4011.

[17] Patrick Kenny, "A small foot-print i-vector extractor," in *Proc. Odyssey*, 2012.

[18] Kong Aik Lee, Anthony Larcher, Chang Huai You, Bin Ma, and Haizhou Li, "Multi-session PLDA scoring of i-vector for partially open-set speaker detection.," in *INTERSPEECH*, 2013, pp. 3651–3655.