Domain Mismatch Compensation for Speaker Recognition Using a Library of Whiteners

Elliot Singer, *Senior Member, IEEE*, and Douglas A. Reynolds, *Fellow, IEEE*

***Abstract*—The development of the i-vector framework for generating low dimensional representations of speech utterances has led to considerable improvements in speaker recognition performance. Although these gains have been achieved in periodic National Institute of Standards and Technology (NIST) evaluations, the problem of domain mismatch, where the system development data and the application data are collected from different sources, remains a challenging one. The impact of domain mismatch was a focus of the Johns Hopkins University (JHU) 2013 speaker recognition workshop, where a domain adaptation challenge (DAC13) corpus was created to address this problem. This paper proposes an approach to domain mismatch compensation for applications where in-domain development data is assumed to be unavailable. The method is based on a generalization of data whitening used in association with i-vector length normalization and utilizes a library of whitening transforms trained at system development time using strictly out-of-domain data. The approach is evaluated on the 2013 domain adaptation challenge task and is shown to compare favorably to in-domain conventional whitening and to nuisance attribute projection (NAP) inter-dataset variability compensation.**

***Index Terms*—Channel compensation, domain mismatch, i-vectors, whitening.**

I. Introduction

THE development of low dimensional vector representations of speech utterances has led to considerable improvement in the performance of speaker recognition systems submitted to the periodic National Institute of Standards and Technology (NIST) speaker recognition evaluations (SREs). In particular, the i-vector method, which models the speech utterance in a total variability subspace, has emerged as the predominant approach due to its state-of-the-art performance, low computational complexity, and compact representation [1].

For maximum effectiveness, the data used in the estimation of system hyperparameters (universal background model, total variability matrix, within and across class covariance matrices) should be matched to the domain of the application (i.e., the enroll and test data). Mismatches due to domain variability can lead to degradation in performance, as has been demonstrated by participants in the 2013 Johns Hopkins University (JHU) speaker recognition workshop [2]. Numerous approaches to domain mismatch in speaker recognition have been proposed, including classical cepstral standardization [3], feature mapping [4], source normalization techniques [5][6], inter-dataset variability compensation via nuisance attribute projection (NAP) [7][8], within-class covariance correction [9], and audio characterization [10]. This paper will present an alternative approach that generalizes the data whitening and vector length normalization method introduced in [11] and specifically addresses applications where in-domain development data is assumed to be unavailable. Rather than relying on a single whitening transform, a library of whitening transforms is created by partitioning a training set into subcomponents to provide a bank of whitening parameters that can be applied on an utterance-by-utterance basis using a maximum likelihood selection process.

The remainder of this paper is organized as follows: Section II describes related compensation methods that have been employed for dealing with mismatched domains, along with the proposed whitening library scheme. Section III describes the i-vector system and experimental corpus used in this study. Section IV compares performance of the library whitening method against a variety of baselines, and Section V concludes with a summary of the results and suggestions for additional work.

II. Background and Related Work

Reducing the impact of domain mismatch has received considerable study within the speaker recognition community. Representative techniques are described below.

A. Feature Standardization and Warping

Acoustic feature standardization is a common and long-standing method of normalization that was originally developed to provide noise robust speech recognition [12] and was then applied to speaker recognition [3]. The technique is applied by using parameters estimated over a sliding segmental window to normalize each element of a signal-processing front-end feature vector to have a zero mean and unit standard deviation. Note that this is equivalent to data whitening assuming that the feature vector elements are uncorrelated. This method was intended to reduce mismatches between the training and testing environments.

Feature warping [13] is an acoustic feature vector post-processing method designed to introduce robustness to linear channel mismatch and additive noise conditions. Robustness is achieved by warping individual cepstral feature streams over a short time interval to conform to a standard Gaussian target distribution. Unlike feature standardization, feature warping guarantees that the warped feature distributions are Gaussian,

Manuscript received May 22, 2015; revised June 29, 2015; accepted June 29, 2015. Date of publication July 01, 2015; date of current version July 09, 2015. This work was 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. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Daniel Povey.

The authors are with the MIT Lincoln Laboratory, Lexington, MA 02420 USA (e-mail: es@ll.mit.edu; dar@ll.mit.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2015.2451591

although this may lead to a suboptimal mapping when the real distribution is non-Gaussian [13].

B. Hyperparameter Model Conditioning

Linear discriminant analysis (LDA) is a common method used to identify a subspace onto which i-vectors are projected so as to simultaneously maximize speaker discrimination while reducing inter-session variability [1]. Several domain compensation techniques have been proposed that modify the conventional computation of the within-speaker and between-speaker scatter matrices used in applying LDA. In source normalization [5][6], the between-speaker scatter matrix is computed as the average of the individual in-domain between-speaker scatter matrices, thereby removing cross-domain bias from the final matrix and avoiding an overestimation of between-speaker variability. The within-speaker scatter is then computed as the difference between the total variability scatter and the between-speaker scatter, thus attempting to introduce cross-domain variability into the within-speaker scatter matrix. The authors in [5] showed that this method provided improved performance in the NIST SRE12 task, which contained cross-domain utterances from telephony, microphone, and interview recordings.

A related method, referred to as within-class covariance correction [9], relies on a conventional computation of the between-speaker scatter matrix but adds a correction term to the within-scatter matrix using the cross-domain scatter obtained from a source-labeled dataset. (This is similar to the matrix estimated in inter-dataset variability compensation below.) Improvements are reported for both the RATS task and the 2013 domain adaptation challenge data. Note that both source normalization and within-class covariance correction rely on the availability of a large speaker-labeled dataset (although not necessarily in-domain) to compute the LDA transformation.

C. Inter-dataset Variability Compensation

Inter-dataset variability compensation (IDVC) [7] was introduced during the JHU 2013 summer workshop as a means of conditioning i-vectors to compensate for dataset mismatches deliberately introduced into the domain adaptation challenge task (DAC13) [2]. As will be described in further detail in Section III, the DAC13 corpus consisted of a fully labeled Switchboard (SWB) dataset, an enroll and test dataset derived from MIXER data, and an unlabeled SRE dataset available for domain adaptation. Results presented in [7] show that domain mismatch compensation can be achieved without using any in-domain adaptation data by applying nuisance attribute projection (NAP) [14] to the i-vectors to remove known between-dataset variability. An extension of the IVDC concept, described in [8], showed additional improvements in performance on the DAC13 challenge task. This method was not evaluated in the present study.

D. Whitening and Length Normalization

Another i-vector conditioning technique is length normalization, which was introduced in [11] as a means of Gaussianizing i-vectors so that they conform to the Gaussian modeling assumptions made in probabilistic linear discriminant analysis (PLDA) [15][16]. The length-normalization approach requires that i-vectors be whitened before being projected onto the unit sphere. This relatively simple process allows conventional i-vector recognizers to match the performance of the more computationally demanding heavy-tailed PLDA model [15]. Length normalization is a common component of published state-of-the-art systems, although descriptions do not always make clear whether the i-vectors were properly whitened prior to normalization.

E. Whitening Library

Whitening is applied to the data via an affine transformation $y = S^{-1/2}(w - \overline{w})$ where w is the raw i-vector, $\overline{w}$ is the i-vector mean, S is the i-vector covariance matrix, and y is the whitened i-vector. Ideally, $\overline{w}$ and S are estimated from i-vectors derived from the same source as w. In practice, however, the data available to train the whitening parameters may be derived from a combination of sources making the underlying distribution multi-modal and thus leading to an improper whitening transformation and suboptimum Gaussianization via length normalization.

To reduce the effects of domain mismatch on i-vector length normalization, this paper proposes using a *whitening library* whose individual components are estimated from labeled source collections. Each i-vector w is compared against a source-specific Gaussian model $\mathcal{N}(w|\mu_j, S_j)$, where parameters (μ_j, S_j) are estimated from source-specific data. The parameters of the model that produces the maximum likelihood are used to whiten the i-vector w and applied via $y = S^{-1/2}(w - \mu_j)$, where

$$i = \underset{j \in (1,\ldots,K)}{\arg\max}\ \mathcal{N}(w|\mu_j, S_j)$$

III. Experiment Setup

A. Speaker Recognition System Description

All experiments were performed using an i-vector speaker recognition system. Mel-frequency cepstral coefficient (MFCC) features were computed from each audio frame using a mel-spaced filterbank comprising 20 triangular filters whose center frequencies spanned the range of 300-3140 Hz. Delta cepstra were computed over a ± 2 frame span and appended to the cepstral vector. Speech segments were then extracted using speech activity detection (SAD) based on GMMs trained on telephone speech, and each component of the resulting 40-dimensional feature vector was standardized to zero mean and unit variance over a 300 ms window. The acoustic feature vectors were converted to i-vectors using a 2048-order gender independent diagonal covariance universal background model (UBM) and a rank-600 total variability (T) matrix. The resulting i-vectors were subjected to various types of domain compensation as described below. Scoring against speaker models was performed using either the cosine distance or PLDA.

B. Corpus

The experimental corpus is taken from the domain adaptation challenge task (DAC13) that was developed for the JHU Center for Language and Speech Processing 2013 summer workshop [2], and was derived from LDC Switchboard (SWB) and MIXER (SRE 2004-2010) telephone data. The NIST SRE10 data and condition 5 protocols (core extended trial

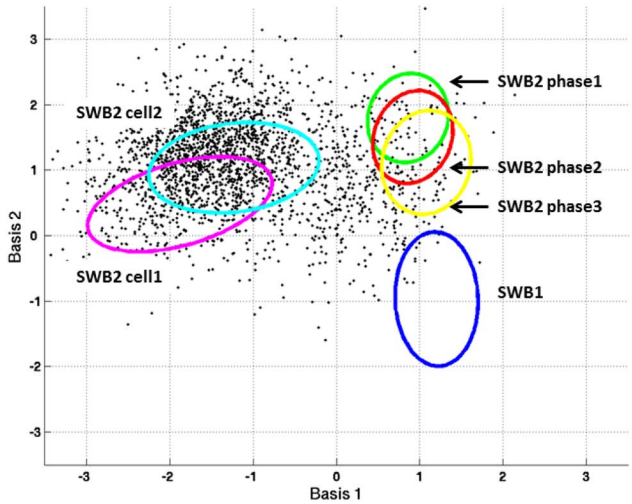

Fig. 1. Projections of the SWB and SRE10 i-vectors onto a basis formed from the between-dataset covariance of the labeled SWB collections. Ovals represent the equal probability contours of the 2-d projections of the individual SWB collections, and the scatter represents the distribution of the 2-d SRE10 projections.

lists involving normal vocal effort conversational telephone speech) were used for enrollment, scoring, and evaluation [17], while the SWB (out-of-domain) utterances comprised the development dataset. The SWB component, for which speaker labels were provided, was used to train the UBM, T-matrix, and PLDA within-speaker and between-speaker matrices. An additional in-domain development dataset derived from SRE 2004-2008 was made available to workshop participants for adaptation; however, none of this material was used in the current study other than to establish a fully-matched ("ideal") performance baseline.

Analysis of the DAC13 corpus provides additional motivation for the proposed strategy. A between-dataset PCA basis was derived from the i-vectors using the labeled collection types (SWB1; SWB2 phases 1 to 3; and SWB2 Cellular 1 and 2) available in the training data. These subsets generally reflect the chronological order in which the data was collected, beginning with landline calls in SWB1 and continuing through the SWB2 cellular collections. Fig. 1 shows contours of equal probability for distributions of the individual collections when mapped onto the first two principal components of the basis. It is readily apparent that the distributions of the individual collections that comprise the development set vary widely. Also shown is a scatter plot of the evaluation data (SRE10) similarly projected onto the between-dataset basis. Although the mismatch between the evaluation data and subsets of the development data is apparent from the figure, it is also clear that there is substantial overlap between the evaluation data and the SWB2 cellular subsets and that the DAC13 corpus does not reflect a fully mismatched scenario.

IV. Experiments and Results

A. System Configurations

Speaker recognition performance was compared using the following compensation methods:

- No whitening. (Unless otherwise stated the term "whitening" as employed here implies both centering and whitening.)

TABLE I
Performance Comparison (EER(%) and New and Old DCF [17]) Using PLDA Scoring. With the Exception of the Ideal (matched) Whitening Condition, only Mismatched Domain (SWB) Data was used to Train the Compensation Methods

COMPENSATION	EER (%)	DCF (old)	DCF (new)
None	6.30	0.4551	0.6560
Whiten mismatched	6.07	0.4528	0.6540
Whiten matched (ideal)	4.12	0.3325	0.5250
Library	3.86	0.3568	0.5880
IDVC [7]	4.13	0.3687	0.5700

- Whitening using mismatched out-of-domain (SWB) training data.
- Whitening using matched domain training data derived from the DAC13 in-domain development set. This represents the "ideal" whitening strategy and provides a baseline against which to compare compensation methods.
- Domain variability compensation using IDVC, as in [7]. The rank of the nuisance subspace is set to 10.
- Domain variability compensation using a whitening library, as proposed in Section II-E.

Inter-dataset variability compensation is applied following the approach of [7], where the SWB (out-of-domain) data is subdivided into partitions using the labeled collection types (SWB1; SWB2 phases 1-3; and SWB2 Cellular 1-2) and its gender- dependent subsets, resulting in a total of 12 domain classes. The first 10 principal components of the resulting scatter matrix are treated as nuisance directions and projected away from all i-vectors. The whitening library is constructed using gender-independent data from the six SWB collection partitions. Use of a gender dependent library did not improve performance. Results are presented for the combined male and female trials by means of equal error rates (EER), and minimum detection cost functions (DCF old and new) as described in [17].

B. Performance Comparison

Speaker recognition performance for the DAC13 evaluation set using PLDA scoring under the aforementioned conditions and summary statistics are shown in Table I. Whitening using data mismatched to the enroll and test sets (that is, data from SWB) has effectively no effect on performance. However, employing an adaptive whitening strategy using a whitening library constructed from the labeled dataset components of SWB produces performance that is roughly comparable to that of both ideal matched whitening and to the IDVC approach of [7].

C. Analysis

Fig. 2 shows the number of times each subset of the extended whitening library was selected for the SRE10 enroll and test utterances. Approximately 79% of the i-vectors best match the SWB2-Cell2 model, with the other 21% scattered among the remainder. Affinity of the SRE10 evaluation data to SWB cellular data is apparent from Fig. 1 and has been observed by others ([18], [19]); it may confidently be attributed to the high proportion of cell calls in SRE10. Regardless, the whitening selection method appears to identify the proper whitening parameters. A similar analysis of the SWB data used to compute the within-class and between-class covariance matrices shows that the whitener selected for data from each of the subsets will be

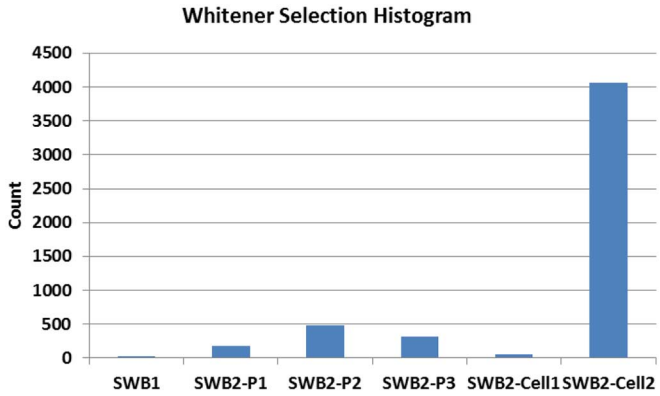

Fig. 2. Whitening library selections for the DAC13 evaluation data (5119 utterances).

TABLE II
PERFORMANCE COMPARISON (EER(%) AND NEW AND OLD DCF [17]) FOR DIFFERENT WHITENING PARAMETERS USING COSINE DISTANCE SCORING

WHITEN	EER (%)	DCF (old)	DCF (new)
SWB1	36.17	0.8987	0.9590
SWB2-p1	37.12	0.9031	0.9560
SWB2-p2	34.04	0.8947	0.9550
SWB2-p3	31.93	0.8859	0.9400
SWB2-Cell1	11.05	0.6856	0.8710
SWB2-Cell2	7.25	0.5209	0.7800
Combined	9.37	0.6210	0.7940
Library	7.06	0.5298	0.7570

properly selected (i.e., matched to the data source) over 99% of the time.

In an additional analysis, performance was evaluated using individual SWB subsets as whiteners and compared to using all of SWB and to the library approach. Results are presented using cosine distance scoring to avoid the complication of the varying training requirements for the within-speaker and between-speaker matrices for each condition. Table II once again indicates a similarity between the evaluation data and the SWB2-Cell2 subset. Also apparent is the degradation in performance observed when the whitener is improperly matched to the data, a phenomenon already observed in [20]. Unlike [20], where the whitening was manually adjusted to the data domain, the whitening library scheme proposed in this study allows the system to adapt the specific whitener to the data automatically.

V. CONCLUSIONS

This paper proposes an approach to domain mismatch compensation for applications where in-domain development data is assumed to be unavailable. Results indicate that a whitening strategy that makes use of individual whiteners selected from a variety of domains may be useful for reducing domain mismatches between system development training data and evaluation data. The utility of the approach was demonstrated on the DAC13 corpus, which was deliberately designed to introduce a training and evaluation domain mismatch, although it was noted that a portion of the development data (SWB2-Cell2) was indeed well matched to the SRE10 evaluation data. The library approach gives results that compare favorably to inter-dataset variability compensation [7] and to whitening using matched domain data.

Several follow-up studies suggest themselves. Applying the whitening library method does not preclude using other domain compensation techniques. Construction of the library made use of domain labels provided with the data distribution, and it would be of interest to evaluate its use on a large unlabeled set using blind clustering to select the library components. Finally, additional investigation is warranted using other mixed domain tasks (e.g., telephone/microphone) to further evaluate the utility of the proposed technique.

REFERENCES

[1] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 19, no. 4, pp. 788–798, May 2010.
[2] JHU 2013 Speaker Recognition Workshop [Online]. Available: http://www.clsp.jhu.edu/workshops/archive/ws13-summer-workshop/groups/spk-13/
[3] J. Koolwaaij and L. Boves, "Local normalization and delayed decision making in speaker detection and tracking," *Dig. Signal Process.*, vol. 10, no. 1/2/3, pp. 113–132, 2000.
[4] D. A. Reynolds, "Channel robust speaker verification via feature mapping," in *Proc. Int. Conf. Acoust., Speech, and Signal Process*, Hong Kong, 2003, pp. II:53–56.
[5] M. McLaren and D. van Leeuwen, "Source-normalized LDA for Robust Speaker Recognition Using I-Vectors From Multiple Speech Sources," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 20, no. 3, pp. 755–766, Mar. 2012.
[6] A. Kanagasundaram, D. Deana, S. Sridharan, M. McLaren, and R. Vogt, "I-vector based speaker recognition using advanced channel compensation techniques," *Comput. Speech Lang.*, vol. 28, pp. 121–140, 2014.
[7] H. Aronowitz, "Inter dataset variability compensation for speaker recognition," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, Florence, Italy, 2014, pp. 4002–4006.
[8] H. Aronowitz, "Compensating inter-dataset variability in PLDA hyperparameters for robust speaker recognition," in *Proc. IEEE Odyssey: Speaker Lang. Recognition Workshop*, Joensuu, Finland, 2014.
[9] O. Glembek, J. Ma, P. Matejka, B. Zhang, O. Plchot, L. Burget, and S. Matsoukas, "Domain adaptation via within-class covariance correction in i-vector based speaker recognition systems," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, Florence, Italy, 2014, pp. 4032–4036.
[10] L. Ferrer *et al.*, "A unified approach for audio characterization and its application to speaker recognition," in *Proc. IEEE Odyssey: Speaker Lang. Recognition Workshop*, Singapore, 2012, pp. 317–323.
[11] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proc. Interspeech*, Florence, Italy, 2011, pp. 249–252.
[12] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Commun.*, vol. 25, pp. 133–147, 1998.
[13] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *A Speaker Odyssey - The Speaker Recognition Workshop*, Crete, Greece, 2001.
[14] A. Solomonoff, W. M. Campbell, and C. Quillen, "Nuisance attribute projection," *Speech Commun.*, 2007.
[15] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Proc. IEEE Odyssey Speaker and Language Recognition Workshop*, Brno, Czech Republic, 2010.
[16] S. J. D. Prince, "Probabilistic linear discriminant analysis for inferences about identity," in *IEEE 11th Int. Conf. Computer Vision (ICCV)*, Rio de Janeiro, Brazil, 2007, pp. 1–8.
[17] NIST Year 2010 Speaker Recognition Evaluation Plan [Online]. Available: http://www.nist.gov/itl/iad/mig/upload/NIST_SRE10_evalplan-r6.pdf
[18] S. Biswas, J. Rohdin, and K. Shinoda, "I-vector selection for effective plda modeling in speaker recognition," in *Proc. IEEE Odyssey: Speaker Lang. Recognition Workshop*, Joensuu, Finland, 2014.
[19] S. H. Shum, D. A. Reynolds, D. Garcia-Romero, and A. McCree, "Unsupervised clustering approaches for domain adaptation in speaker recognition systems," in *Proc. Odyssey Speaker Lang. Recognition Workshop*, Joensuu, Finland, 2014.
[20] D. Garcia-Romero, A. McCree, S. Shum, N. Brümmer, and C. Vaquero, "Unsupervised domain adaptation for i-vector speaker recognition," in *Proc. IEEE Odyssey Speaker Lang. Recognition Workshop*, Joensuu, Finland, 2014.