IDIAP RESEARCH REPORT

CONTENT NORMALIZATION FOR TEXT-INDEPENDENT SPEAKER VERIFICATION

Subhadeep Dey Marc Ferras Petr Motlicek
Srikanth Madikeri

Idiap-RR-31-2017

DECEMBER 2017

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

CONTENT NORMALIZATION FOR TEXT-INDEPENDENT SPEAKER VERIFICATION

Subhadeep Dey[1,2]*, Marc Ferras*[1]*, Petr Motlicek*[1] *and Srikanth Madikeri*[1]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
{subhadeep.dey,marc.ferras, petr.motlicek, srikanth.madikeri}@idiap.ch

ABSTRACT

In the past few years, Deep Neural Network (DNN) based i-vector Speaker Verification (SV) systems have shown to provide state-of-the-art performance. However, error rates increase drastically for short duration recordings. In this paper, we improve the i-vector approach for short utterances, (i) by using smoothed DNN posteriors for i-vector extraction, and (ii) by normalizing the content of the enrollment data to match the test data. The quality of DNN posteriors is enhanced by employing an Automatic Speech Recognition (ASR) system to generate phone recognition lattices. These lattices incorporate content information through the use of lexical and language models across a number of hypothesized paths. Content normalization is then performed by estimating i-vectors on phonetic units. The largest similarity scores across phonetic units co-occurring in enrollment and test are taken. Experiments on a modified protocol of the RSR database show that the proposed approach achieves 67% relative improvement in equal error rate over a DNN-based i-vector baseline system in a condition where the content of the test data has been seen during enrollment.

Index Terms— speaker verification, i-vectors, content mismatch, posteriors

1. INTRODUCTION

Over the last decade, the state-of-the-art techniques in Speaker Verification (SV) such as i-vector and Joint Factor Analysis (JFA) have shown to provide high performance for a variety of conditions including long duration utterances [1, 2]. When applied to forensic or access control, systems are asked to deal with short recordings of speech. However, the performance of SV systems on short test utterances is far from being acceptable for any deployable system [3]. Such poor performance can be mainly attributed to the difficulty of the SV systems to deal with the content mismatch, i.e. the choice of words differ from enroll to test [4].

In text-dependent SV systems using fixed phrases, the test and enrollment content is expected to be the same. In case it is not, the system can reliably detect the mismatch and reject the claim. In text-independent SV, the test and enrollment content is unconstrained, thus being unlikely for the content to be matched. In this scenario, systems are asked to make decisions independently of such mismatch, i.e. focusing on the speaker factor only.

Various techniques have been explored that aim at exploiting the content information of the test data for text-independent SV systems [5, 6, 7]. In [5], content information is used by extracting an i-vector for every linguistic unit of the utterance. Experimental results show that a significant gain in performance can be achieved using this approach. In [6, 7], content-matching is performed by transforming the enrollment data to match the lexical content of the test utterance. In [6], a few top-scoring posteriors from a Gaussian Mixture Model (GMM) are selected to transform a conversational utterance into a text-dependent one. In [7], posteriors estimated using a Deep Neural Network (DNN) are used for content-matching, prior to i-vector extraction. This approach outperforms a GMM based i-vector system, probably due to the use of DNN being trained for content discrimination. Furthermore, an approach that scales enrollment zero-th order statistics to match test statistics is proposed as a way to successfully deal with content mismatch [7].

The conventional approaches described above perform content matching in the i-vector framework using context-dependent state (senone) posteriors estimated using DNN. In our work, we use senone posteriors estimated from Automatic Speech Recognition (ASR) phone recognition lattices. These senone posteriors incorporate the information of both the acoustic (incorporating also lexical model) and language models, thus increasing phone classification accuracy. We leverage on recent work where the performance of i-vector based SV systems was shown to be directly linked to such accuracy [8].

We also present a method to perform content normalization by selecting regions in the enrollment data to match the test data by employing i-vectors. In this approach, we assume that estimating i-vectors on linguistic units, such as phones, of the speech signal can contain sufficient speaker and content information. The common phonetic units between the enrollment and test data are obtained by using cosine distance metric.

The paper is organized as follows. Sections 2 and 3 describe the scenarios considered in this paper and the baseline system while Section 4 describes SV using posteriors generated by ASR and the content normalization technique. Sections 5 and 6 describe the experimental setup for the evaluating the system and discuss the achieved results by various systems. Finally, the paper is concluded in Section 7.

2. TEXT-INDEPENDENT SCENARIO

Content variability has a strong impact on the performance of a SV system when short utterances are involved [4, 9]. For text-dependent SV, the enrolled content is expected to be the same as the test content, as shown in Figure 1 (a) (**Matched**).

In the text-independent case, the system may have to deal with the content mismatch to produce improved similarity scores. Regardless of the enrollment data being used, two scenarios can be considered to better understand the effect of content mismatch:

- **Seen**: the test content has been pronounced during enrollment, and thus, lexical content of the test data is a subset of

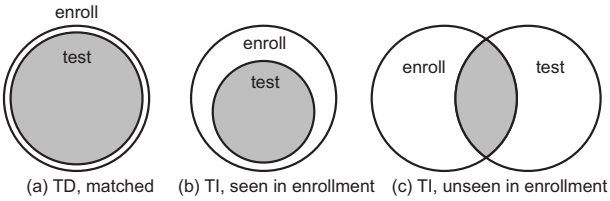

Fig. 1: Various scenarios of sharing content for the Text-Dependent (TD) and Text-Independent (TI) systems.

the enrollment data. This is shown in Figure 1 (b).

- **Unseen**: the test content has not been spoken during enrollment. Figure 1 (c) illustrates this as well as a possible intersection between the phonemes or words co-occurring between enrollment and test data.

In this paper, we aim at improving speaker recognition performance of a SV system in the scenarios described above.

3. BASELINE SYSTEM

The state-of-the-art text-independent SV approach to model speakers is built around i-vectors [2]. This approach assumes that the invariant speaker characteristics lie in a low dimensional subspace of mean GMM supervectors. A speaker model is represented by a fixed-dimensional vector called *i-vector*.

In [7], DNNs were used to cluster the acoustic space into linguistic units such as senones, making it easier to focus on the content of each utterance. The posterior probabilities of each of the senones were then used for i-vector extraction. A posterior normalization technique was further proposed to scale the zero-th and first order statistics of the enrollment data to match those in the test data [7]. The technique is described as follows. Let N_e, N_t, $\mathbf{F}_e$ and $\mathbf{F}_t$ be the zero-th order first order statistics of the enrollment and test utterances respectively. The new statistics of the enrollment are obtained as

$$\mathrm{N}_e^{'} = \beta \mathrm{N}_e \tag{1}$$

$$\mathbf{F}_e^{'} = \beta \mathbf{F}_e, \tag{2}$$

where β is a normalization constant and is defined as $\mathrm{N}_t/\mathrm{N}_e$. When N_e or N_t is 0, β is set to zero as well. The details of the technique can be found in [7]. The DNN based i-vector system using this technique is among the baseline systems in this paper.

4. POSTERIORS AND CONTENT MATCHING

In this work, we use two techniques to perform content normalization, one based on DNN posterior estimation and the other using online i-vectors. Both are described in the following sections:

4.1. Posteriors from ASR decoder

A DNN based i-vector system involves the estimation of zero-th and first order statistics as a prior step to computing the i-vectors. The state-of-the-art SV systems compute these statistics using the senone posteriors obtained at the output of the DNN [7, 10]. Therefore, the DNN acts as a short-term content estimator in terms of senones.

In this work, senone posteriors are obtained after decoding using language and lexical models, in the context of an ASR system. In [8], it was shown that senone posteriors obtained after ASR decoding performed better than those obtained after a DNN forward pass. The former posteriors are smoothed by using language constraints and drastically improve the phone accuracy.

In our work, we use a lattice decoder [11], based on a Weighted Finite State Transducer (WFST), that outputs a graph of hypothesized sequences of phones (although the acoustic model is trained using a word dictionary, we applied a phone Language Model (LM)). Senone posterior probabilities are estimated from the acoustic scores at the nodes of the lattice, after the forward-backward recursion, for each frame. These are used for i-vector extraction. For content normalization, we use the posterior normalization technique as proposed for the baseline system [7].

4.2. Content normalization using i-vectors

In this paper, an alternative content normalization technique is further proposed. This approach decodes phone sequences for each utterance and computes i-vectors on the acoustic features aligned with each phone class instance in an utterance.

Enrollment and test content are matched by computing the maximum similarity scores from each phone class instance in test to all instances in enrollment. As many scores as the number of phone class instances in test are obtained. Finally, these scores are averaged to obtain a global similarity score. The rationale behind this approach is to choose the best phone class instance in the enrollment data. The accumulated global score is obtained as follows

$$s(\mathbf{X}, \mathbf{Y}) = \frac{1}{C} \sum_j min_i d(\mathbf{x}_i, \mathbf{y}_j), \tag{3}$$

where $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_R\}$ and $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_C\}$ represent set of i-vectors for the enrollment and test data, the function $\mathrm{d}(\mathbf{x}_i, \mathbf{y}_j)$ computes the distance between the i-vectors $\mathbf{x}_i$ and $\mathbf{y}_j$. The score $s(\mathbf{X}, \mathbf{Y})$ represents the accumulated distance between the closest phone units. We used the cosine distance metric to compute the dissimilarity between two i-vectors. A threshold on the cosine distance is applied to detect when a test phone unit is not present in the enrollment data. This threshold is optimized for the development data set.

5. EXPERIMENTAL SETUP

In this section, we describe the experimental setup for the baseline and proposed systems.

5.1. Evaluation and Training Data

We performed experiments on the evaluation Part1 portion of the RSR data set [9, 12, 13], restricting to female speakers only. This subset contains 49 speakers with speech utterances from 30 unique pass-phrases. The protocol described in [9] was adopted to perform text-dependent and text-independent SV. We created three conditions based on those found in [7] to evaluate the proposed systems:

- **Matched**: Speaker-mismatch trials involving 15 pass-phrases, 735 speaker models, 4'410 and 211'680 target and impostor trials, as part of the original RSR text-dependent protocol.
- **Seen**: Text-independent trials involving 15 pass-phrases with each speaker model being built using three sessions for each pass-phrase, for a total of 45 utterances. The test data is same as in **Matched** condition. The same number of trials as in **Matched** condition are scored.

- **Unseen**: Text-independent trials with the enrollment data being the same as in the **Seen** condition. The test data are taken from utterances involving pass-phrases that are not seen during enrollment. This condition contains 4'405 target and 211'310 impostor trials.

The Fisher female English Part I and II data was used as the training data. It contains about 13k utterances with 1000 hours. The NIST datasets - SRE 2004, 2005, 2006, 2008 and 2008 extended, Switchboard Part II and Part III, and Switchboard Cellular Part I and II - were used as the development data. All speech files were downsampled to 8 kHz for compatibility with other datasets used for system development.

5.2. i-vector system

The front-end SV system extracts Mel Frequency Cepstral Coefficients (MFCC) of 20 dimensions from 25 ms of frame of speech signal with 10 ms sliding window with the delta and double delta features appended to it. Short time gaussianization is applied to the features using a 3 s sliding window [14]. The Hungarian phoneme recognizer is used to detect voice activity. It compares the sum of posteriors over all phone classes with the posterior of the silence class to classify each frame as speech or non-speech [15]. This is used to mark the start and end points of the speech region in the audio. The training data is used to estimate the parameters of the i-vector model. The dimensionality of i-vector extractor is set to 400. Linear discriminant analysis (LDA) and Probabilistic Linear discriminant analysis (PLDA) models are trained on the development data.

5.3. ASR system

DNN acoustic model is trained as a part of the ASR system. It is trained with MFCCs with 6 hidden layers each of dimension 1200. The output layer has 1530 senone units including 20 silence units. The ASR system employs a CMU dictionary with 42k words, similar to [3]. The ASR system is validated on a separate subset consisting of 200 utterances from the Fisher database with 3gram word LM. The Word Error Rate (WER) on the validation set is 24.4%.

For generating phone recognition lattices, we used a 2gram phone LM trained on transcripts of the training data. The extracted senone posteriors are used to estimate the parameters of the i-vector model.

6. EXPERIMENTAL RESULTS AND DISCUSSIONS

In this section, we describe the results obtained with the baseline and the proposed SV systems. The various systems considered in this paper are the following

- **Ivec$_{\text{PLDA}}$**: the conventional i-vector systems for speaker recognition. The systems using GMM, DNN and decoded ASR (phone) lattice posteriors are referred to as **Ivec$^{\text{GMM}}_{\text{PLDA}}$**, **Ivec$^{\text{DNN}}_{\text{PLDA}}$** and **Ivec$^{\text{DNN-dec}}_{\text{PLDA}}$** respectively.
- **PN**: the systems using posterior normalization technique as explained in Section 4.1. The systems using GMM, DNN and decoded ASR (phone) lattice posteriors for i-vector extraction are referred to as **PN-Ivec$^{\text{GMM}}_{\text{PLDA}}$** , **PN-Ivec$^{\text{DNN}}_{\text{PLDA}}$** and **PN-Ivec$^{\text{DNN-dec}}_{\text{PLDA}}$** respectively.
- **minD**: the SV systems applying content normalization technique using i-vectors as explained in Section 4.2. The systems using GMM, DNN and decoded ASR (phone) lattice posteriors for i-vector extraction are referred to as **minD-Ivec$^{\text{GMM}}$** , **minD-Ivec$^{\text{DNN}}$** and **minD-Ivec$^{\text{DNN-dec}}$** respectively.

Table 1: Performance of the various baseline systems in terms of EER(%). The **PN-Ivec$^{\text{DNN}}_{\text{PLDA}}$** provides the best performance among the baseline systems in **Seen** condition.

Systems/Conditions	**Matched**	**Seen**	**Unseen**
Ivec$^{\text{GMM}}_{\text{PLDA}}$	3.7	16.5	**22.5**
Ivec$^{\text{DNN}}_{\text{PLDA}}$	2.8	11.6	23.4
PN-Ivec$^{\text{GMM}}_{\text{PLDA}}$	3.1	12.3	27.1
PN-Ivec$^{\text{DNN}}_{\text{PLDA}}$	**2.4**	**8.6**	28.4

6.1. Baseline SV systems

Table 1 shows the performance of various i-vector based SV systems. We observe that the performance of the systems on **Matched** condition (column 1 of rows 1 and 2) is worse compared to the results published in [16]. It may be due to the case that parameters of the systems in [16] are tuned for the RSR dataset, while in this paper we rather used the text-independent SV setup (without using RSR data set for system development).

From Table 1, we observe that incorporating the linguistic information from a DNN benefits performance in **Matched** and **Seen** conditions. The **Ivec$^{\text{DNN}}_{\text{PLDA}}$** consistently performs better than **Ivec$^{\text{GMM}}_{\text{PLDA}}$** by relative Equal Error Rate (EER) of about 24% (from 3.7 % to 2.8% absolute) and 30% (from 16.5% to 11.6%) in **Matched** and **Seen** conditions respectively. Furthermore, we observe that the performance of both these systems in **Seen** condition is significantly worse than in the **Matched** condition. We observe that even though content of the test data is contained in the enrollment, the i-vector system is unable to exploit this information. To use the content information of test data, posterior normalization technique as described in Section 3 is used. We observe that the technique helps i-vector system in **Matched** and **Seen** conditions considerably. The **PN-Ivec$^{\text{DNN}}_{\text{PLDA}}$** improves upon **Ivec$^{\text{DNN}}_{\text{PLDA}}$** by relative EER of about 14% (from 2.8% to 2.4 % absolute) and 26% (from 11.6% to 8.6% absolute) in **Matched** and **Seen** conditions respectively. In **Unseen** condition, the systems based on posterior normalization perform worse than the conventional i-vector system. The **PN-Ivec$^{\text{DNN}}_{\text{PLDA}}$** will act as the baseline system for the following experiments.

6.2. SV systems using ASR lattice posteriors

We explore the application of posteriors estimated from phone recognition ASR lattices in an i-vector framework. Table 2 shows the performance of the i-vector systems using these posteriors. We observe that **Ivec$^{\text{DNN-dec}}_{\text{PLDA}}$** outperforms **Ivec$^{\text{DNN}}_{\text{PLDA}}$** in **Seen** condition by 0.7% absolute improvement in EER. Significant gain in performance is achieved by the **PN-Ivec$^{\text{DNN-dec}}_{\text{PLDA}}$** compared to **PN-Ivec$^{\text{DNN}}_{\text{PLDA}}$** , with about 37% relative EER (from 8.6% to 5.4% absolute) for **Seen** condition. This indicates the importance of more accurate senone alignments in obtaining better SV performance.

6.3. SV systems based on content normalization technique

As opposed to using posterior normalization, we explore content normalization using i-vectors, as described in Section 4.2. Table 3 shows the performance of the proposed content normalization based

Table 2: Performance of the various SV systems (using posteriors from decoded ASR (phone) lattices) in terms of EER(%). The **PN-Ivec$^{\text{DNN-dec}}_{\text{PLDA}}$** performs the best among the other systems in **Seen** condition.

Systems/Conditions	**Matched**	**Seen**	**Unseen**
Ivec$^{\text{DNN-dec}}_{\text{PLDA}}$	2.6	10.9	**24.1**
PN-Ivec$^{\text{DNN-dec}}_{\text{PLDA}}$	**2.4**	**5.4**	27.3

Table 3: Performance of the various SV systems (using content normalization technique) in terms of EER(%). The **minD-Ivec$^{\text{DNN-dec}}$** performs the best among the other systems in **Seen** condition.

Systems/Conditions	**Match**	**Seen**	**Unseen**
minD-Ivec$^{\text{GMM}}$	1.8	4.1	**27.5**
minD-Ivec$^{\text{DNN}}$	**1.1**	2.8	28.1
minD-Ivec$^{\text{DNN-dec}}$	**1.1**	**2.7**	27.9

SV systems using posteriors from GMM, DNN and decoded ASR (phone) lattices. We observe that the proposed systems outperform the posterior normalization based systems in **Matched** and **Seen** conditions. The **minD-Ivec$^{\text{DNN-dec}}$** performs better than **PN-Ivec$^{\text{DNN}}_{\text{PLDA}}$** by relative EER of about 54% (from 2.4% to 1.1% absolute) and 67% (from 8.6% to 2.7% absolute) in **Matched** and **Seen** conditions respectively. This shows the importance of the content normalization technique using i-vectors. The proposed system decreases performance in **Unseen** condition. This may be due to that the phone units that co-occur between the enrollment and test data do not contain sufficient speaker information.

7. CONCLUSIONS

In this paper, we addressed content mismatch problem for short duration SV using the i-vector approach. An i-vector system tackles this problem by incorporating content information using senone posteriors. A posterior normalization technique is applied to scale the sufficient statistics of the enrollment data to match the statistics of the test data. Significant gain in performance is observed for the **Matched** and **Seen** conditions. The DNN based i-vector system applying posterior normalization is considered as the baseline system.

We proposed to improve upon the baseline system by, (a) enhancing the senone prediction accuracy of the DNN posteriors, and (b) normalizing the content of the enrollment to match the test using i-vectors. We observe that proposed approach improves upon the baseline system by 67% relative EER in **Seen** condition. This shows the i-vectors contain sufficient speaker and content information. However in **Unseen** condition, the proposed systems did not give any improvement over the baseline systems. This is probably due to that the linguistic units co-occurring between the enrollment and test data are insufficient to produce speaker discriminating scores.

8. REFERENCES

[1] Daniel Garcia Romero and Carol Y. Espy Wilson, "Analysis of ivector length normalization in speaker recognition systems," in *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27 to 31, 2011*, 2011, pp. 249–252.

[2] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, May 2011.

[3] Petr Motlicek et al., "Employment of subspace gaussian mixture models in speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4445–4449.

[4] Kong Aik Lee, Anthony Larcher, Guangsen Wang, Patrick Kenny, Niko Brümmer, David van Leeuwen, Hagai Aronowitz, Marcel Kockmann, Carlos Vaquero, Bin Ma, et al., "The reddots data collection for speaker recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[5] Liping Chen, Kong Aik Lee, Bin Ma, Wu Guo, Haizhou Li, and Li-Rong Dai, "Phone-centric local variability vector for text-constrained speaker verification," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[6] Hagai Aronowitz and Oren Barkan, "On leveraging conversational data for building a text dependent speaker verification system.," in *INTERSPEECH*, 2013, pp. 2470–2473.

[7] Nicolas Scheffer and Yun Lei, "Content matching for short duration speaker recognition.," in *INTERSPEECH*, 2014, pp. 1317–1321.

[8] Hang Su and Steven Wegmann, "Factor analysis based speaker verification using asr," *Interspeech 2016*, pp. 2223–2227, 2016.

[9] Anthony Larcher, Kong-Aik Lee, Bin Ma, and Haizhou Li, "Text-dependent speaker verification: Classifiers, databases and RSR2015," *Speech Communication*, vol. 60, pp. 56–77, 2014.

[10] Yun Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 1695–1699.

[11] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Nagendra Goel, Mirko Hannemann, Yanmin Qian, Petr Schwarz, and Georg Stemmer, "The kaldi speech recognition toolkit," in *In IEEE 2011 workshop*, 2011.

[12] A. Larcher, Kong Aik Lee, Bin Ma, and Haizhou Li, "Modelling the alternative hypothesis for text-dependent speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 734–738.

[13] A. Larcher, Kong Aik Lee, Bin Ma, and Haizhou Li, "Phonetically-constrained plda modeling for text-dependent speaker verification with multiple short utterances," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, May 2013, pp. 7673–7677.

[14] Jason Pelecanos and Sridha Sridharan, "Feature warping for robust speaker verification," 2001, pp. 213–218, In Proc. of Speaker Odyssey.

[15] Niko Brummer, Lukas Burget, P Kenny, P Matejka, E de Villiers, M Karafiat, M Kockmann, O Glembek, O Plchot, D Baum, et al., "Abc system description for nist sre 2010," *Proc. NIST 2010 Speaker Recognition Evaluation*, pp. 1–20, 2010.

[16] S. Dey, S. Madikeri, M. Ferras, and P. Motlicek, "Deep neural network based posteriors for text-dependent speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, March 2016.