A Study of X-vector Based Speaker Recognition on Short Utterances

A. Kanagasundaram[1,2]*, S. Sridharan*[2]*, G. Sriram*[3]*, S. Prachi*[3]*, C. Fookes*[2]

[1]University of Jaffna
[2]Speech and Audio Research Lab, SAIVT, Queensland University of Technology
[3]LEAP Lab, Indian Institute of Science

ahilan@eng.jfn.ac.lk, sriramg@iisc.ac.in, c.fookes,s.sridharan@qut.edu.au

Abstract

The aim of this work is to gain insights into how the deep neural network (DNN) models should be trained for short utterance evaluation conditions in an x-vector based speaker verification system. The study suggests that the speaker embedding can be extracted with reduced dimensions for short utterance evaluation conditions. When the speaker embedding is extracted from deeper layer which has lower dimension, the x-vector system achieves 14% relative improvement over baseline approach on EER on NIST2010 5sec-5sec truncated conditions. We surmise that since short utterances have less phonetic information speaker discriminative x-vectors can be extracted from a deeper layer of the DNN which captures less phonetic information. Another interesting finding is that the x-vector system achieves 5% relative improvement on NIST2010 5sec-5sec evaluation condition when the back-end PLDA is trained using short utterance development data. The results confirms the intuitive expectation that duration of development utterances and the duration of evaluation utterances should be matched. Finally, for the duration mismatch condition, we propose a variance normalization approach for PLDA training that provides a 4% relative improvement on EER over baseline approach.

Index Terms: Speaker verification, PLDA, DNN, x-vector, speaker embedding, short utterance

1. Introduction

The development of state-of-the-art speaker verification systems for short utterances is an active area of research since potential users of the system prefer short utterance for enrollment and authentication. For wider development of speaker verification technology, high accuracy need to be achieved with short utterances; however current state-of-the-art such as x-vector based systems still require significant amount of speech for enrollment and verification. Researchers have been working to improve the performance of speaker verification on short utterance evaluation conditions [1, 2, 3, 4].

Previously, joint factor analysis (JFA), i-vector, probabilistic linear discriminant analysis (PLDA) based speaker recognition systems were studied on short utterances [1, 5, 2, 3, 4]. These studies have shown that when the evaluation utterance length is reduced, it significantly affects the performance [1, 2, 4].

Recently, the deep learning approaches have been incorporated into i-vector-based speaker recognition systems using two main approaches: (1) A speech-based Deep Neural Network (DNN) is used to extract bottleneck (BN) features from the middle layer restricted in dimensionality [6]; (2) DNN senone approach, where the calculation of Baum-Welch statistics is based on speech-based DNN [7, 8]. Though DNN senone based speaker verification systems achieve the state-of-the-art performance, this approach is computationally expensive and it is infeasible to use in practical applications. More recently, researchers proposed the end-to-end x-vector speaker recognition systems [9, 10, 11]. In the end-to-end x-vector approach, deep neural network is fed with a variable length utterance and maps it to a speaker embedding [10]. Snyder *et al* proposed data augmentation to achieve further improvement on DNN speaker embedding based speaker recognition [12]. Existing DNN architectures work well with long utterances but the performance drops when the utterance length is reduced [12].

In this paper, we investigate approaches for improving the performance of time delay DNN architecture in x-vector based speaker verification under short utterance evaluation conditions. Since short utterances are likely to contain less phonetic information compared to long utterances, we hypothesize that it is sufficient to use lower dimensional speaker embedding which can capture all the variations present in the short utterances. In an x-vector based speaker recognition system, the speaker embedding is normally extracted from the layer adjacent to the stats pooling layer of the DNN as this high dimensional layer is expected capture all the speaker discriminative information at phonetic level. Since the phonetic information in short utterances is small, we argue that it would be sufficient to capture speaker embedding from deeper low dimensional layers of the DNN.

We also investigate the training of the PLDA which is used as the back-end of the x-vector based speaker verification system. When PLDA is trained using full-length data and the speaker recognition system is evaluated on short-length data, the data duration mismatch significantly affects the performance. To overcome this mismatch, the PLDA can be trained using short-length data. However, this causes mismatch when evaluated on longer utterances. In this paper, novel utterance variance transformation is proposed to compensate the data duration mismatch and improve the performance of under utterance length mismatch evaluation conditions.

This paper is structured as follows: Section 2.1 details the extraction of speaker embedding features. Section 2.2 provides the details of PLDA classifier. The proposed short utterance variance transformation is detailed in Section 2.3. The experimental protocol and corresponding results are given in Section 3 and Section 4. Section 5 concludes the paper.

Table 1: *The network architecture for DNN embedding. For core-core condition we follow the standard process of extracting the x-vector from segment 6, next to stats pooling [12]. For the short utterance 5sec-5sec condition we propose to extract a deeper lower dimensional x-vector from segment 7. N in the softmax layer corresponds to the number speakers in the training set. The values in bold indicate the proposed optimal selection for long and short utterance conditions.*

Layer	**Layer context**	**Input × output**	
		Core-core condition	5sec-5sec condition
TDNN - frame 1	$[t-2, t+2]$	115×512	115×512
TDNN - frame 2	$\{t-2, t, t+2\}$	1536×512	1536×512
TDNN - frame 3	$\{t-3, t, t+3\}$	1536×512	1536×512
frame 4	$\{t\}$	512×512	512×512
frame 5	$\{t\}$	512×1500	512×1500
stats pooling	$[0, T]$	1500×3000	1500×3000
segment 6	0	$\mathbf{3000} \times \mathbf{512}$	3000×150
segment 7	0	512×512	$\mathbf{150} \times \mathbf{150}$
softmax	0	$512 \times$ N	$150 \times$ N

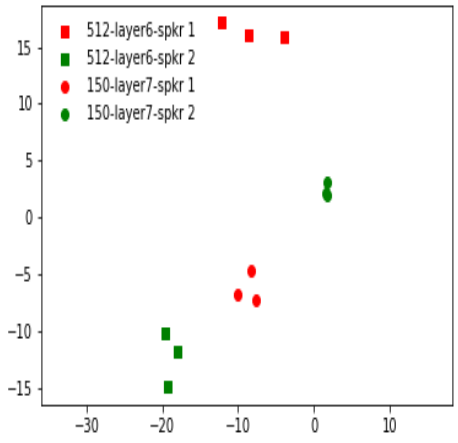

Figure 1: *Distribution of first two dimensions of PCA projected space of x-vector features of two different speakers for different short utterance durations.*

2. The DNN x-vector speaker embedding system

In an x-vector speaker embedding system, a time-delay neural network (TDNN) is used to compute the speaker embedding from variable length utterances. Once the fixed-length speaker embedding (x-vector) is obtained from speech segments, PLDA is used as back end to classify the speakers.

2.1. Speaker embeddings extraction

The time-delay DNN is trained using a large amount of training data to discriminate between the speakers [10]. The Table 1 shows the proposed network architecture for full-length and short-length evaluation conditions. The first four layers of the networks operate at the frame-level. If the t is the current time step, $t-2, t-1, t, t+1, t+2$ frames are spliced together at the input layer. As the 23 MFCC features are extracted for our experiments, the input layer dimension is 115. The size of the output layer is 512. In the first hidden layer, $t-2, t, t+2$ frames are spliced together and the size input is 1536 (512×3). In the second hidden layer, $t-3, t, t+3$ frames are spliced together and the size of the input is 1536 (512×3). The dimensions of the third and fourth layers are respectively 512 and 1500. The fifth layer is stats pooling where all the frames are aggregated together and the mean and standard deviation are estimated and concatenated together (1500×2). From this layer onward, the utterance level parameters are estimated. The sixth and seventh layers have a dimension of 512 for long utterance evaluation conditions and a dimension of 150 for short utterance evaluation conditions. Finally, the softmax layer is trained to classify the speakers from the training dataset. After the network training, the x-vector speaker embedding (512) can be extracted from layer 6 or from layer 7. We argue that for long utterance evaluations the embedding should be extracted from layer 6 as customarily done but for short utterance evaluations it would be advantageous to use layer 7.

The x-vectors extracted from short utterances can also vary considerably with changes in phonetic content between the utterances. We illustrate variation of short utterances spoken by two speakers by comparing the most significant x-vector features as shown using the first two dimensions of PCA-projected x-vectors in Figure 1. It can be observed that when the speaker embedding is extracted from deeper layer (7^{th} layer) which has lower dimension (150), the intra-speaker variations of the x-vector for each speaker is less which would contribute to higher accuracy.

2.2. PLDA classifier

After fixed length speaker embedding is extracted from variable length speaker utterances as explained in Section 2.1, standard linear discriminant analysis (LDA) projection is applied on speaker embedding. The LDA dimension is selected as 150 and 75 respectively based on performance for NIST2010 core-core and 5sec-5sec evaluation conditions. Subsequent to this dimension reduction, length normalization is applied and PLDA model parameters are estimated [13, 14]. Scores are calculated using the batch likelihood ratio between a target and test x-vectors.

2.3. Short utterance variance based transformation approach

When the PLDA is trained on full-length data and evaluated on short-length data, the data duration mismatch significantly affects the performance. To overcome this problem, for 5sec-5sec matched evaluation conditions, the PLDA can be trained on short-length duration utterances. However, when the PLDA training data is short-length but evaluation data is core-5sec, there is a mismatch between evaluation data and development data and the performance of speaker verification system drops. To overcome this problem a short utterance variance based transformation approach is proposed to compensate mismatch between the duration of the development data and evaluation data.

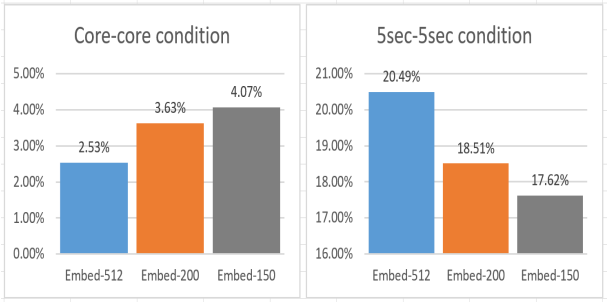

Figure 2: *Performance comparison of Speaker recognition systems on NIST 2010 core-core and 5sec-5sec conditions when DNN is trained using different embedding size.*

The short utterance variance (SUV) is captured as follows,

$$\mathbf{SUV} = \sum_{s=1}^{N} (\mathbf{w}_{full} - \mathbf{w}_{short})(\mathbf{w}_{full} - \mathbf{w}_{short})^T, \quad (1)$$

Where $\mathbf{w}_{full}$ and $\mathbf{w}_{short}$ represent the full-length and short-length x-vectors. N is total number of x-vectors in development set. The transformation matrix, $\mathbf{D}$ is estimated using the Cholesky decomposition of $\mathbf{DD}^T = \mathbf{SUV}$.

Following this PLDA training and evaluation data are transformed as follows,

$$\mathbf{w}_{SUV} = \mathbf{D}^T \mathbf{w}, \quad (2)$$

where $\mathbf{w}$ is raw x-vector and $\mathbf{w}_{SUV}$ is duration mismatch compensated x-vector.

3. Experimental methodology

The experiments were done using kaldi toolkit [15]. The speaker embedding based experiments were evaluated using the NIST 2010 corpora [16]. For NIST 2010, the performance was evaluated using the equal error rate (EER). For short utterance evaluation conditions, NIST 2010 core-core was truncated into 5sec-5sec.

The DNN architecture was trained using NIST and Switchboard data. The NIST data consists of NIST SRE 04, 05, 06, 08, and Switchboard data consists of Switchboard 2 Phases 1, 2, 3, Switchboard Cellular [17]. The DNN speaker embedding system is trained with 23 MFCCs without their first derivative. DNN speaker embedding system is trained using the Kaldi recipe. The PLDA classifiers are trained on NIST SRE 04, 05, 06 and 08 dataset.

For the augmented system, noise and reverberation are added into the training data to increase the amount and the diversity of the existing data as proposed in [12].

4. Results and Discussion

4.1. DNN architecture training for short utterance evaluation conditions

We first investigate the effect on speaker verification performance when the speaker embedding is trained with the different embedding dimensions. For these experiments, the DNN is trained using the 10sec chunks. It can be seen from Figure 2 that the speaker verification achieves about 15% relative improvement on 5sec-5sec evaluation condition when the DNN embedding is trained with lower dimension (150). It was also found from experiments that further reduction of x-vector embedding size below 150 offer no performance gain; in fact the accuracy decreases with reduction of the size below 150. On the other hand under the core-core condition, when the size of embedding is reduced, it significantly affects the performance of speaker verification system. These results suggest that the embedding should be trained using lower dimension for short utterance evaluation conditions. We believe that this is because the short utterances have less phonetic information and the low dimensional embedding is optimal to capture the speaker variations in short utterance conditions.

The speaker verification performance is also investigated when the DNN is trained using 10sec, 5sec and 1sec chunk lengths. When the chunk length is reduced to 1sec, it degrades the performance of NIST2010 core-core condition. We believe that this is due to fact that the speaker discriminating phonetic information is captured using time-delay network become less as the chunk length reduces. On the other hand, when the chunk length is reduced for evaluation with 5sec-5sec evaluation condition, the performance is marginally improved.

Table 2: *Performance comparison of DNN speaker embedding speaker recognition systems when DNN is trained using argumentation approach.*

DNN training	**5sec-5sec** EER
Without augmentation	17.62%
With argumentation	**15.57%**

4.2. Improving short utterance performance using data augmentation approach

It was well known that the DNN architecture needs to be trained using substantial amount of data [12]. However, it is hard to collect significant amount of labeled data for speaker verification task. To overcome this problem, data augmentation approach has been proposed [12]. However, it has not been confirmed that the data argumentation approach is effective for short evaluation conditions. It can be seen from results in Table 2 that data augmentation based DNN approach is effective for short utterance as well and shows over 11% improvement on NIST2010 5sec-5sec truncated evaluation condition.

The speaker embedding approach achieves significant improvement in long utterance evaluation conditions as speaker discriminative embedding is extracted. DNN lower layers have more phonetic information and when the speaker embedding is extracted from 6^{th} layer, x-vector system achieves state-of-the-art performance on long utterance evaluation conditions. However, short utterance data has less phonetic information. It's therefore surmised that the speaker discriminative x-vectors can be extracted in the case of short utterances from a deeper layer in accordance with this assumption. We extract the speaker embedding are extracted from the 7^{th} layer instead of 6^{th} layer for short utterance evaluations.

Table 3 compares the effects of the embedding selection on the performance of speaker recognition on NIST2010 5sec-5sec truncated conditions. It can be observed from results that the x-vector speaker recognition achieves over 14% relative improvement on EER on 5sec-5sec truncated conditions when the

Table 3: *Performance comparison of speaker recognition systems on 5sec-5sec conditions when speaker embedding is extracted from 6^{th} and 7^{th} layer.*

DNN training	**5sec-5sec** EER
6^{th} layer	15.57%
7^{th} layer	**13.35%**

Table 4: *Performance comparison of speaker recognition systems on 5sec-5sec condition when PLDA is trained using full-length, short-length and dataset transformation approach.*

PLDA training	**Full-5sec** EER	**5sec-5sec** EER
Full-length	7.02%	13.35%
Short-length	7.49%	**12.68%**
Proposed transformation	**6.74%**	-

speaker embedding are extracted from the 7^{th} layer instead of 6^{th} layer. The experiment results confirm that for short utterance evaluations it is better to use the layer 7 rather than 6 as short utterance have less phonetic information.

4.3. Compensating the mismatch between development data and short utterance evaluation data

When the speaker recognition is developed on one database and evaluated on another database, the dataset mismatch between PLDA training data and evaluation data significantly affects the performance. Several domain compensation approaches have been proposed to address the mismatch. Similarly, when the PLDA is trained on full-length utterances and evaluated on short-short or full-short evaluation conditions, the duration mismatch between PLDA training data and evaluation data affects the performance.

Table 4 compares the performance of speaker recognition on full-5sec and 5sec-5sec conditions when the PLDA is trained using full-length, short-length and the short length transformation approach proposed in Section 2.3 to compensate the utterance duration mismatch. When the PLDA is trained using short-length data, the x-vector system achieves over 5% relative improvement on 5sec-5sec evaluation conditions. We conclude that PLDA should be trained short-length data for evaluation on short utterance data, so that there is no utterance duration mismatch between development and evaluation data.

However, if the trained system needs to be evaluated on both full and short utterances, the utterance duration mismatch will occur between development data and evaluation data. To overcome this problem, we have proposed in Section 2.3 a short utterance variance based transformation approach to compensate mismatch between full-length PLDA training data and full-5sec evaluation data. The proposed short utterance variance transformation approach achieves 4% relative improvements on full-5sec evaluation condition.

5. Conclusion

In this paper we have studied the effects of utterance duration mismatch in the design of a x-vector time delay DNN based speaker verification system. The study suggests that the speaker embedding can be extracted using lower dimensions for short utterance evaluation conditions as the short utterances have less phonetic information and the low dimensional embedding is enough to capture the speaker variations. When the embedding was extracted from low dimensional deeper layer, x-vector system achieved over 14% relative improvement over the baseline approach on EER on NIST2010 5sec-5sec truncated conditions. To compensate for utterance duration mismatch between development data and evaluation data in the back-end PLDA of the x-vector system, we proposed a short utterance variance based transformation which shows 4% relative improvement on full-5sec mismatch evaluation conditions over the baseline approach.

6. Acknowledgments

This research was supported by the India Science and Research Fellowship (ISRF) programme.

7. References

[1] R. Vogt, B. Baker, and S. Sridharan, "Factor analysis subspace estimation for speaker verification with short utterances," in *Interspeech 2008*, Brisbane, Australia, September 2008.

[2] M. McLaren, R. Vogt, B. Baker, and S. Sridharan, "Experiments in SVM-based speaker verification using short utterances," in *Proc. Odyssey Workshop*, 2010.

[3] A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, and M. Mason, "i-vector based speaker recognition on short utterances," in *Proceed. of INTERSPEECH*. International Speech Communication Association (ISCA), 2011, pp. 2341–2344.

[4] A. Kanagasundaram, R. J. Vogt, D. B. Dean, and S. Sridharan, "PLDA based speaker recognition on short utterances," in *The Speaker and Language Recognition Workshop (Odyssey 2012)*. ISCA, 2012.

[5] R. Vogt, C. Lustri, and S. Sridharan, "Factor analysis modelling for speaker verification with short utterances," in *Odyssey: The Speaker and Language Recognition Workshop*, 2008.

[6] M. McLaren, Y. Lei, and L. Ferrer, "Advances in deep neural network approaches to speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4814–4818.

[7] F. Richardson, D. Reynolds, and N. Dehak, "A unified deep neural network for speaker and language recognition," *arXiv preprint arXiv:1504.00923*, 2015.

[8] D. Snyder, D. Garcia-Romero, and D. Povey, "Time delay deep neural network-based universal background models for speaker recognition," in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*. IEEE, 2015, pp. 92–97.

[9] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, "Deep neural network-based speaker embeddings for end-to-end speaker verifi-

cation," in *Spoken Language Technology Workshop (SLT), 2016 IEEE*. IEEE, 2016, pp. 165–170.

[10] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *Proc. Interspeech*, 2017, pp. 999–1003.

[11] C. Li, X. Ma, B. Jiang, X. Li, X. Zhang, X. Liu, Y. Cao, A. Kannan, and Z. Zhu, "Deep speaker: an end-to-end neural speaker embedding system," *arXiv preprint arXiv:1705.02304*, 2017.

[12] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," *Submitted to ICASSP*, 2018.

[13] P. Kenny, "Bayesian speaker verification with heavy tailed priors," in *Proc. Odyssey Speaker and Language Recogntion Workshop, Brno, Czech Republic*, 2010.

[14] D. Garcia-Romero and C. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *International Conference on Speech Communication and Technology*, 2011, pp. 249–252.

[15] 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.

[16] A. F. Martin and C. S. Greenberg, "The nist 2010 speaker recognition evaluation," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[17] C. Cieri, D. Miller, and K. Walker, "The fisher corpus: a resource for the next generations of speech-to-text." in *LREC*, vol. 4, 2004, pp. 69–71.