Inter dataset variability compensation for robust speaker recognition

Hagai Aronowitz

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

Abstract

Recent advances in speaker recognition, namely the introduction of i-vectors and Probabilistic Linear Discriminant Analysis (PLDA) resulted in low error rates in the recent NIST speaker recognition evaluations (SREs). However, the success of i-vector based PLDA is dependent on the availability of a large development set. 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 was more successful, unless an unrealistically large text-dependent development dataset is available.

In the summer of 2013, a speaker recognition workshops was held at the Johns Hopkins University (JHU). The cross domain speaker recognition task was addressed in the workshop and was named the domain adaptation challenge. The challenge was motivated by preliminary experiments that showed that a PLDA system built on the Switchboard 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).

In this work reported we address the domain adaptation challenge. We introduce a novel method named IDVC (inter dataset variability compensation) which explicitly learns inter dataset variability from heterogonous training data. The learnt variability is used to improve the robustness of the i-vectors with respect to domain mismatch. The proposed approach obtains an error reduction of 62% for the 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 for estimating the PLDA hyper-parameters. Moreover, the development data must be matched to the evaluation data.

For domains that differ from the standard NIST SREs, the use of i-vector based PLDA is less 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 (DAC). The challenge was motivated by preliminary experiments that showed that a PLDA system built on the Switchboard [7] corpus had 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 in the framework of the DAC. 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 has been previously addressed using some amount of adaptation data [6, 8, 9]. In contrast to these methods, source normalization (SN) [10] and inter dataset variability compensation (IDVC) [15] do not require adaptation data to be effective. SN addresses the case when the development data originates from several sources but most speakers lack samples from each source. This causes dataset differences to be captured as inter-speaker variability rather than within-speaker (channel) variability. IDVC aims at explicitly modeling dataset shift variability in the i-vector space and compensating it as a pre-processing cleanup step, and was shown in [15] to outperform SN in the framework of the DAC.

In this paper we investigate types of mismatch other than mere shifts in i-vector space, and generalize the IDVC method to cope with dataset variability in additional hyper-parameters of the PLDA model, namely the within-speaker covariance matrix and the between speaker covariance matrix. We present empirical results that indicate that compensating the variability of these hyper-parameters across datasets improves robustness to dataset mismatch under the DAC setup.

The rest of the paper is organized as follows: Section 2 provides an overview of the experimental setup. Section 3 describes the proposed method. Section 4 reports the experiments and results. In Section 5 we discuss the results and compare to related work. Finally, Section 6 concludes.

2. Experimental setup

We use the JHU-2013 speaker recognition workshop DAC setup which can be downloaded from [12]. 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. This dataset serves as the mismatched development dataset. The dataset consists of 3114 speakers and 33039 sessions.

2.2. The MIXER dataset

The MIXER dataset consists of a subset of telephone calls taken from SREs 2004-2008. For SRE 2008 only interview data is included. This dataset serves as the matched development dataset. The dataset consists of 3790 speakers and 36470 sessions.

2.3. The NIST-2010 dataset

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, and may be downloaded from [12]. 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 standard processing method for i-vector-based systems is to center the i-vectors of given datasets at the origin. In our baseline system we compute the center of the development data and use it to center both the development and evaluation data.

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 rank between and within covariance matrices.

2.7. The Domain robustness task

In the domain robustness task, SWB is used for system training and NIST-2010 is used for evaluation. No use of MIXER (not even for i-vector centering) is allowed whatsoever. In this work we address the domain robustness task. Note that most other works addressing the DAC use the unlabeled MIXER for i-vector centering and unsupervised adaption of the PLDA system.

2.8. Evaluation measures

We report results by pooling male and female trials. Three error measures are used: EER, minDCF (old) and minDCF (new) as specified in [11].

3. Inter Dataset Variability Compensation

IDVC aims at estimating and removing dataset mismatch in the i-vector domain. This is done by first partitioning the development data into subsets corresponding to different sources, and then training a PLDA model for each subset. The variability in the PLDA hyper-parameters across the subsets is analyzed and a low-dimensional subspace in i-vector space accounting for most of that variability is pursued. The estimated low-subspace is then removed from all i-vectors as a pre-processing step before other processing (such as length normalization, LDA), PLDA training and scoring. The method is described in detail in the following subsections.

3.1. Two covariance model

The PLDA framework assumes that the i-vectors distribute according to Equation (1):

$$\phi = \mu + s + c \tag{1}$$

where ϕ denotes an i-vector, s denotes a speaker component, c denotes a channel (or within-speaker variability) component, and μ denotes the center of the i-vector space. Components s and c are assumed to distribute normally with zero mean and covariance matrices B (between speaker) and W (within-speaker) respectively.

The PLDA model is thus parameterized by $\{\mu, B, W\}$ and the goal of any PLDA training or adaptation algorithm is to estimate (or adapt) these hyper-parameters.

3.2. Motivation for the proposed method

We hypothesize that some directions in the i-vector space are more sensitive to dataset mismatch than other directions. In order to make a PLDA system robust to dataset mismatch, we aim at finding and removing a low-dimensional subspace which is spanned by directions in i-vector space which are relatively sensitive to dataset mismatch.

For a homogenous development dataset, it is unclear if the mismatch-sensitive subspace can be estimated without any use of adaptation data. However, in our setup (SWB as a development set) and in many other setups, the development dataset is heterogeneous (it contains for instance both landline and cellular data) and consists of relatively homogenous subsets. These relatively homogenous subsets of the development dataset may be used to estimate the PLDA hyper-parameters for each subset independently, and the mismatch-sensitive subspace may be estimated from the collection of PLDA models.

The hope (verified in the experimental section) is that the subspace estimated from the development dataset can be generalized to the unseen evaluation dataset.

3.3. Outline of the proposed method

Following is an outline of the proposed method. Details are given in following subsections.

Inter-dataset variability subspace estimation

1. Partition the development dataset into n homogenous subsets.

2. Estimate PLDA hyper-parameters $\{\mu_i, B_i, W_i\}$ for each subset *i*.
3. Estimate i-vector subspace S_μ corresponding to the set $\{\mu_i\}$
4. Estimate i-vector subspace S_W corresponding to the set $\{W_i\}$
5. Estimate i-vector subspace S_B corresponding to the set $\{B_i\}$
6. Join subspaces to form a single subspace: $S = S_\mu \cup S_W \cup S_B$

PLDA training

1. Remove subspace *S* from the i-vectors of the development set.
2. Train PLDA using the standard scheme.

PLDA scoring

1. Remove subspace *S* from the i-vectors of the evaluation set.
2. Score using the standard scheme.

3.4. Defining homogenous subsets

Given a development dataset (such as SWB), the dataset is split into subsets according to available metadata. In this work SWB was divided into 12 subsets (6 per gender). The subsets were defined according to the different LDC distributions (Table 1). Similarly, for some additional experiments, MIXER was also partitioned into 8 gender dependent (GD) subsets for SRE 2004, 2005, 2006 and 2008 (interview).

Table 1. 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

3.5. Estimation of the PLDA hyper-parameters for each subset

PLDA hyper-parameters are estimated independently for each subset using the standard method [2] . Limited training data may become an issue as data is partitioned. In our setup, it is no longer possible to reliably estimate the full rank B matrices without some sort of smoothing. We therefore smooth the estimation of the B matrix by linearly interpolating with its estimated diagonal giving the diagonal a weight of 0.1.

3.6. Estimation of i-vector subspace S_μ

As done in [15] we apply Principal Component Analysis (PCA) to the set of vectors $\{\mu_i\}$. The choice of the dimension of the subspace is investigated in the experimental section.

3.7. Estimation of i-vector subspaces S_W and S_B

The following subsection addresses the estimation of S_W. The estimation of S_B is done in a similar manner. For a set of *n* covariance matrices $\{W_i\}$ we denote the mean of the set by W. We use the following recipe for estimating subspace S_W.

Estimating subspace S_W of dimension *d*:

1. Whiten the i-vector space with respect to matrix W by applying $W_i \rightarrow TW_iT^t$ with $TT^t=W^{-1}$
2. Compute $\Omega = \frac{1}{n}\sum W_i^2$
3. Find the *k* largest eigenvalues of Ω. The corresponding eigenvectors span subspace S_W

The motivation for the proposed scheme is that after whitening with the mean within covariance W, the within-speaker variance along every axis equals to one for the whitened mean within covariance W. For whitened covariance matrix W_i, the variance along an axis defined by unit vector *v* equals to v^tW_iv. The quantity we choose to maximize is the variance of the variances along the axis. This is formulated in Equation (2):

$$\begin{aligned} v^* &= \arg\max \mathrm{var}\left(v^t W_i v\right) \\ &= \arg\max\left\{\frac{1}{n}\sum_i \left(v^t W_i v\right)^2 - 1\right\} \\ &= \arg\max v^t\left(\frac{1}{n}\sum_i W_i^2\right)v \end{aligned} \quad (2)$$

The solution for Equation (2) is the first eigenvector of matrix $\frac{1}{n}\sum_i W_i^2$.

3.8. IDVC with unlabeled data

Estimation of IDVC for the W and B hyper-parameters requires the availability of speaker labels. However, a common use case for IDVC would be when labeled data is available for PLDA training from one source only, and the multi-source data available for IDVC training is unlabeled. In this case we propose to apply IDVC on the μ hyper-parameter and on the total covariance matrix denoted by T which is the covariance of the i-vectors distribution.

4. Experiments and results

4.1. Baseline results

The effect of dataset mismatch is illustrated in Table 2 which shows the degradation due to using mismatched data for estimating the PLDA hyper-parameters. The degradation due to mismatch in estimating μ is up to 50% relative when B and W are estimated on SWB. The degradation due to mismatch in estimating B and W is up to a factor of 3 when μ is mismatched and a factor of 2 when μ is matched.

4.2. IDVC applied to the μ hyper-parameter

The effect of IDVC applied to the μ hyper-parameter is reported in Table 3. The results indicate ~50% relative reduction in EER and minDCF(old), and a very significant reduction in minDCF(new).

Table 2. The effect of dataset mismatch for estimating PLDA hyper-parameters. Results are for pooled male and female trials.

W and B	μ	EER (in %)	minDCF (old)	minDCF (new)
SWB	SWB	8.20	0.325	0.687
	MIXER	7.03	0.297	0.676
	NIST-10 training data	4.58	0.218	0.606
	NIST-10[1]	3.96	0.189	0.546
MIXER	MIXER	2.41	0.119	0.374
	NIST-10 training data	2.30	0.110	0.345
	NIST-10	2.27	0.110	0.346

4.3. IDVC applied independently for the W and B hyper-parameters

The effects of IDVC applied independently for the W and B hyper-parameters are reported in Table 3.

The results indicate quite similar results for applying IDVC for both the W and B hyper-parameters. The results indicate even larger accuracy improvements compared to those achieved for the μ hyper-parameter.

4.4. IDVC applied jointly for the μ, W and B hyper-parameters

The effect of IDVC applied jointly for the μ, W and B hyper-parameters is reported in Table 3. The results show that additional reduction in error can be obtained by applying IDVC on all three hyper-parameters.

Table 3. Selected results for IDVC applied on the hyper-parameters μ, B and W. Training of PLDA and IDVC is on SWB. The first three columns list subspace dimension.

μ	W	B	EER (in %)	minDCF (old)	minDCF (new)
0	0	0	8.20[2]	0.325	0.687
10	0	0	3.75[3]	0.192	0.533
0	100	0	3.37	0.155	0.496
0	0	100	3.53	0.164	0.510
10	50	0	3.09	0.138	0.454
10	0	50	3.39	0.154	0.469
10	30	30	3.15	0.138	0.463

4.5. Exploring different data partitions for IDVC training

The experiments reported so far were made using a partition of SWB into 12 subsets, following the work in [15]. In this subsection we investigate the sensitivity of the IDVC algorithm to other partitions which may represent other useful setups.

We denote the partition used so far by GD-12. We consider two additional partitions. First, we create a gender independent version of GD-12, denoted by GI-6. Second, we select from GI-6 two subsets (SWB Cellular Part 2 and SWB-1 Release 2) and create a partition containing these two subsets only. We denote this partition by GI-2. We repeat selected experiments training IDVC on the two new partitions.

Table 4 reports the results for IDVC training on the 3 different partitions. For μ-based IDVC, the full partition GD-12 is better than GI-6 and clearly outperforms GI-2. However, for IDVC applied on the hyper-parameters sets {μ, W} and {μ, W, B} the differences in accuracy between the partitions are quite small.

These results indicate that IDVC was able to successfully estimate the IDVC subspace from 2 data subsets only.

Table 4. Selected results for IDVC applied on the hyper-parameters μ, B and W. Training of PLDA and IDVC is on SWB. Three different partitions for IDVC training are evaluated.

Partition	μ	W	B	EER (in %)	minDCF (old)	minDCF (new)
Baseline	0	0	0	8.20	0.325	0.687
GD-12	10	0	0	3.75	0.192	0.533
GI-6	5	0	0	4.22	0.197	0.562
GI-2	1	0	0	5.28	0.231	0.612
GD-12	10	50	0	3.09	0.138	0.454
GI-6	5	50	0	3.05	0.141	0.450
GI-2	1	50	0	3.38	0.154	0.487
GD-12	10	30	30	3.15	0.138	0.463
GI-6	5	30	30	3.00	0.138	0.466
GI-2	1	30	30	3.30	0.152	0.494

4.6. IDVC applied without speaker labels

When speaker labels are unavailable IDVC may be applied for the μ and T hyper-parameters. Table 5 reports results for SWB training (PLDA and IDVC) on the GI-2 partition. Comparison of Tables 4 and 5 shows that replacing B and W with T results in comparable error rates.

Table 5. Selected results for IDVC applied on the hyper-parameters μ, and T on the GI-2 partition. Training of PLDA and IDVC is on SWB.

μ	T	EER (in %)	minDCF (old)	minDCF (new)
1	0	5.28	0.231	0.612
1	25	3.49	0.146	0.477

5. Discussion

Standard PLDA addresses two types of variability: between speaker variability and within-speaker variability. Gaussian-PLDA which is currently the state-of-the-art assumes that both of these types of variability are multivariate Gaussian distributions.

For heterogeneous data, the Gaussian assumptions are inappropriate. Real life data is usually heterogeneous and may contain different channels (landline, cellular, microphone), different audio durations, different textual content (for text dependent speaker verification), etc.

Source normalization [10] and the preliminary IDVC work [15] have addressed dataset shifts differently. SN estimates the

[1] NIST-10 training data is used to center the training i-vectors, and NIST-10 test data is used to center the test i-vectors.

[2] Baseline results without using IDVC

[3] Result from our previous ICASSP paper [15]

inter-dataset covariance in i-vector space from labeled subsets of the data and transfers it from the between speaker covariance matrix to the within-speaker covariance matrix. IDVC [15] estimates the same inter-dataset covariance matrix in i-vector space and removes the subspace containing that variability. IDVC is therefore more effective when the evaluation dataset is highly mismatched to the development dataset.

In this work we extended our original IDVC work to cope with dataset mismatch that is not purely an additive shift in i-vector space. We look for directions in the i-vector space that have between or within-speaker variances that highly differ across subsets of the development set. These directions are evidently not robust to dataset mismatch and we therefore remove them. More generally, we strive at removing directions in i-vector space that contribute to high variability in PLDA hyper-parameters across different datasets.

The tuning of our proposed method is still an open issue. It is clear from our experiments that a good configuration for the highly mismatched SWB-based PLDA system is not suitable for the reasonably matched MIXER-based PLDA system. This issue will be a topic for future research.

Finally, we have shown that IDVC applied for the W and B or T hyper-parameters can be sufficiently estimated from two subsets only. This is a positive indication that IDVC may be useful for other setups.

6. Conclusions

In this work we extended the inter dataset variability compensation method firstly introduced in [15] to capture variability in the hyper-parameters of the PLDA model, namely the center μ, the within-speaker covariance matrix W, and the between speaker covariance matrix B.

IDVC has shown to effectively reduce the influence of dataset variability on the investigated i-vector PLDA system in the context of the JHU-2013 domain adaptation challenge. When evaluated on a system trained on the Switchboard corpus, EER was decreased by 62%, DCF (old) by 58% and DCF (new) by 33%. These error reductions recover 85% of the degradation due to mismatched PLDA training (MIXER training is considered to be the matched condition).

Experiments conducted with 3 different data partitions for IDVC training indicate that the method works well even when trained on two subsets of data only and without speaker labels.

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 Linguistic 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://www.clsp.jhu.edu/workshops/archive/ws13-summer-workshop/groups/spk-13/.

[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, "Inter dataset Variability compensation for speaker recognition", in *ICASSP*, 2014.