INTER DATASET VARIABILITY COMPENSATION FOR SPEAKER RECOGNITION

Hagai Aronowitz

IBM Research - Haifa
Haifa, Israel
hagaia@il.ibm.com

ABSTRACT

Recently satisfactory results have been obtained in NIST speaker recognition evaluations. These results are mainly due to accurate modeling of a very large development dataset provided by LDC. However, for many realistic scenarios the use of this development dataset is limited due to a dataset mismatch. In such cases, collection of a large enough dataset is infeasible. In this work we analyze the sources of degradation for a particular setup in the context of an i-vector PLDA system and conclude that the main source for degradation is an i-vector dataset shift. As a remedy, we introduce inter dataset variability compensation (IDVC) to explicitly compensate for dataset shift in the i-vector space. This is done using the nuisance attribute projection (NAP) method. Using IDVC we managed to reduce error dramatically by more than 50% for the domain mismatch setup.

Index Terms— speaker recognition, robust speaker recognition, i-vector, inter dataset variability compensation, domain adaptation challenge

1. INTRODUCTION

Recent advances in speaker recognition, namely the introduction of i-vectors [1] and Probabilistic Linear Discriminant Analysis (PLDA) [2, 3] resulted in very low error rates in the recent NIST speaker recognition evaluations (SREs) [3]. However, the success of i-vector based PLDA is dependent on the availability of a large development set with thousands of multi session speakers. Moreover, the development data must match the evaluation data.

For domains that differ from the standard NIST SREs, the use of i-vector based PLDA is not so successful. For instance, for text-dependent speaker recognition it has been shown that the NAP framework [4] was more successful [5], unless an unrealistically large text-dependent development dataset is available [6].

In the summer of 2013, two speaker recognition workshops were concurrently held at the Johns Hopkins University (JHU) [12, 13]. The cross domain speaker recognition task was addressed in both workshops and was named the **domain adaptation challenge**. The challenge was motivated by preliminary experiments that showed that a PLDA system built on the Switchboard [7] corpus gave a 3 times larger equal error rate (EER) on the NIST 2010 SRE (condition 5), compared to a system built on a subset of the MIXER corpus (NIST 2004-2008 SREs).

The work reported in this paper was done at the JHU workshop in the framework of the domain adaptation challenge. The main goal addressed in this paper was improving the accuracy of a system built on Switchboard and evaluated on NIST 2010 SRE, without any adaptation stage (using MIXER) whatsoever.

The research challenge of coping with dataset mismatch is usually addressed using some amount of adaptation data [6, 8, 9]. In [10] source normalization (SN) was proposed to improve the robustness of i-vector-based speaker recognition for under-resourced and unseen cross-source evaluation conditions. The technique of source-normalization aims at removing the effect of cross-dataset variation from the estimate of the between-speaker covariance matrix, and pushing it into the estimate of the within-speaker covariance matrix. This makes the recognition system more robust to dataset mismatch. Contrary to SN, we aim at explicitly modeling dataset variation in the i-vector space and compensating it as a pre-processing cleanup step. We empirically compare our method to SN in Section 4 and discuss the differences between our proposed method and SN.

The rest of the paper is organized as follows: Section 2 provides an overview of the experimental setup. Section 3 motivates the proposed method and reviews related work. Section 4 describes the proposed method. Finally, Section 5 concludes.

2. EXPERIMENTAL SETUP

We use the official JHU 2013 speaker recognition workshop experimental setup. Following is a description of the datasets used, the speaker recognition system baseline, and the experimental protocol.

2.1. The SWB dataset

The SWB dataset consists of all telephone calls taken from Switchboard-I and Switchboard-II (all phases) corpora. This dataset serves as the mismatched development dataset. The dataset consists of 3114 speakers and 33039 sessions. For score normalization 2000 male sessions and 2000 female sessions were randomly selected.

2.2. The MIXER dataset

The MIXER dataset consists of a subset of telephone calls taken from SREs 2004-2008. For SRE 2008 interview data only is selected. This dataset serves as the matched development dataset. The dataset consists of 3790 speakers and 36470 sessions. For score normalization 2000 male sessions and 2000 female sessions were randomly selected.

2.3. The NIST-2010 dataset

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

The NIST 2010 SRE [11] condition 5 core extended trial list (single telephone conversations for both test and train with normal vocal effort) is used for evaluation. The dataset consists of 7169 target trials and 408956 impostor trials.

2.4. I-vector extractor

The i-vectors used in this work were created by the organizers of the workshop for the common use of the participants. A detailed description of the i-vector extractor is given in [14]. The i-vector extractor uses 40-dimensional MFCCs (20 base + deltas) with short-time mean and variance normalization. It uses a 2048 mixture gender independent (GI) UBM to obtain 600 dimensional GI i-vectors. The UBM and i-vector extractor were trained using the whole SWB dataset.

2.5. I-vector centering

A common technique for i-vector-based systems is to center the i-vectors of given datasets to a common center. In our baseline system we compute the center of the development data and use it to center both the development and evaluation data (the use of the center of the evaluation data is allowed only in the adaptation setup).

2.6. PLDA based back end

Prior to PLDA modeling [3], the dimensionality of the i-vectors is reduced using GI-LDA to 400. The next steps are within class covariance normalization (WCCN) [3] and length normalization [3]. Standard gender-dependent (GD) PLDA is then used with full between and within covariance matrices. Finally, ZT-norm for score normalization (which we found to outperform S-norm) is optionally performed.

2.7. The Domain robustness task

In the **domain robustness task** SWB is used for system building and NIST-2010 is used for evaluation. No use of MIXER (not even for i-vector centering) is allowed whatsoever.

2.8. The Domain unsupervised normalization task

In the **domain unsupervised normalization task**, SWB is used for system building. Unlabeled MIXER is used for normalization, and NIST-2010 is used for evaluation. Contrary to the domain adaptation challenge in which the MIXER data is assumed to contain many multi-session speakers, no such assumption is made in the normalization task. The normalization step may be followed by a domain adaptation step (involving clustering of the data into speakers and retraining the PLDA system as done extensively by other JHU 2013 participants) but this is not in the scope of this paper.

2.9. Evaluation measures

We report results by pooling male and female trials. For the main contribution we also report separate male and female results. Three error measures are used: EER, minDCF (old) and minDCF (new) – as specified in [11].

3. MOTIVATION AND RELATED WORK

Table 1 shows the degradation due to using the mismatched SWB dataset for building the PLDA system instead of using MIXER. EER increases by a factor of 3 (from 2.41% to 8.2%) and DCF increases very significantly too. Interestingly, when the i-vectors of the evaluation set (NIST 2010) are centered using the mean of the train set, EER drops from 8.2% to 4.58%. Furthermore, when the i-vectors of the train and test set are centered using the center of each set correspondingly, EER goes down to 3.96%. The fact that EER is cut by 50% by just doing proper centering (though breaking the NIST protocol) motivates our IDVC approach.

Table 1. A comparison of a system built on MIXER to systems built on SWB using different centering strategies. Results are for pooled male and female trials.

Devset	EER(in %)	minDCF(old)	minDCF(new)
MIXER	2.41	0.119	0.374
SWB	8.20	0.325	0.687
SWB center using train set	4.58	0.218	0.606
SWB center using train/test sets	3.96	0.189	0.546

Note that the dataset shift observed in Table 1 is different than standard intra-speaker variability. Standard within-speaker variability is assumed to distribute normally and is a source of variability between enrollment (train) and verification (test) data. Dataset shift may distribute quite differently than normal (as seen in Figure 1 below), and many times is a source of variability between the development data and the evaluation data, but not a major source of variability between enrollment and verification data (as can be seen in Table 1).

Note that the PLDA framework is much more vulnerable to dataset shift than the NAP framework. For instance, if both enrollment and verification sessions are shifted identically, the NAP framework is indifferent. However, in PLDA, length normalization and speaker space modeling (using the between speaker covariance matrix) are sensitive to dataset shift.

Furthermore, dataset shift may not be represented properly in the development data (when mismatch between development and evaluation data is much stronger than the internal mismatch within the development data), so standard modeling techniques may need some manual intervention or tuning.

Finally, contrary to standard within-speaker variability, dataset shift variability may be modeled without the need of a labeled multi-session multi-speaker dataset, as an unlabeled single-session multi-speaker dataset is enough to estimate dataset shift variability.

The above analysis gives a general motivation why the standard PLDA framework and natural extensions such as source normalization [10] do not optimally cope with dataset shift.

3.1. Source normalization

SN is an extension of the standard PLDA training framework which does take into account the fact that the development data may originate from several different sources, each one represented by a different dataset shift. SN modifies the PLDA training

framework in such a way that cleans up the estimation of the between-speaker covariance matrix from dataset variability and pushes it to the scope of the within-speaker covariance matrix. This is indeed an improved mean of modeling dataset variability but lacks some of the required properties listed above. In Section 4 we provide experimental results for the SN method in comparison to the proposed IDVC method.

3.2. Total variability subspace removal (TVSR)

TVSR is a technique that estimates a low dimensional subspace representation of the total variability space and removes that subspace as a preprocessing step.

TVSR was first proposed in [16] where it was named common speaker subspace removal. It was successfully used in [17, 8] for building a system with a development set consisting of single-session speakers only. In the NAP framework, TVSR was found to be beneficial even when multi-session speakers are available, both for text-dependent speaker verification [5] and for speaker recognition in summed (two-wire) conversations [18] where it was named two-wire-NAP.

The motivation for using TVSR is as follows: the low dimensional (~15) total variability subspace may consist of sources of variability such as gender variability and dataset or channel variability that are not properly represented in the multi-session speaker development data. Removing these sources of variability may improve the correctness of the Gaussianity assumptions in PLDA in mismatched conditions.

We investigated TVSR in the context of the domain adaptation challenge and found it to be beneficial. In Section 4 we provide experimental results for the TVSR method in comparison to the proposed IDVC method.

4. INTER DATASET VARIABILITY COMPENSATION

Inter dataset variability compensation aims at directly estimating and removing dataset shift in the i-vector domain. Dataset shift vectors are estimated for subsets of the development data (corresponding to different sources) and a low-dimensional subspace is estimated from these shift vectors. The estimated low-subspace is then removed from all i-vectors as a pre-processing step before PLDA training and scoring. The method and experimental analysis are described in detail in the following subsections.

4.1. Algorithm

Given a development dataset (such as SWB), the dataset is split into subsets according to available metadata (in principle, automatic clustering may also be used). In this work SWB was divided into 12 subsets (6 per gender). The subsets were defined according to the different LDC distributions (Table 2). Similarly, MIXER was partitioned into 8 gender dependent (GD) subsets for SRE 2004, 2005, 2006 and 2008 (interview).

For each subset all i-vectors are averaged and the resulting i-vector is the center of the subset. Given the set of 12 centers for SWB, principal component analysis (PCA) is used to find a basis for the subspace spanned by the 12 centers. This subspace is named the inter dataset variability subspace, and is removed from all the i-vectors for both the development and the evaluation data as a pre-processing stage. Figure 1 shows the first four dimensions of the projected centers of the subsets of SWB, MIXER and NIST-2010 in the dataset shift subspace. Note that the first eigenvector corresponds to gender.

Table 2. SWB is partitioned into 6 subsets. Each subset is then partitioned into two GD subsets.

Code	Description
97S62	SWB-1 Release 2
98S75	SWB-2 Phase I
99S79	SWB-2 Phase II
2001S13	SWB Cellular Part 1
2002S06	SWB-2 Phase III
2004S07	SWB Cellular Part 2

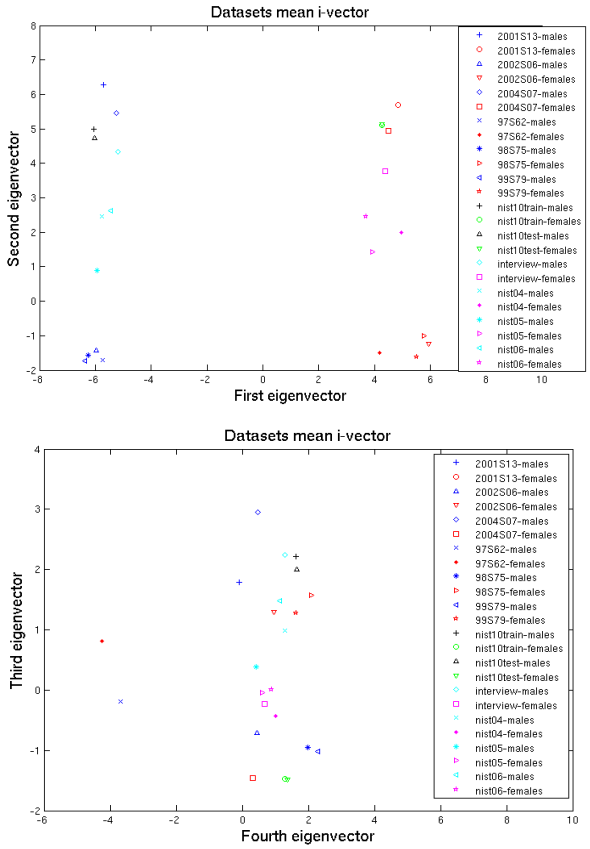

Figure 1: *Projection of the subset-based centers of SWB, MIXER and the evaluation set on the first four eigenvectors of the estimated dataset shift subspace.*

4.2. Results

Table 3 reports the results using IDVC for a PLDA system built on SWB only (named the domain robustness task). Results are without score normalization. We use the whole subspace spanned by the 12 centers for IDVC removal. We observe a 54% reduction in EER for pooled male and female trials. Note that the improvement for females (59%) is larger than for males (48%). For minDCF (old) cost reduction is 41% in average. For minDCF

(new), cost reduction is 22% in average.

Table 4 reports our investigation of the use of IDVC when the unlabeled MIXER data is available for unsupervised normalization. We name this task as the **domain unsupervised normalization task**. We use the term normalization and not adaptation in order to emphasize the lack of an assumption that the MIXER data contains multi-session speakers and the lack of use of clustering techniques. We use the MIXER data jointly with SWB to estimate the IDVC subspace. We use the whole subspace spanned by the 20 centers (12 SWB + 8 MIXER) for IDVC removal. We also explore the possibility of using the MIXER data for score normalization (which was found in [15] to be beneficial for the domain adaptation challenge). We observe that using MIXER jointly with SWB for IDVC estimation gives a small improvement. We also observe that score normalization gives a large improvement for the baseline system, but only small improvements (if any) in conjunction with IDVC.

Next, we compare the IDVC method to other techniques in the framework of the domain robustness task (Table 5) and the framework of the domain unsupervised normalization task (Table 6). Finally we report the results of using IDVC when training the system on the labeled MIXER. As can see in Table 7, IDVC slightly degrades performance for the matched domain task.

Table 3. Results using IDVC with a PLDA system built on SWB (without score normalization).

Eval dataset	IDVC training dataset	EER (in %)	minDCF (old)	minDCF (new)
All	-	8.20	0.325	0.687
	SWB	3.75	0.192	0.533
Males	-	6.55	0.299	0.640
	SWB	3.43	0.165	0.462
Females	-	9.75	0.342	0.706
	SWB	4.00	0.210	0.581

Table 4. Results on pooled male and female trials using IDVC with a PLDA system built on SWB. The use of MIXER for score normalization is explored.

IDVC training dataset	Score normalization	EER (in %)	minDCF (old)	minDCF (new)
-	-	8.20	0.325	0.687
SWB		3.75	0.192	0.533
SWB+MIXER		3.48	0.169	0.520
-	MIXER	5.87	0.227	0.715
SWB		3.53	0.170	0.521
SWB+MIXER		3.42	0.165	0.541

Table 5. Results for the domain robustness task. IDVC outperforms all other methods.

System	EER (in %)	minDCF (old)	minDCF (new)
Baseline	8.20	0.325	0.687
TVSR	7.42	0.301	0.669
SN	5.33	0.254	0.640
IDVC	3.75	0.192	0.533

Table 6. Results for the domain unsupervised normalization task. IDVC outperforms all other methods.

System	EER (in %)	minDCF (old)	minDCF (new)
Baseline	8.20	0.325	0.687
Centering and score normalization using MIXER	5.84	0.223	0.631
SN	4.73	0.206	0.581
TVSR	4.05	0.176	0.549
IDVC	3.48	0.169	0.520

Table 7. Results on pooled male and female trials using IDVC on a PLDA system built on MIXER.

IDVC training dataset	EER (in %)	minDCF (old)	minDCF (new)
-	2.41	0.119	0.374
SWB	2.45	0.122	0.403
SWB+MIXER	2.60	0.125	0.395

5. CONCLUSIONS

The inter dataset variability compensation technique for speaker recognition (using i-vector PLDA) was introduced and analyzed in this study. The shortcomings of the standard i-vector PLDA algorithm for coping with dataset mismatch were illustrated. Contrary to standard within-speaker variability, dataset shift may be highly non-Gaussian. Furthermore, in many cases dataset shift is a source of variability between development and evaluation data but not between enrolment and verification data.

IDVC has shown to effectively reduce the influence of dataset variability on the investigated i-vector PLDA system in the context of the domain adaptation challenge. When evaluated on a system trained on the Switchboard corpus, EER was decreased by 54%, DCF (old) by 41% and DCF (new) by 22%. When unlabeled MIXER data is used for adaptation, some more gains are achieved.

An unexplored additional use of the IDVC framework would be for robust data shift adaptation (centering) using a small amount of unlabeled adaptation data.

6. ACKNOWLEDGEMENTS

This work was part of the SMART EU project, partly funded by the European Commission in the scope of the 7th ICT framework.

7. REFERENCES

[1] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis For Speaker Verification," IEEE *Trans. on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788 - 798, 2010.

[2] S. J. D. Prince, "Probabilistic linear discriminant analysis for inferences about identity", in Proc. *International Conference on Computer Vision (ICCV)*, 2007.

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

[4] A. Solomonoff, W. M. Campbell, and C. Quillen, "Nuisance Attribute Projection", *Speech Communication*, Elsevier Science BV, 1 May, 2007.
[5] H. Aronowitz, R. Hoory, J. Pelecanos, D. Nahamoo, "New Developments in Voice Biometrics for User Authentication", in Proc. *Interspeech*, 2011.
[6] H. Aronowitz, O. Barkan, "On Leveraging Conversational Data for Building a Text Dependent Speaker Verification System", in Proc. *Interspeech*, 2013.
[7] The Lingusitic Data Consortium (LDC) catalog. Available online: http://catalog.ldc.upenn.edu/project_index.jsp
[8] H. Aronowitz, "Text Dependent Speaker Verification Using a Small Development Set", in Proc. *Speaker Odyssey*, 2012.
[9] M. Senoussaoui, P. Kenny, N. Dehak, and P. Dumouchel, "An i-vector Extractor Suitable for Speaker Recognition with both Microphone and Telephone Speech", in Proc *Speaker Odyssey*, 2010.
[10] M. McLaren and D. van Leeuwen, "Source-normalized LDA for robust speaker recognition using i-vectors from multiple speech sources", IEEE *Trans. Audio, Speech and Language Processing*, 20(3):755–766, march 2012.
[11] NIST 2010 SRE evaluation plan. Available online: http://www.nist.gov/itl/iad/mig/upload/NIST_SRE10_evalplan-r6.pdf.
[12] JHU 2013 speaker recognition workshop. Available online: http://speech.fit.vutbr.cz/short-news/world-top-scientists-speaker-recognition-together-again-year-johns-hopkins.
[13] JHU SCALE 2013 workshop: Robust Speaker Recognition for Real Data. Available online: http://hltcoe.jhu.edu/research/scale-workshops/.
[14] D. Garcia-Romero and A. McCree, "Supervised domain adaptation for i-Vector based speaker recognition", submitted to *ICASSP*, 2014.
[15] H. Aronowitz, "PLDA-based speaker recognition in a mismatched domain", submitted to *ICASSP*, 2014.
[16] H. Aronowitz, "Speaker recognition using Kernel-PCA and Intersession Variability Modeling", in Proc. *Interspeech*, 2007.
[17] D. Matrouf, J.-F. Bonastre, S. E. Mezaache, "Factor analysis multi-session training constraint in session compensation for speaker verification", in Proc. *Interspeech*, 2008.
[18] Y. A. Solewicz, H. Aronowitz, "Two-Wire Nuisance Attribute Projection", in Proc. *Interspeech* 2009.