Domain Adaptation for Text Dependent Speaker Verification

Hagai Aronowitz, Asaf Rendel

IBM Research – Haifa, Haifa, Israel
hagaia@il.ibm.com, asafren@il.ibm.com

Abstract

Recently we have investigated the use of state-of-the-art text-dependent speaker verification algorithms for user authentication and obtained satisfactory results mainly by using a fair amount of text-dependent development data from the target domain. In this work we investigate the ability to build high accuracy text-dependent systems using no data at all from the target domain. Instead of using target domain data, we use resources such as TIMIT, Switchboard, and NIST data. We introduce several techniques addressing both lexical mismatch and channel mismatch. These techniques include synthesizing a universal background model according to lexical content, automatic filtering of irrelevant phonetic content, exploiting information in residual supervectors (usually discarded in the i-vector framework), and inter dataset variability modeling. These techniques reduce verification error significantly, and also improve accuracy when target domain data is available.

Index Terms: speaker verification, text-dependent, domain adaptation

1. Introduction

The introduction of i-vectors [1] and Probabilistic Linear Discriminant Analysis (PLDA) [2] resulted in very low error rates in the recent NIST text-independent (TI) speaker recognition evaluations (SREs) [2]. However, the success of i-vector based PLDA is dependent on the availability of a large development set with thousands of multi session speakers, to estimate the PLDA hyper-parameters. Moreover, the development data must be matched to the evaluation data.

For text-dependent speaker recognition, the use of the i-vector framework has not been as successful as for TI speaker recognition. In [3-5] the Nuisance Attribute Projection (NAP) [6] framework clearly outperformed the i-vector framework. Moreover, results in [3-5] were obtained using in-domain (target) data for development (at least 100 sessions). In practice, in-domain data is not always available, or may be available in very small amounts. In these cases resources such as NIST SRE data may be used for the sake of system development.

In [5] it was concluded that NIST SRE data was adequate for training an i-vector-LDA system as long as some in-domain data was used for training the universal background model (UBM) and for score normalization. Best results were achieved by adapting the NIST data to the target passphrase by filtering out frames that didn't match well the UBM.

In this paper we present our efforts for performing text-dependent speaker verification without the use of in-domain data. The types of mismatch needed to be addressed are channel and lexical mismatch. We investigate both the i-vector and the NAP frameworks and propose four novel methods to enhance text-dependent speaker recognition accuracy. First, we propose to train a passphrase dependent UBM by selecting the appropriate context-dependent phonetic HMM states trained from a phonetically transcribed corpus. Second, we propose a novel method for adapting TI development data to the passphrase lexical content. Third, we show that the residual of the i-vector extraction process is extremely useful for text-dependent speaker recognition. Fourth, we propose to use inter dataset variability compensation (IDVC) [7,8] for reducing additive mismatch.

The remainder of this paper is organized as follows: Section 2 describes the datasets. Section 3 describes our speaker verification baseline systems. Section 4 describes the proposed methods. Section 5 presents the experiments and results. Finally, Section 6 concludes.

2. Datasets

2.1. The WF corpus

The WF corpus consists of 750 speakers which are partitioned into a development dataset (200 speakers) and an evaluation dataset (550 speakers). Each speaker has 2 sessions using a landline phone and 2 sessions using a cellular phone. The data collection was accomplished over a period of 4 weeks.

Four authentication conditions were defined and collected (global, speaker-dependent and prompted passphrases as well as free text), and experimental results were reported for them in [3]. In this work we limit ourselves to the global (passphrase is shared among all speakers) and the speaker (passphrase is speaker dependent) conditions for which the same passphrase is used for both enrollment and verification. We report results for the 10-digit pass phrase 0-1-2-3-4-5-6-7-8-9 which we name ZN.

In the WF dataset each session contains 3 repetition of ZN. For each enrollment session we use all 3 repetitions for enrollment, and for each verification session we use only a single repetition. For some controlled experiments we use ~30-seconds long TI utterances (one per sessions) which we denote by WF-TI. These utterances are used to differentiate between the effect of lexical mismatch and the effect of channel mismatch.

A comprehensive description of the WF data can be found in [3].

2.2. Standard telephony development set

We use the following standard conversational telephony datasets [9]: Switchboard-II, NIST 2004, 2005, 2006 and 2008 speaker recognition evaluations (SREs). We denote this development set by NIST.

2.3. Phonetically transcribed datasets

We use the TIMIT corpus which contains broadband recorded read speech [9] for training phonetically-inspired UBMs. TIMIT contains 6,300 phonetically annotated utterances read by 630 speakers. In this work, we use the standard training set, which consists of 3,696 sentences.

3. Speaker verification baseline system

In this section we describe the baseline speaker verification systems we use in this work.

3.1. I-vector-based system

We use standard i-vector extraction with length normalization followed by LDA (Linear Discriminant Analysis) and WCCN (Within Class Covariance Normalization). We use cosine-based similarity scoring and normalize using ZT-norm, which we found to be significantly superior to s-norm under domain mismatch. We trained the i-vector extractor, LDA and WCCN on the Switchboard-II, NIST 2004, 2006 and 2008 SREs. We did not use PLDA, due to results from preliminary experiments which indicated PLDA was slightly inferior to LDA+WCCN under domain mismatch.

3.2. GMM-NAP-based system

Our GMM-NAP system is described in detail in [3]. Contrary to past system, we use neither a geometric mean comparison kernel [10] nor two-wire NAP [11] as we found these methods to degrade accuracy when development data is highly mismatched to the target domain. Instead of the geometric mean kernel we use the dot product kernel:

$$C_{dot}\left(E,T\right)=m_E^t\left(\lambda_{UBM}^{1/2}\otimes I_n\right)\Sigma^{-1}\left(\lambda_{UBM}^{1/2}\otimes I_n\right)m_T \qquad (1)$$

where E and T stand for the enrollment and test sessions, m_E and m_T are the corresponding concatenated GMM means, λ_{UBM} stands for the UBM weights, $\sum$ is a block matrix with covariance matrices from the UBM on the diagonal, n is the feature vector dimension, $\otimes$ is the Kronecker product, and In is the identity matrix of rank n.

4. Domain adaptation for text-dependent speaker recognition

In this section we describe the methods we propose for coping with the channel and lexical mismatch between our development data (mainly the NIST SREs data) and the target text dependent data (WF data).

4.1. Synthesizing a passphrase dependent UBM by pooling context-dependent phonetic HMM states

An important conclusion of the works in [3-5] is that for UBM-based systems, it is important to train the UBM from data that matches the lexical content of the passphrase.

We therefore utilize the phonetically aligned TIMIT data in the following way: first a context-dependent HMM model is trained based on TIMIT; then, Gaussians that are not associated with states (phonetic contexts) that appear in the state-graph induced by the passphrase are removed. Finally, the number of Gaussians is reduced to 512 by k-means clustering (using KL-divergence distance), Mixture weights are distributed uniformly between these Gaussians.

4.2. Transferring the text-independent resources (NIST) to better match the text-dependent (WF) task

Our aim is to find techniques to reduce the mismatch between the text-independent (NIST) and text-dependent (WF) data by transforming the text-independent data to better match the text-dependent task. Previously [5], we have filtered the NIST data (for the purpose of extracting sufficient statistics) by selecting only likely frames (having high posterior probability) based on a text-dependent UBM trained on the WF development data.

In this work, we assume no text-dependent training data, and explore filtering the NIST data based on the considered passphrase's phonetic transcript alone. To this end, we utilize the phonetically aligned TIMIT data to train a UBM having each Gaussian labeled as either belonging or not belonging to the passphrase. Then, we use this UBM for extracting sufficient statistics from the NIST data. After sufficient statistics are computed, we throw away the statistics corresponding to the Gaussians that are labeled as not belonging to the passphrase. Figure 1 illustrates the sufficient statistics extraction process.

For processing the target domain data (WF enrollment and verification data) we remove all the Gaussians from the UBM that are un-associated to the passphrase, and renormalize the mixture weights.

The UBM is generated from the TIMIT HMM model by simply clustering all Gaussians (from all states) to 1024 Gaussians (maintaining the state to Gaussian mapping), and labeling Gaussians as belonging-to-passphrase if they are associated with states (phonetic contexts) that appear in the state-graph induced by the passphrase. A total of 454 Gaussians are labeled as passphrase Gaussians for the ZN passphrase. Mixture weights are distributed uniformly between these Gaussians.

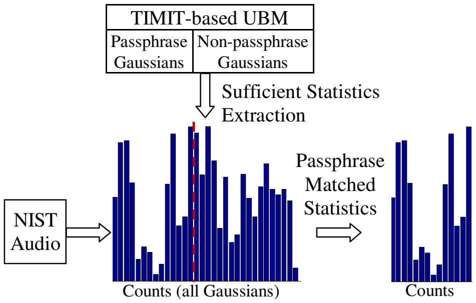

Figure 1: Passphrase-specific sufficient statistics extraction. The process of computation of the GMM counts is illustrated. Computation of the first order statistics (sums) is similar.

4.3. I-vector residual supervectors

The general inferiority of the i-vector based framework compared to NAP in the context of text-dependent speaker recognition implies that some standard assumptions are not suitable for the text-dependent setup.

I-vector extraction is based on an assumption that most of the relevant speaker information resides in a low dimensional subspace (400-1000 dimensional), and that suitable development data is available for estimating that subspace. We hypothesize that for text-dependent speaker recognition, either the small amount of development data (when domain data is used) or the lexical mismatch (when out of domain data is used) breaks the above assumptions.

In order to validate our hypothesis we define the residual of the i-vector extraction process as follows. In the i-vector framework, a session is represented by a supervector of stacked GMM means denoted by *s* and is modeled as:

$$s = m + \mathrm{T}w + \varepsilon \qquad (2)$$

where *m* is the UBM supervector, T is a low-rank rectangular matrix of column vectors spanning the *total variability* subspace that is supposed to capture most of the variability in the supervector space, *w* is a low dimensional vector (the i-vector) and *ε* is an error or residual supervector which is discarded in the standard i-vector framework. Supervector *s* is therefore a sum of three supervectors: the mean *m*, the *total variability* supervector which resides in the *total variability* subspace spanned by the columns of T, and the residual which resides in the complement of the *total variability* subspace.

Given sufficient statistics for an audio session, we extract the i-vector *w* as usual. We extract the residual supervector *ε* by estimating the supervector *s* using relevance-MAP (as done in the NAP framework) and removing from *s* its projection on the *total variability subspace*. The method for extraction of the residual supervector can be summarized as:

1. Estimate *s*
2. $\varepsilon = (s - m)(\mathrm{I} - \mathrm{VV'})$
   where V denotes an orthonormal basis for the span of the columns of T

In Section 5 we report results for using the residual supervectors for speaker recognition. We use the dot product for scoring, followed by ZT-score normalization. The normalized score obtained using the residual supervectors is further fused with the normalized scores of the i-vector system using a weighted average.

Note that use of the residual in the complement of the *total variability* subspace has been already done for speaker recognition in 2007 [12], when kernel-PCA (Principal Component Analysis) was used to capture the *total variability* subspace (instead of the modern Factor Analysis approach).

4.4. Inter dataset variability compensation (IDVC)

Recently, a method called IDVC has been proposed [13, 14] for reducing the effect of channel mismatch in the context of i-vector-PLDA based TI speaker recognition. IDVC compensates dataset shifts in the i-vector space by constraining the shifts to a low dimensional subspace. The subspace is estimated from a collection of distinct datasets (not necessarily having speaker labels).

In its simplest version [13], the means of the i-vectors of each dataset are assumed to reside in a low dimensional subspace which is estimated using PCA. This subspace is estimated from the set of distinct datasets and then removed from all i-vectors as a pre-processing step before PLDA training and scoring. In [13,14], the mismatch subspace we estimated from the development datasets happened to generalize well to the mismatch observed in the evaluation data.

In this work we use the simple version [13] for compensating dataset mismatch for the i-vector system. For the NAP system, we apply the same method on supervector means instead of i-vector means. The datasets we use to train IDVC are as follows: Switchboard-II phase 2, Switchboard-II phase 3, NIST 2004, NIST 2005, NIST 2006, NIST 2008, NIST 2008 microphone, TIMIT sentence SA1, TIMIT sentence SA2 and TIMIT excluding sentences SA1 and SA2.

5. Experiments and results

The WF evaluation set consists of 17994 target trials and 81720 imposter trials. We report results for the following four conditions: all trials (All), same gender trials (SG), matched channel trials (MC), and different channel trials (DC). Results for the i-vector system are reported in Table 1, and results for the NAP system are reported in Table 2. Table 3 reports results for an analysis of the relative magnitudes of the effect of channel and lexical mismatch.

5.1. Baseline results

Use of the out-of-domain data instead of in-domain data causes error rates to double for the i-vector system and to triple for the NAP system. Note that the NAP error rates are 50% lower than those of the i-vector system when trained on in-domain data. For the rest of this section, baseline results are considered to use the out-of-domain (NIST) data only.

5.2. TIMIT based UBMs

The use of a passphrase-matched UBM trained on TIMIT (denoted by **TIMIT based UBM**) results in a ~9% relative reduction of EER for the i-vector system and no significant change for NAP. The use of a UBM trained on TIMIT with adaptation of the NIST data (denoted by **TIMIT based UBM with adaptation**) results in a ~13% relative reduction of error rate for the i-vector system and a ~4% relative reduction of error rate for the NAP system.

5.3. Residual supervectors

Scoring the residual supervectors results in a ~5% relative reduction of error rate compared to the i-vector baseline. Fusing the ZT-normalized scores of the baseline i-vector system and the residual-based system yields a ~22% error reduction compared to the baseline.

Residual supervectors were also evaluated when training is done on in-domain data (WF). An error reduction of 26% is obtained compared to the in-domain baseline and a relative reduction of 36% when fused with the i-vector baseline.

5.4. IDVC

IDVC applied on the baseline gave very modest improvements: 1% relative for both systems. However, when applied jointly with **TIMIT based UBM with adaptation**, it gave a relative improvement of 3% (16% in total) for the i-vector system and 2% (6% in total) for the NAP system.

A further improvement is achieved when fused with a residual system (with **TIMIT based UBM with adaptation),** which results in a total relative improvement of ~31% compared to the baseline.

Table 1. I-vector-based results for NIST (out-of-domain) and WF (in-domain) training. The baseline system is contrasted to the proposed methods. Results are in EER (in %).

Method	All	SG	MC	DC
Out-of-domain (NIST) training				
Baseline	3.90	5.15	1.88	4.64
TIMIT based UBM	3.60	4.78	1.68	4.24
TIMIT based UBM with adaptation	3.48	4.62	1.53	4.17
Residual	3.69	4.91	1.78	4.40
I-vector + residual	3.03	3.96	1.47	3.62
IDVC	3.84	5.09	1.85	4.59
TIMIT based UBM with adaptation + IDVC	3.36	4.45	1.52	3.98
TIMIT based UBM with adaptation + i-vector + residual + IDVC	2.70	3.55	1.32	3.23
In domain (WF-ZN) training				
Baseline	1.97	2.63	1.04	2.28
Residual	1.48	1.92	0.74	1.75
I-vector + residual	1.25	1.66	0.66	1.51

Table 2. NAP-based results for NIST (out-of-domain) and WF (in-domain) training. The baseline system is contrasted to the proposed methods. Results are in EER (in %).

Method	All	SG	MC	DC
Out-of-domain (NIST) training				
Baseline	3.25	4.28	1.40	3.79
TIMIT based UBM	3.26	4.25	1.41	3.78
TIMIT based UBM with adaptation	3.05	4.08	1.41	3.61
IDVC	3.16	4.22	1.35	3.76
TIMIT based UBM with adaptation + IDVC	3.00	4.00	1.37	3.56
In domain (WF-ZN) training				
Baseline with WF data	0.96	1.22	0.54	1.11

Table 3. NAP-based results for using different datasets for centering and score normalization. NIST data is both channel and lexically mismatched. WF-TI is only lexically mismatched

Centering	Score-norm	All: EER (in %)
NIST	NIST	3.25
WF-TI	NIST	2.85
WF-ZN	NIST	2.34
NIST	WF-TI	2.68
NIST	WF-ZN	1.92

5.5. Analysis of the magnitudes of the effect of channel and lexical mismatch

Both channel and lexical mismatch are tackled in this paper. The experiments reported in Table 3 give some indications about the relative effect of each type of mismatch in our setup. Regarding i-vector centering, the degradation due to lexical mismatch (2.34→2.85) is roughly equal to the degradation due to channel mismatch (2.85→3.25). A similar phenomenon is observed for score normalization: the degradation due to lexical mismatch (1.92→2.68) is roughly equal to the degradation due to channel mismatch (2.68→3.25).

We hypothesize that the modest improvement obtained using IDVC (especially when applied without a TIMIT-based UBM) are explained by noting that lexical mismatch is highly non-linear and IDVC is a linear approach. When IDVC is applied jointly with **TIMIT based UBM with adaptation**, the lexical mismatch is partly compensated and therefore the channel effect may become more dominant with regard to the centering problem, therefore enhancing the effectiveness of IDVC.

6. Conclusions

In this work we have explored the possibility of building text-dependent speaker verification systems using out-of-domain text-independent data (NIST). In our preliminary work [5] we investigated the i-vector framework and focused on the UBM, i-vector extraction, LDA and WCCN components but used limited amounts of in-domain (WF) data for score normalization. In this work, we addressed the scenario were no in-domain data is available at all.

We investigated both the i-vector framework and the NAP framework. Although NAP is inferior to i-vector-PLDA in the context of NIST TI evaluations, it has been found in the past [3-5] to outperform the i-vector framework for text-dependent speaker recognition.

We have several contributions in this work. First, we propose to synthesize a UBM from a phonetic HMM trained on TIMIT keeping track of Gaussians that correspond to the lexical content of the passphrase, and use this UBM to extract sufficient statistics from out-of-domain NIST data. These statistics are adapted to the passphrase by removing the irrelevant Gaussians, and the in-domain enrollment and verification data is processed using the truncated GMM. This method outperformed the simpler method of synthesizing a passphrase matched UBM, and reduced EER by 13% for the i-vector system and 4% for the NAP system.

Second, we propose to use the recently introduced IDVC method for improving the robustness of our systems to additive mismatch in the i-vector or supervector domains. Using IDVC reduced EER by 2-3% on top of the improvements obtained using the TIMIT-based UBM and adaptation method.

Third, we show that the i-vector representation fails to capture important information which can be recovered by the residual supervector. Augmenting the i-vectors with the residual supervectors improved the accuracy significantly for both out-of-domain training (22%) and in-domain training 36%).

Finally, using all three contributions jointly reduced the EER of the out-of-domain based system i-vector system by 31% relatively (3.9% reduced to 2.7%), which outperforms the NAP framework (3%).

7. Acknowledgements

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

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

[3] H. Aronowitz, R. Hoory, J. Pelecanos, D. Nahamoo, "New Developments in Voice Biometrics for User Authentication", in Proc. *Interspeech*, 2011.

[4] H. Aronowitz, "Text Dependent Speaker Verification Using a Small Development Set", in Proc. *Speaker Odyssey*, 2012.

[5] H. Aronowitz, O. Barkan, "On Leveraging Conversational Data for Building a Text Dependent Speaker Verification System", in Proc. *Interspeech*, 2013.

[6] A. Solomonoff, W. M. Campbell, and C. Quillen, "Nuisance Attribute Projection"**,** *Speech Communication*, Elsevier Science BV, 1 May, 2007.

[7] H. Aronowitz, "Inter dataset Variability compensation for speaker recognition", to appear in *ICASSP*, 2014.

[8] H. Aronowitz, "Compensating Inter-Dataset Variability in PLDA Hyper-Parameters for Robust Speaker Recognition", submitted to *Speaker Odyssey*, 2014.

[9] J.S. Garofolo, L.F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "The DARPA TIMIT acoustic phonetic continuous speech corpus CD-ROM," NIST, 1993.

[10] W. Campbell, Z. Karam, "Simple and Efficient Speaker Comparison using Approximate KL Divergence", in Proc. *Interspeech*, 2010.

[11] Y.A. Solewicz, H. Aronowitz, "Two-Wire Nuisance Attribute Projection", in Proc. *Interspeech* 2009.

[12] H. Aronowitz, "Speaker recognition using Kernel-PCA and Intersession Variability Modeling", in Proc. *Interspeech*, 2007.

[13] H. Aronowitz, "Inter dataset Variability compensation for speaker recognition", to appear in *ICASSP*, 2014.

[14] H. Aronowitz, "Compensating Inter-Dataset Variability in PLDA Hyper-Parameters for Robust Speaker Recognition", submitted to *Speaker Odyssey*, 2014.