DOMAIN ADAPTATION VIA WITHIN-CLASS COVARIANCE CORRECTION IN I-VECTOR BASED SPEAKER RECOGNITION SYSTEMS

Ondřej Glembek[1,2], *Jeff Ma*[1], *Pavel Matějka*[1,2], *Bing Zhang*[1], *Oldřich Plchot*[2], *Lukáš Burget*[2], *Spyros Matsoukas*[3]

[1]Raytheon BBN Technologies, Cambridge, USA
[2]Speech@FIT group, Brno University of Technology, Czech Republic
[3]Amazon, Boston, USA
{oglembek,jma,pmatejka,bzhang}@bbn.com,
{iplchot,burget}@fit.vutbr.cz, matsouka@amazon.com

ABSTRACT

In this paper we propose a technique of Within-Class Covariance Correction (WCC) for Linear Discriminant Analysis (LDA) in Speaker Recognition to perform an unsupervised adaptation of LDA to an unseen data domain, and/or to compensate for speaker population difference among different portions of LDA training dataset. The paper follows on the study of source-normalization and inter-database variability compensation techniques which deal with multi-modal distribution of i-vectors. On the DARPA RATS (Robust Automatic Transcription of Speech) task, we show that, with two hours of unsupervised data, we improve the Equal-Error Rate (EER) by 17.5%, and 36% relative on the unmatched and semi-matched conditions, respectively. On the Domain Adaptation Challenge we show up to 70% relative EER reduction and we propose a data clustering procedure to identify the directions of the domain-based variability in the adaptation data.

Index Terms— speaker recognition, i-vectors, source normalization, LDA, inter-dataset variability compensation

1. INTRODUCTION

I-vector based systems have recently become the state-of-the-art framework in Speaker Recognition. They provide an elegant way of reducing the large-dimensional variable-length input data to a small-fixed-dimensional feature vector while retaining most of the relevant information. The technique was originally inspired by Joint Factor Analysis framework introduced in [1, 2].

The objective in speaker verification calls for robust extraction of the relevant speaker information. However, an i-vector contains not only the *speaker* information—resulting in *wanted* variability in the i-vector space—but also all kinds of *unwanted* information—resulting in what is commonly referred to as *channel*.

There are various techniques to deal with these two types of variability which all aim at suppressing as much of the channel variability and emphasizing as much of the speaker variability as possible, such as Nuisance Attribute Projection (NAP [3]) or Linear Discriminant Analysis (LDA). Note that Probabilistic Linear Discriminant Analysis (PLDA [4])—as used in this work—also deals with this issue, but it is applied at the scoring level and it is not studied in this paper.

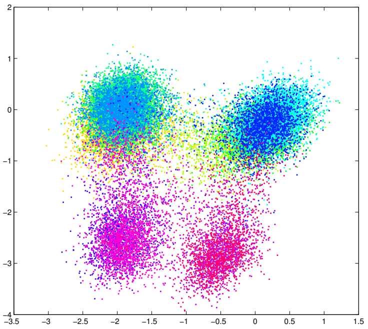

Fig. 1. Multi-modal distribution of i-vectors taken from multiple Switchboard databases, projected into first two bases of PCA estimated using the between-dataset covariance. The top two clusters belong to the SWB-Cellular collection, while the bottom two belong to SWB-Phase 2 collection. From left to right, we see two male, and two female clusters, respectively.

In various studies, it was observed that, a distribution of a collection of i-vectors can be multi-modal and it was shown that the modes correspond to different data sub-collections [5, 6, 7]. Such data mismatch is also often typical between the training and the test-set. For sake of clarity, we will refer to the sources as datasets in this work and use capital letter D for reference. Figure 1 shows an analysis of one of our training sets of i-vectors. Such multi-modality can lead to misinterpretation of the channel and speaker information. In the studies referenced above, it was shown how to compensate for this phenomena by the use of within- and between-speaker covariance matrices. In this work we present a similar technique of Within-Speaker Covariance Correction (WCC) and we show how it can be extended in unsupervised adaptation of the LDA matrix to compensate for the mismatch of the training and the test datasets.

2. THEORETICAL BACKGROUND

Let us first take a look at the anatomy of our recognition system. It is based on a comparison of a pair of pre-processed i-vectors. The

This work was partly supported by the DARPA RATS Program under Contract No. D10PC20015, by Technology Agency of the Czech Republic grant No. TA01011328, and by IT4Innovations Centre of Excellence CZ.1.05/1.1.00/02.0070. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government. Spyros Matsoukas was affiliated with Raytheon BBN Technologies at time of cooperation.

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

comparison is facilitated via Probabilistic Linear Discriminant Analysis (PLDA) model [8, 4]. Given a pair of i-vectors, PLDA allows to compute the log-likelihood for the same-speaker hypothesis and for the different-speaker hypothesis.

The pre-processing consists of applying LDA to reduce the dimensionality and will be discussed further in this work. Such processed i-vectors are then followed by global mean and variance normalization, followed by length-normalization [9, 10].

Let us recall that LDA is based on computing the between-class and withing-class covariance matrices $\mathbf{\Sigma}_\text{B}$ and $\mathbf{\Sigma}_\text{W}$, whose Maximum-Likelihood (ML), or un-equalized estimation is given as

$$\mathbf{\Sigma}_B = \frac{1}{N}\sum_{c=1}^{C} N_c(\boldsymbol{\mu}_c - \boldsymbol{\mu})(\boldsymbol{\mu}_c - \boldsymbol{\mu})' \tag{1}$$

$$\mathbf{\Sigma}_W = \frac{1}{N}\sum_{c=1}^{C}\sum_{n=1}^{N_c}(\boldsymbol{\phi}_{n,c} - \boldsymbol{\mu}_c)(\boldsymbol{\phi}_{n,c} - \boldsymbol{\mu}_c)' \tag{2}$$

where $\boldsymbol{\phi}_{n,c}$ is the n-th i-vector in class c, C is number of classes, N_c is number of data-points in class c, $\boldsymbol{\mu}_c$ is the mean of the data belonging to class c:

$$\boldsymbol{\mu}_c = \frac{1}{N_c}\sum_{n=1}^{N_c}\boldsymbol{\phi}_{n,c}, \tag{3}$$

where $\boldsymbol{\phi}_{n,c}$ is n-th data-point in class c, and $\boldsymbol{\mu}$ is the global mean of the data, computed as

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\phi}_n. \tag{4}$$

LDA emphasizes discrimination of data belonging to different classes and it does so by solving the generalized eigen-value problem:

$$\mathbf{\Sigma}_\text{B}\mathbf{v}_m = \lambda_m\mathbf{\Sigma}_\text{W}\mathbf{v}_m, \tag{5}$$

with $\mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_{\hat{M}}]$ for $\hat{M}$ largest eigen-values λ_m, and applying $\mathbf{V}$ as

$$\boldsymbol{\phi}_\text{LDA} = \mathbf{V}'\boldsymbol{\phi}. \tag{6}$$

Class separability for each basis is often expressed by the Fisher ratio and is equal to the basis' corresponding eigen-value.

2.1. Within-Class Covariance Correction

Let us decompose the within-speaker variability as

$$\mathbf{\Sigma}_\text{WS} = \mathbf{\Sigma}_\text{BD} + \mathbf{\Sigma}_\text{IS}, \tag{7}$$

where $\mathbf{\Sigma}_\text{BD}$ is the between-dataset covariance, and $\mathbf{\Sigma}_\text{IS}$ is the inter-session covariance, describing an average speaker variability within a dataset and assumed to be shared across datasets. It can be expressed as a within-class covariance where the classes are pairs of speaker and datasets (d, s):

$$\mathbf{\Sigma}_\text{IS} = \frac{1}{N}\sum_{d=1}^{D}\sum_{s=1}^{S_d}\sum_{n=1}^{N_{d,s}} N_{d,s}(\boldsymbol{\phi}_{n,d,s} - \boldsymbol{\mu}_{d,s})(\boldsymbol{\phi}_{n,d,s} - \boldsymbol{\mu}_{d,s})', \tag{8}$$

where D is number of datasets, s is a speaker instance for dataset d. Other variables have obvious meanings. We can decompose the total variability $\mathbf{\Sigma}_\text{T}$ as

$$\begin{aligned}\mathbf{\Sigma}_\text{T} &= \mathbf{\Sigma}_\text{BS} + \mathbf{\Sigma}_\text{WS} &(9)\\ &= \mathbf{\Sigma}_\text{BS} + \mathbf{\Sigma}_\text{BD} + \mathbf{\Sigma}_\text{IS}. &(10)\end{aligned}$$

Fig. 2 depicts this situation.

It is important to note that, if the speakers do not overlap across datasets, the "speaker" class will effectively (without any change in meaning) be understood as "speaker and dataset" class. As a result, $\mathbf{\Sigma}_\text{WS}$ will have the same form as $\mathbf{\Sigma}_\text{IS}$ in (8). The $\mathbf{\Sigma}_\text{BD}$ term from (10) will diminish and will be absorbed by $\mathbf{\Sigma}_\text{BS}$, incorrectly emphasizing discriminability power in LDA computation.

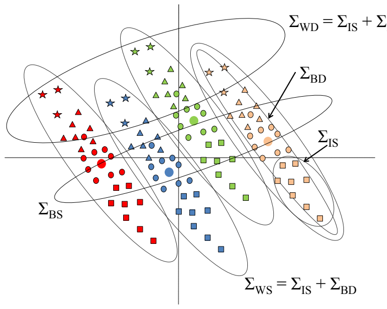
$\Sigma_\text{WD} = \Sigma_\text{IS} + \Sigma_\text{BS}$
Σ_BD
Σ_IS
Σ_BS
$\Sigma_\text{WS} = \Sigma_\text{IS} + \Sigma_\text{BD}$

Fig. 2. Illustration of decomposition of within-speaker covariance $\mathbf{\Sigma}_\text{WS}$ into inter-session covariance $\mathbf{\Sigma}_\text{IS}$, and between-dataset covariance $\mathbf{\Sigma}_\text{BD}$. The within-dataset covariance is decomposed into inter-session covariance and between-speaker covariance. Note that different colors represent different speakers and different shapes represent different datasets.

Note that, multi-modality over datasets does not necessarily have to be a problem if the speakers in the datasets overlap. In this case, the between-dataset variability is correctly included in the within-speaker (channel) covariance and (7) is satisfied.

Source Normalization (SN) [5, 6] solves this issue by computing separate between-speaker covariance matrices for each dataset and averaging these over all datasets: $\mathbf{\Sigma}_\text{BS} = \sum_{d\in\mathcal{D}}\mathbf{\Sigma}_{\text{BS},d}$, where $\mathcal{D}$ is a set of all datasets. This way, the data from different sources are effectively centered around a global mean and the between-dataset variability is compensated for in the between-speaker covariance computation.

Inter-dataset Variability Compensation (IDVC [7]) relies on a complete removal of known between-dataset variability by the means of NAP [3]. Note that, this requires an additional step in i-vector pre-processing.

In our work we use a very similar approach, i.e. we use the between-dataset covariance matrix $\mathbf{\Sigma}_\text{BD}$ (as in IDVC) to update our within-speaker variability (unlike updating the BS covariance in SN). To distinguish among the methods, we will refer to our approach as to Within-Class Correction (WCC). Note that the rank of $\mathbf{\Sigma}_\text{BD}$ matrix would typically be low, since the computation of the covariance is over means of data associated with the labeled datasets. Our assumption is that the directions of dataset shift are exclusive, meaning not carrying any useful speaker information and we can give them very high (negative) importance by scaling the $\mathbf{\Sigma}_\text{BD}$ up, so that the Fisher ratio in LDA for these directions will be very small. As a consequence, these directions will not be included the final LDA projection. This leads to an update formula

$$\mathbf{\Sigma}_\text{WS}^\text{(new)} = \mathbf{\Sigma}_\text{WS} + \alpha\mathbf{\Sigma}_\text{BD}^\text{(WCC)}, \tag{11}$$

where $\mathbf{\Sigma}_\text{BD}^\text{(WCC)}$ is the correction matrix estimated as a between-class covariance matrix with datasets as classes. In this case, however, instead of using the ML estimate (1), we equalize the number of data in each dataset, therefore

$$\mathbf{\Sigma}_\text{BD}^\text{(WCC)} = \frac{1}{C}\sum_{c=1}^{C}(\boldsymbol{\mu}_c - \hat{\boldsymbol{\mu}})(\boldsymbol{\mu}_c - \hat{\boldsymbol{\mu}})' \tag{12}$$

where $\hat{\boldsymbol{\mu}}$ is the equalized (not weighted) average of dataset means:

$$\hat{\boldsymbol{\mu}} = \frac{1}{C}\sum_{c=1}^{C}\boldsymbol{\mu}_c. \tag{13}$$

In our experiments we have seen that, the performance is very insensitive to the α constant once it is set to a large-enough number. Our typical approach was to set it to 10. Also note that, for the case when speakers do not overlap among datasets, $\boldsymbol{\Sigma}_{\text{BD}}$ is still absorbed by $\boldsymbol{\Sigma}_{\text{BS}}$, but by artificially adding it to $\boldsymbol{\Sigma}_{\text{WS}}$ with a large-enough weight, we make it disappear from the LDA projection. On the other hand, if the speakers do overlap across the datasets, the variability specified by $\boldsymbol{\Sigma}_{\text{BD}}$ is implicitly part of $\boldsymbol{\Sigma}_{\text{WS}}$ and, as we will see, adding it to $\boldsymbol{\Sigma}_{\text{WS}}$ ("again") does not hurt the performance.

2.2. Adaptation to New Dataset

The problem of dataset mismatch becomes even more serious when using an already built system on an unknown test data. As we will show in our experimental section, the degradation of recognition performance due to the dataset mismatch can be very significant.

Our main task was to develop an unsupervised technique, which would adapt the LDA matrix in order to compensate for the new dataset shift. The most straightforward way to do this is to re-estimate LDA using adapted within- or between- speaker covariance matrices $\boldsymbol{\Sigma}_{\text{WS}}$, $\boldsymbol{\Sigma}_{\text{BS}}$ in (5). Adapting the between-speaker covariance would require either speaker labels (which are not provided in unsupervised adaptation), or perform a speaker clustering—which is studied in [11, 12, 13]. Instead, we assume that our current estimate of the within-dataset (and thus between-speaker and inter-session) covariance is good enough and we will update the within-speaker covariance using a new estimate of the between-dataset covariance in the same way as was described in the previous section, i.e. using linear combination as in (11). The adaptation can be understood as within-class correction using new dataset directions.

Let us recall that the correction between-class covariance matrix is computed using (12). Identifying a new class, i.e. new dataset, transmission channel, domain, etc., we can compute the mean over its data-points and add it to collection of means in (12). Assuming that the adaptation data is a signle dataset, we can compute its mean and use it for WCC. We will show an improvement on the RATS task. Note that such operation requires re-estimation of the global mean $\hat{\boldsymbol{\mu}}$, however, we have found that re-estimation of $\hat{\boldsymbol{\mu}}$ had negligible impact on the result.

2.2.1. *Dataset Clustering*

Let us note, however, that the adaptation data can also suffer of multimodality. Assuming that the within-dataset covariance matrix is known and well-estimated, we can use a GMM with shared covariance—set to our within-dataset covariance—to cluster the adaptation data space. The means of the individual Gaussian components would be those added to the collection of means in (12). The clustering itself is not the main topic of this work, and we have used it rather empirically as natural choice, but, as we will see, it shows an interesting direction in future research.

3. EXPERIMENTS

We have carried out our experiments in two different scenarios. Originally, the task was to develop a system which would be easily adaptable to different transmission channels in the RATS problem, i.e. recognizing heavily distorted data, having limited amount of new-channel adaptation data. After seeing improvements in the RATS domain, we carried a set of experiments in the concurrently going research known as "Domain Adaptation Challenge" (DAC) [14, 11, 12, 13]. While in RATS the task was to adapt a system to an unseen transmission channel, in DAC the task was to adapt the system to an unknown domain simulated by having the Switchboard data collection in training and various MIXER databases in testing and adaptation sets.

From a dataset point of view, the main difference between these tasks is that in RATS we have multiple utterances per speaker across various channels, while in DAC, speakers do not overlap among the train datasets. As for the adaptation set, in RATS we have close to one utterance per speaker, while in DAC, there are multiple utterances per speaker.

3.1. RATS

The Linguistic Data Consortium (LDC) provided the training and test data for the RATS participants. The audio recordings were selected from existing and new data sources as follows: NIST SRE 2004 (Eng., Ara., Chin., Rus., Span.); RATS-LDC (Lev. Arabic, Farsi); RATS-Appen (Lev. Ara., Farsi, Pash., Dari, Urdu); Call-Friend Farsi; Fisher (Lev. Ara. and Eng.); NIST LRE (various languages).

All recordings were retransmitted through 8 different noisy communication channels, labeled by the letters A through H [15]. A "push-to-talk" (PTT) transmission protocol was used in all channels except G. PTT states produce some regions where multiple non-transmission (NT) segments may occur. As a result, the amount of usable audio decreases after retransmission.

It should be noted that among the data sources listed above, only the first three were annotated with speaker labels. Data from the other sources was used to train universal background models and i-vector extractors. We used the "dev" subset of the RATS-LDC and RATS-Appen corpora [1] to define speaker enrollment and testing samples. The rest of the RATS-LDC and RATS-Appen data, along with the NIST SRE 2004 set, was used for speaker modeling.

We have held out channel B data to serve as unknown-channel adaptation data (analogy to in-domain in the context of DAC). Channels C through H served as training data for all our hyper-parameter estimation. Let us note that speakers heavily overlap across the channels. We have excluded channel A data completely due to similarity of the A and B channels. We made sure the speakers do not overlap between the training and the adaptation sets. The adaptation set contains 2 hours of speech, 1164 utterances, and 1000 speakers (which is close to one utterance per speaker).

Trial-wise, the test protocol is defined as 6-conversational enrollment, and 1-conversation test. We have defined two test conditions based on the presence (semi-matched) or absence (unmatched) of unknown-channel utterance in enrollment—i.e. whether one of the 6 utterances comes from channel B. The test utterances come from the unknown channel.

We report our results at three operating points, as defined by the RATS, to target the extreme areas of the DET curves—false alarm probability at miss probability of 10%, miss probability at false alarm of 2.5%, and equal-error rate. The results are based on pools of male and female scores, however, only same-gender tests were performed.

3.1.1. *System Description*

We band-limit the audio to the 125-3750Hz range and extract 14 perceptual linear predictive (PLP) coefficients plus normalized energy using a 25ms Hamming window with a 10ms frame shift. We augment the PLPs with their first and second derivatives, yielding 45-dimensional feature vectors, which are then subjected to feature warping using a 3s sliding window over the detected speech regions.

The voice activity detection (VAD) is a variant of the GMM-based VAD described in [16]. RASTA-based [17] normalization was applied to PLPs. The 15-dimensional feature vector at each frame was augmented with the corresponding features from the preceding 15 and following 15 context frames, and then projected down

[1] LDC catalog ids: LDC2012E49, LDC2012E63, LDC2012E69.

Table 1. *Results of WCC for RATS. The asterisk records correspond to the trials containing the unknown channel in the enrollment.*

	fa@miss10	miss@fa2p5	EER
baseline	20.46	47.69	14.87
baseline*	15.70	40.90	12.66
WCC	20.70	47.81	14.85
WCC*	15.64	40.64	12.62
WCC adapt	15.61	34.10	12.24
WCC adapt*	6.82	21.40	8.07
known data	3.66	13.31	6.26
known data*	4.51	17.17	6.58

to 45 dimensions using heteroscedastic linear discriminant analysis (HLDA). Two 2048-component GMMs (speech/non-speech) were trained in the resulting feature space so as to maximize the mutual information between the training observations and their respective speech/non-speech labels.

We gender-independent UBM-GMM with 2048 components and we used 400-dimensional i-vector extractor, further reduced to 200 dimensions using LDA. Trial scoring was performed by the means of PLDA trained for full-rank speaker- and channel-covariance matrices. For a detailed description, see [18].

3.1.2. Results

Tab. 1 shows the overall results. The asterisk (*) marks the semi-matched condition. The first line shows the results of the system with no WCC applied. The second set shows the case, when WCC was computed using the training data only—we see no improvements as the speakers among the datasets overlap. The third set shows the case when a single mean—computed on the adaptation data—is added to the collection of means for BD covariance computation. Note, that this is a mere rank-1 update of the LDA within-class covariance matrix and a similar result was obtained, when only $\alpha(\boldsymbol{\mu}_{\text{adapt}} - \hat{\boldsymbol{\mu}})(\boldsymbol{\mu}_{\text{adapt}} - \hat{\boldsymbol{\mu}})'$ was added to the original unadapted within-class covariance matrix. For reference, we provide the results on how the system performs on data that match the training set (channels C-H, "known data"). Dataset clustering degraded performance, therefore it is not included in the results.

3.2. "Domain Adaptation Challenge"

DAC protocol is based on the Linguistic Data Consorotium's (LDC) telephone corpora: the MIXER and the Switchboard data sets. The test set is based on the NIST SRE 2010 telephone data, condition 5 (normal vocal effort). The training portion of the corpus is divided into two subsets: a) the *in-domain* (referred to as SRE) part is a collection of all telephone conversations from all speakers from NIST's SRE 2004 through 2008 task (which is believed to match the SRE 2010 data) and is used for system-parameter adaptation, and b) the *out-of-domain* (referred to as SWB) part based on the Switchboard collections (Phase 1, Phase 2, Cellular) is used for the base hyper-parameter training. There were 1461 male, and 1653 female speakers in 32921 SWB files (in average 10.6 files per speaker), and 1531 male and 2276 female speakers in 36498 SRE files (in average 9.6 files per speaker).

3.2.1. System Description

We used cepstral features, extracted using a 25 ms Hamming window. 19 Mel frequency cepstral coefficients together with log energy were calculated every 10 ms. This 20-dimensional feature vector was subjected to short time Gaussianization [19] using a 3 s sliding window. Delta and double delta coefficients were then calculated using a five-frame window giving a 60-dimensional feature vector.

Table 2. *Results on Domain Adaptation Challenge.* * *indicates WCC was perfomed on the train data (no adaptation).*

	DCF_{new}	DCF_{old}	EER
baseline	0.7412	0.3836	9.94
WCC*	0.6408	0.2368	5.04
WCC 1G	0.7233	0.3329	7.93
WCC 2G	0.7181	0.3315	8.07
WCC 4G	0.5474	0.1682	3.57
WCC 8G	0.4811	0.1424	2.91
WCC 16G	0.4701	0.1386	2.91
WCC 32G	0.4589	0.1405	3.03
WCC oracle	0.3563	0.1181	2.41

Speech/silence segmentation is performed by the BUT Czech phoneme recognizer [20], where all phoneme classes are linked to the *speech* class. The recognizer was trained on the Czech CTS data, but we have added noise with varying SNR to the 30% of the database.

We used gender-independent, 2048G UBM with diagonal covariance matrices trained by subsequential Gaussian splitting. I-vector extractor's dimensionality was set to 600, with LDA reducing the dimensionality to 200. PLDA used full-rank speaker- and channel-covariance matrices.

3.2.2. Results

We report our min-DCFs, and EER results on the female portion of the test set, although the male condition follows similar trends. Tab. 2 shows the result when applying the techniques in this paper. Our baseline does not include any WCC. Correcting the within-class covariance using the training data labels with $\alpha = 10$, we already see a big improvement (WCC*), which correlates with the fact that the speakers among the datasets do not overlap. The classes for the BD covariance computation were set to each of the switchboard collection and gender, similarly as in [7], i.e. {m,f}×{sw1, sw2p1, sw2p2, sw2p3, swcell1, swcell2}, resulting in 12 classes. Adapting the system using the same approach as in the RATS case degraded the result (WCC 1G) w.r.t. the WCC. However knowing that the test data comes from different datasets (SRE04-08), we performed GMM clustering improving the result. We see that the two-Gaussian system still degrades the performance. Our hypothesis is that the Gaussian components got seriously misaligned in a saddle between two major data clusters defining wrong dataset directions. Finer clustering brought significant improvement to our system. Note that in the RATS experiments, such clustering only degraded the performance. "WCC oracle" shows the result when computing the update covariance as a within-speaker covariance using true speaker labels.

4. CONCLUSIONS

In this paper, we have shown a technique of within-class correction for Linear Discriminant Analysis estimation. We have shown that when correct dataset clustering is used, adapting the within-class covariance of LDA by low-rank between-dataset covariance matrix can lead to significant improvement of the system, namely up to 70% in the Domain Adaptation Task, and 17.5% and 36% relative in the RATS unmatched and semi-matched tasks, respectively. The dataset clustering problem gave us an interesting direction for future research.

5. REFERENCES

[1] P. Kenny, “Joint factor analysis of speaker and session variability : Theory and algorithms - technical report CRIM-06/08-13. Montreal, CRIM, 2005,” 2005.

[2] P. Kenny, G. Boulianne, P. Oullet, and P. Dumouchel, “Joint factor analysis versus eigenchannels in speaker recognition,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2072–2084, 2007.

[3] W.M. Campbell, D.E. Sturim, D.A. Reynolds, and A. Solomonoff, “SVM based speaker verification using a GMM supervector kernel and nap variability compensation,” in *Proceedings of ICASSP 2006*, may 2006, vol. 1, p. I.

[4] Patrick Kenny, “Bayesian speaker verification with heavy–tailed priors,” in *Proc. of Odyssey 2010*, Brno, Czech Republic, June 2010, http://www.crim.ca/perso/patrick.kenny, keynote presentation.

[5] M. McLaren and D. 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*, 2011, pp. 5456–5459.

[6] M. McLaren and D. 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.

[7] Hagai Aronowitz, “Inter dataset variability compensation for speaker verification,” in *Submitted to ICASSP, 2014.*

[8] S. J. D. Prince and J. H. Elder, “Probabilistic linear discriminant analysis for inferences about identity,” in *11th International Conference on Computer Vision*, 2007, pp. 1–8.

[9] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. PP, no. 99, pp. 1 –1, 2010.

[10] Daniel Garcia-Romero, “Analysis of i-vector length normalization in Gaussian-PLDA speaker recognition systems,” in *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, 2011.

[11] Stephen H. Shum, Douglas A. Reynolds, Daniel Garcia-Romero, and Alan McCree, “Unsupervised clustering approaches for domain adaptation in speaker recognition systems,” in *Submitted to IEEE Signal Processing Letter 2014*, 2014.

[12] Daniel Garcia-Romero and Alan McCree, “Supervised domain adaptation for i-vector based speaker recognition,” in *Submitted to ICASSP, 2014*, 2014.

[13] Daniel Garcia-Romero, Alan McCree, Stephen Shum, Niko Brummer, , and Carlos Vaquero, “Unsupervised domain adaptation for i-vector speaker recognition,” in *Submitted to ICASSP, 2014*, 2014.

[14] “Domain adaptation challenge,” http://www.clsp.jhu.edu/workshops/archive/ws-13/, 2014.

[15] K. Walker and S. Strassel, “The RATS radio traffic collection system,” in *ISCA Speaker Odyssey*, 2012.

[16] Tim Ng, Bing Zhang, Long Nguyen, Spyros Matsoukas, Karel Vesely, Pavel Matějka, Xinhui Zhu, and Nima Mesgarani, “Developing a speech activity detection system for the DARPA RATS program,” in *Proc. of Interspeech 2012*, Sept. 2012.

[17] H. Hermansky, “RASTA processing of speech,” *IEEE Trans. Speech and Audio Processing*, vol. 2, pp. 578–589, 1994.

[18] Oldrich Plchot, Spyros Matsoukas, Pavel Matejka, Najim Dehak, Jeff Ma, S. Cumani, O. Glembek, H. Hermansky, S.H. Mallidi, N. Mesgarani, R. Schwartz, M. Soufifar, Z.H. Tan, S. Thomas, B. Zhang, and X. Zhou, “Developing a speaker identification system for the DARPA RATS project,” in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 6768–6772.

[19] J. Pelecanos and S. Sridharan, “Feature warping for robust speaker verification,” in *Proceedings of Odyssey 2006: The Speaker and Language Recognition Workshop*, 2006, pp. 213–218.

[20] Pavel Matějka, Lukáš Burget, Petr Schwarz, and Jan Černocký, “Brno university of technology system for NIST 2005 language recognition evaluation,” in *Proceedings of Odyssey 2006: The Speaker and Language Recognition Workshop*, 2006, pp. 57–64.