INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

AUT System for SITW Speaker Recognition Challenge

Abbas Khosravani, Mohammad Mehdi Homayounpour

Laboratory for Intelligent Multimedia Processing (LIMP)
Amirkabir University of Technology (AUT)
{a.khosravani,homayoun}@aut.ac.ir

Abstract

This document intends to present AUT speaker recognition system submitted to SITW (Speakers in the Wild) speaker recognition challenge. This challenge aims to provide real world data across a wide range of acoustic and environmental conditions in the context of automatic speaker recognition so as to facilitate the development of new algorithms. The presented system is based on the state-of-the-art *i*-vector/PLDA and source normalization techniques. The system has been developed on publically available databases and evaluated on the data provided by SITW challenge. Taking advantage of the challenge development data, our experiments indicate that source normalization can help speaker recognition system to better adapt to the evaluation condition. Post evaluation analysis is conducted on the conditions of SITW database.

Index Terms: speaker recognition, *i*-vector, probabilistic linear discriminant analysis, source normalization, SITW challenge

1. Introduction

The main theme in the SITW speaker recognition challenge [1] is to use the data acquired without constrain on recording equipment and environmental conditions for the task of speaker detection. The database contains speech utterances from 299 well-known public figures on open-source media channels which offers considerable mismatch in audio conditions [2]. Any noise, reverbration, overlapping speech, laughter and acoustic artifacts in audio files are natural and there is no constrain on the duration of speech utterances. Moreover, unlike most of available databases in the field of text-independent speaker verification, this database has been released free of charge for research purposes.

Speaker verification systems have been tailored to conversational telephony speech due to the availability of an abundance of corresponding data for system development. The majority of these databases are focused on constrained conditions such as clean microphone or telephone speeches. This makes robust speaker verification challenging during evaluation in unconstrained conditions where any kind of natural degradation is presented in audio files. This mismatch motivates us to utilize recent advances in source normalization [3] to bridge the mismatch between development and evaluation conditions.

In recent years speaker verification systems based on *i*-vector features [4] have yielded state-of-the-art performance. This fixed-length and low-dimensional feature vector captures speaker specific information from any arbitrary speech segment. The sufficient statistics of *i*-vectors can be extracted from GMM or phonetically aware DNN posteriors of frame-level features [5] such as MFCC or more recently bottleneck features (frame-level features extracted from a phonetically aware DNN with a special bottleneck layer) [6]. The significant performance gain obtained using DNN is due to the incorporation of speech content into *i*-vector modeling. However, any variability due to acoustic and environmental conditions is still captured by *i*-vectors. In order to robustly compensate for these variability, inter-session compensation techniques such as Within-Class Covariance Normalization (WCCN) [7], Linear Discriminant Analysis (LDA) and Probabilistic LDA (PLDA) [8] have been developed.

Source normalization was recently developed to compensate for speech source variation through improving the estimation of within-speaker scatter matrix from a training database with insufficient variety of speaker utterances from different sources [3]. The within-speaker variability is computed as the residual total variability in *i*-vector space that is not captured by between-speaker variability. The between speaker variability is then computed on a source conditioned basis to remove the bias toward a specific source. This technique has been successfully incorporated into WCCN as well as LDA which offers significant improvement in cross-speech source conditions [3] [9] [10].

The AUT system submission to SITW speaker recognition challenge is based on *i*-vector/PLDA framework (Figure 1). The system has been developed on publically available databases as well as the development portion of the SITW database, and evaluated on the evaluation portion of the SITW database. We have incorporated Source-Normalized WCCN (SN-WCCN) [3] as an *i*-vector pre-processing stage prior to PLDA modeling. In order to mitigate the adverse bias attributed to the mismatch condition of available databases and that of SITW database, we consider the development portion of the SITW database as a different source of variability. Due to the fact that the development and evaluation portion of the SITW database represent similar conditions, it is expected that SN-WCCN results in better adaptation of the system to the evaluation condition.

We outline in this paper the system and experimental results of the submitted system as well as post evaluation analysis. Section 2 describes our speaker recognition system. In Section 3, we describe experimental evaluation on SITW data. Finally we present some post evaluation analysis of the data in Section 4.

2. Speaker Recognition System

In this section we will provide a description of the main components of AUT speaker recognition system including *i*-vector extraction, pre-processing, modeling and scoring. A schematic block diagram of the system is depicted in Figure 1.

2.1. *i*-Vectors

i-Vectors are low-dimensional representation of GMM supervectors in a subspace spanned by the columns of a low-rank

http://dx.doi.org/10.21437/Interspeech.2016-1378

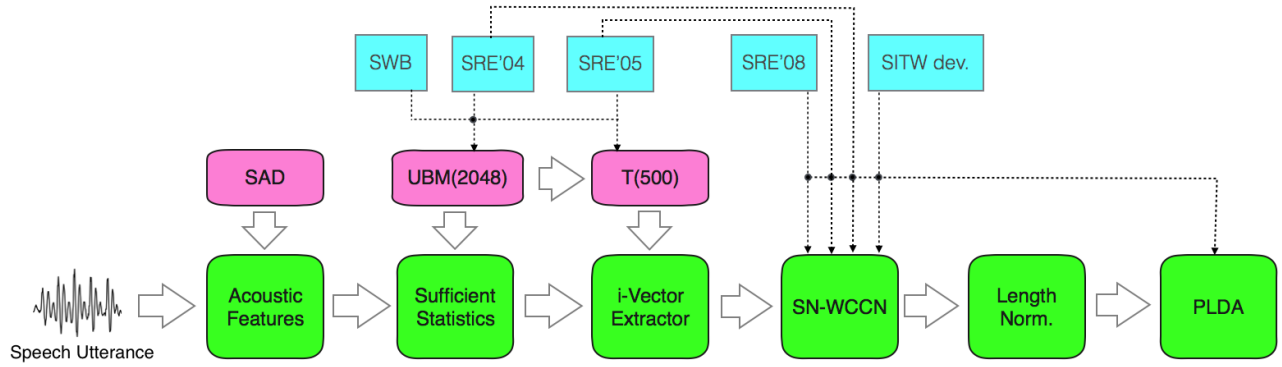

Figure 1: *Block diagram of the AUT speaker recognition system.*

rectangular matrix [4], entitled *total variability matrix*, which preserves characteristics of speaker and inter-session variability. Mathematically, for a given speaker utterance s the adapted supervector can be formulated as

$$s = m + \mathbf{T}w + \varepsilon, \tag{1}$$

where m is the Universal Background Model (UBM) supervector, essentially a speaker-independent GMM supervector, w with standard normal distribution is referred to as the *i-vector*, and ε is the residual term which account for the variability not captured by $\mathbf{T}$. The extraction of *i*-vectors in the proposed system is based on Baum-Welch statistics calculated for a given utterance with respect to UBM components and speech frame-level Mel-Frequency Cepstral Coefficients (MFCC). Inspired by the success of DNN models in automatic speech recognition (ASR) recently, it is also possible to compute the sufficient statistics in a supervised fashion (e.g. incorporating phonetic information) using phonetically aware DNN senone posteriors of frame-level features [5]. Another approach is the use of DNN for extracting phonetically aware frame-level features called Bottleneck Features (BF) with the bottleneck being a small hidden layer usually at the end of the network [6] and use them in *i*-vector extraction process.

2.2. Pre-processing

In order to achieve the state-of-the-art performance, a number of techniques have been proposed as pre-possessing steps prior to PLDA modeling. The common pre-processing includes Within-Class Covariance Normalization (WCCN) [7] followed by length normalization of *i*-vectors [11]. However, the way WCCN estimates the within-speaker scatter matrix does not adequately represents the directions of variation due to the mismatch condition of development and evaluation databases. In order to mitigate the adverse bias attributed to this mismatch, we proposed to use Source-Normalized WCCN (SN-WCCN). The source to be normalized in this context is different databases used during system development. If we consider the development portion of the SITW database as a different source of variability, we expect the speaker recognition system to better adapt to the evaluation condition as the development and evaluation portion of the SITW database represent similar conditions.

2.2.1. *Source-Normalized WCCN (SN-WCCN)*

Source normalization is an effective technique to compensate for speech source variation (i.e. microphone vs telephone sourced speech) in state-of-the-art *i*-vector/PLDA speaker recognition system [3]. Source-Normalized WCCN (SN-WCCN) [3] can be implemented by using the source-normalized within-speaker scatter matrix $\hat{\mathbf{S}}_W$ which is estimated as the variability not captured by the between speaker scatter matrix as

$$\hat{\mathbf{S}}_W = \mathbf{S}_T - \hat{\mathbf{S}}_B. \tag{2}$$

where $\mathbf{S}_T$ is the total scatter matrix and $\hat{\mathbf{S}}_B$ is the normalized between-speaker scatter matrix which are formulated as,

$$\mathbf{S}_T = \sum_{n=1}^{N} w_n {w_n}^T, \tag{3}$$

where N is the total number of *i*-vectors available for development (assuming zero-mean *i*-vectors), and

$$\hat{\mathbf{S}}_B = \sum_{k=1}^{K} \sum_{s=1}^{S_k} N_s^k (m_s^k - m_k)(m_s^k - m_k)^T. \tag{4}$$

where K is the number of sources available in development data, S_k is the number of speakers available for source k, m_s^k is the mean of N_s^k *i*-vectors from speaker s and source k and finally m_k is the mean of *i*-vectors for source k.

2.2.2. *Length normalization*

Due to Gaussian assumption made by PLDA, it has been shown that length normalization of *i*-vectors can approximately Gaussianize their distribution [11]. This has been shown to improve the performance of Gaussian PLDA to that of heavy-tailed PLDA [12].

2.3. Probabilistic Linear Discriminant Analysis (PLDA)

Probabilistic LDA (PLDA) provides a powerful mechanism to distinguish between-speaker variability which characterizes speaker information from other sources of undesired variability that characterizes distortions. However, to achieve this, it is required to provide PLDA with enough labeled data which contain multiple utterances of a speaker under different distortions.

A standard Gaussian PLDA assumes that an *i*-vector w, is modeled according to

$$w = m + \mathbf{V}y + z. \tag{5}$$

where, m is the mean of *i*-vectors, y is the speaker latent variable with standard normal prior and the residual z is normally distributed with zero mean and full covariance matrix Σ_z. In order to estimate the parameters of the model $(\mathbf{V}, \Sigma_z)$, PLDA uses the expectation-maximization (EM) algorithm [8].

After parameter estimation, for each two trial *i*-vectors w_1 and w_2, the verification score will be computed using the log likelihood ratio of the hypothesis $\mathcal{H}_s$, that both *i*-vectors are from the same speaker and the hypothesis $\mathcal{H}_d$, that they are from two different speakers,

$$score = \log \frac{p(w_1, w_2|\mathcal{H}_s)}{p(w_1, w_2|\mathcal{H}_d)}. \tag{6}$$

Considering the Gaussian assumption, the PLDA score can be computed in closed-form solution

$$\begin{aligned} score = &\log \mathcal{N}([{}^{w_1}_{w_2}]; [{}^{m}_{m}], [{}^{\mathbf{S}_T\ \mathbf{S}_B}_{\mathbf{S}_B\ \mathbf{S}_T}]) \\ &- \log \mathcal{N}([{}^{w_1}_{w_2}]; [{}^{m}_{m}], [{}^{\mathbf{S}_T\ \mathbf{0}}_{\mathbf{0}\ \ \mathbf{S}_T}]) \end{aligned}. \tag{7}$$

where, $\mathbf{S}_B = \mathbf{V}\mathbf{V}^T$ and $\mathbf{S}_T = \mathbf{S}_B + \Sigma_z$. For a clear exposition and a fast method to compute the score we refer you to [11].

3. Experiments

3.1. Experimental protocol

Experiments were performed on the SITW16 speaker recognition challenge database [2]. The database includes two non-overlapping portions of development and evaluation. However, the conditions observed in both portions are similar. There are 119 and 180 target speakers presented in development and evaluation portions respectively. Speech data were collected from a variety of microphone (podium, handheld, lapel, video and studio) types and degradation (clean, noise, codec, phone, reverb) conditions. Evaluation protocol is based on the challenge evaluation plan [1]. There were 6 trial conditions which are formed by different combination of enrollment (Core, Assist, AssistClean) and test (Core, Multi) conditions. The Core condition contains contiguous speech segment from a target speaker whereas in Assist, AssistClean or Multi, there might be more speakers including the target speaker. Annotations for target speaker is provided in Assist and AssistClean conditions but not for Multi condition with the aim of minimizing the labor intensive task of manual annotation. Table 1 summarizes trial conditions. Only results on the first three conditions including the required Core-Core set of trials, are reported in this paper.

Table 1: *Description of the 6 conditions of the evaluation portion of SITW database.*

enrolment	test	#tgt trials	#imp trials
$Core$	$Core$	$3,658$	$718,130$
$Assist$	$Core$	$18,444$	$3,546,040$
$AssistClean$	$Core$	$3,076$	$631,048$
$Core$	$Multi$	$10,045$	$2,000,638$
$Assist$	$Multi$	$34,596$	$6,711,932$
$AssitClean$	$Multi$	$5,828$	$1,194,400$

3.2. System configuration

The system has been trained using publically available databases including LDC releases of Switchboard Cellular Part II (SWBC2), NIST speaker recognition evaluation (SRE) 2004, 2005 and 2008 databases. For UBM and total variability matrix training, we used SWBC2, NIST SRE04 and SRE05. These data include 15740 speech segments from 1193 speakers. For SN-WCCN and PLDA modeling we included data from NIST SRE08. Therefore, training of both SN-WCCN and PLDA has been performed on 26616 speech segments from 2513 speakers.

For acoustic features, we used 20 MFCC features along with first and second order derivatives. These features were then passed through an energy based speech activity detector, followed by a cepstral mean and variance normalization (CMVN). We have trained a full covariance, gender-independent UBM model with 2048 Gaussians. We then trained a 500-dimensional *i*-vector extractor using the open-source Kaldi software [13]. The parameters of the PLDA model were tuned using the SITW development protocol and was set to 200-dimensional subspace for the eigenvoice latent components.

3.3. Calibration

In litrature, the performance of speaker recognition is usually reported in terms of calibrated-insensitive equal error rate (EER) or minimum decision cost function (C_{det}^{min}). It is only recently that it was required to submit scores as calibrated log-likelihood ratio in NIST SRE12. However, in real applications of speaker recognition there is a need to present recognition results in terms of calibrated log-likelihood-ratios. We have utilized BOSARIS Toolkit [14] for calibration of scores. C_{det}^{min} provides an ideal reference value for judging calibration. If $C_{det} - C_{det}^{min}$ is minimized, then the system can be said as to be well calibrated. To evaluate the calibrated performance of the speaker recognition system we can also use calibrated-sensitive criterion C_{llr} which is the cost of log-likelihood ratio and measures calibration over the entire range of effective priors [15].

3.4. Performance metrics

Several performance measures are used to determine system performance as indicated in the SITW16 evaluation plan. The primary metric for SITW challenge is based on the following detection cost function

$$C_{det} = C_{miss} \times P_{miss} \times P_{tar} + C_{fa} \times P_{fa} \times (1 - P_{tar}). \tag{8}$$

Equal costs between miss and false alarms ($C_{miss} = C_{fa} = 1.0$) has been used and target prior was set to $P_{tar} = 0.001$. In calculation of C_{det}, an optimal theoretical threshold equal to 6.90675 can be applied to calibrated log-likelihood ratio scores to determine P_{miss} and P_{fa} for performance reporting. Other metrics including, cost of log-likelihood ratio C_{llr}, equal error rate (EER) and minimum decision cost function C_{det}^{min} are also used to report performance. An alternative performance measure, average R-precision ($\bar{R}_{prec}$) which accounts for the proportion of relevant test segments for a given target speaker has been introduced for performance reporting.

4. Results and Discussion

Table 2 summarizes the system results for different conditions of both the development and evaluation portion of the SITW challenge. From the results, we can make the following observations. First, the lower performance in development portion

Table 2: *Performance comparison on different conditions of both development and evaluation portion of the SITW16 database with post evaluation improvement. The results are shown for both WCCN and SN-WCCN.*

		Development					**Evaluation**				
	Condition	EER	C_{det}	C_{det}^{min}	C_{llr}	$\bar{R}_{prec}$	EER	C_{det}	C_{det}^{min}	C_{llr}	$\bar{R}_{prec}$
WCCN	Core-Core	11.30	0.7101	0.7097	0.3761	0.6001	10.61	0.7156	0.7087	0.4207	0.5630
	Assist-Core	11.44	0.7644	0.7607	0.3812	0.5207	11.29	0.7683	0.7598	0.5024	0.4799
	AssistClean-Core	11.13	**0.6883**	0.6829	0.3694	**0.5749**	9.66	**0.6598**	0.6554	0.3910	0.5475
SN-WCCN	Core-Core	**10.92**	**0.7067**	**0.7023**	**0.3619**	**0.6025**	**10.49**	**0.6950**	**0.6793**	**0.3528**	**0.5696**
	Assist-Core	**11.06**	**0.7491**	**0.7476**	**0.3669**	**0.5273**	**10.65**	**0.7644**	**0.7517**	**0.4545**	**0.4857**
	AssistClean-Core	**10.49**	0.6950	**0.6793**	**0.3528**	0.5696	**9.00**	0.6624	**0.6471**	**0.3514**	**0.5529**

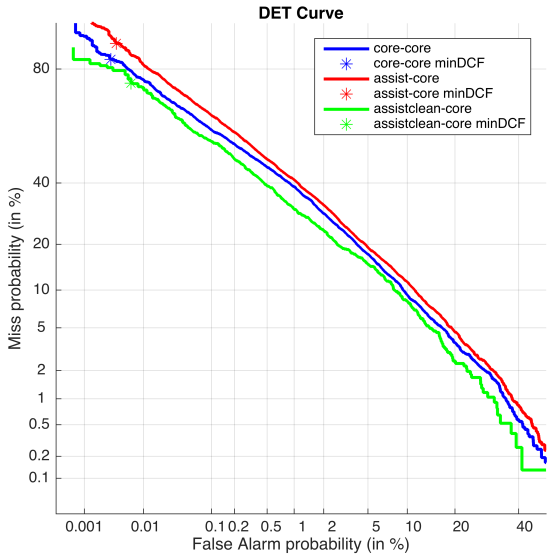

Figure 2: *Detection Error Trade-off for different conditions of evaluation portion of SITW16 using SN-WCCN for adaptation.*

of SITW database compared to the evaluation portion is mainly due to the use of development portion in system training to report results for the evaluation protocol. This is an indication of mismatch channel between NIST SREs and SITW database conditions and the system could benefit from the SITW development data for better adaptation to the evaluation condition. Second, the performance of the system could almost always be improved by applying SN-WCCN instead of WCCN prior to PLDA. This is due to the ability of SN-WCCN in better adaptation of the system to the evaluation condition. We have used different evaluations of SREs as different sources of variability to report results on development set. We then added the development portion of SITW database as a new source of variability to report results on evaluation set. The results for the Core-Core condition indicates much improvement in terms of decision cost function on the evaluation set. However, in Assist-Core or AssistClean-Core conditions we see better improvement in terms of EER. As could be expected in the AssistClean condition, SN-WCCN is less or not effective since we have a more clean condition. We have also reported system performance for different combination of microphone types and degradation trials from the Core-Core condition of the evaluation portion of the

Table 3: *System performance for different combination of microphone types on the core-core condition of the evaluation portion of SITW16 database.*

enrol/test	*podium*	*handheld*	*lapel*	*video*	*studio*
podium	0.3750	0.5351	0.4195	0.5154	0.0000
handheld	0.4985	0.6070	0.5334	0.8110	0.6521
lapel	0.5501	0.6421	0.3929	0.5636	0.3512
video	0.8069	0.8763	0.8149	0.8593	0.8780
studio	0.5417	0.6435	0.3757	0.5554	0.2966

Table 4: *System performance for different combination of degradation from the core-core sets of trials of the evaluation portion of SITW16 database.*

enrol/test	*clean*	*codec*	*noise*	*phone*	*reverb*
clean	0.3341	0.5317	0.5009	0.5000	0.4141
codec	0.6957	0.7899	0.7709	0.7649	0.7311
noise	0.3826	0.5837	0.6722	0.5833	0.5112
phone	–	0.3333	0.2838	–	–
reverb	0.3474	0.5435	0.6247	0.5000	0.4333

SITW'16 database in Table 3 and Table 4 respectively. These tables indicate which types of degradation or microphone type affect the performance the most. Results indicate that in codec degradation condition we have the worst recognition. The DET plot for different conditions of the evaluation portion in also shown in Figure 2. Interestingly, we do not see much improvement when evaluated on AssistClean-Core condition in comparison to Core-Core condition.

5. Conclusions

We have presented the speaker recognition system used for the SITW16 speaker recognition challenge. We investigated the impact of source normalization as a technique for system adaptation on the performance of the system. The proposed system is based on *i*-vectors extracted in unsupervised fashion using classic GMM. Post-evaluation analysis showed that using development portion of the SITW database as a different source of variability for Source-Normalized WCCN (SN-WCCN) results in improvement of speaker recognition system through better adaption of the system to evaluation condition.

6. References

[1] M. McLaren, A. Lawson, L. Ferrer, D. Castan, and M. Graciarena, "The speakers in the wild speaker recognition challenge plan," 2016.

[2] A. L. L. F. Mitchell McLaren, Diego Castan, "The speakers in the wild (sitw) speaker recognition database," in *INTERSPEECH 2016 – 16th Annual Conference of the International Speech Communication Association, September 812, San Francisco, California, USA, Proceedings*, 2016.

[3] M. Mclaren and D. Van Leeuwen, "Source-normalized lda for robust speaker recognition using i-vectors from multiple speech sources," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 3, pp. 755–766, 2012.

[4] 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, 2011.

[5] Y. Lei, L. Ferrer, M. McLaren *et al.*, "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*. IEEE, 2014, pp. 1695–1699.

[6] Y. Zhang, E. Chuangsuwanich, and J. R. Glass, "Extracting deep neural network bottleneck features using low-rank matrix factorization." in *ICASSP*, 2014, pp. 185–189.

[7] A. O. Hatch, S. S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for svm-based speaker recognition." in *Interspeech*, 2006.

[8] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1–8.

[9] M. McLaren, M. I. Mandasari, and D. A. van Leeuwen, "Source normalization for language-independent speaker recognition using i-vectors," 2012.

[10] M. McLaren and D. A. Van Leeuwen, "Gender-independent speaker recognition using source normalisation," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4373–4376.

[11] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems." in *Interspeech*, 2011, pp. 249–252.

[12] P. Kenny, "Bayesian speaker verification with heavy-tailed priors." in *Odyssey:Speaker and Language recognition workshop*, 2010, p. 14.

[13] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[14] N. Brümmer and E. de Villiers, "The bosaris toolkit user guide: Theory, algorithms and code for binary classifier score processing," *Documentation of BOSARIS toolkit*, 2011.

[15] D. A. Van Leeuwen and N. Brümmer, *An introduction to application-independent evaluation of speaker recognition systems*. Springer, 2007.